



**UNIVERSIDADE FEDERAL DO CEARÁ**  
**CENTRO DE TECNOLOGIA**  
**DEPARTAMENTO DE ENGENHARIA ELÉTRICA**  
**PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA**  
**DOUTORADO EM ENGENHARIA ELÉTRICA**

**EUGENIO PEIXOTO JUNIOR**

**DIAGNÓSTICO MULTIMODAL DA DOENÇA DE PARKINSON COM**  
**ARQUITETURA AGENTIVA COLABORATIVA BASEADA EM INTERNET DE**  
**MODELOS DE LINGUAGEM MÉDICA**

**FORTALEZA**

**2026**

EUGENIO PEIXOTO JUNIOR

DIAGNÓSTICO MULTIMODAL DA DOENÇA DE PARKINSON COM ARQUITETURA  
AGENTIVA COLABORATIVA BASEADA EM INTERNET DE MODELOS DE  
LINGUAGEM MÉDICA

Tese apresentada ao Programa de Pós-Graduação em Engenharia Elétrica do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de doutor em Engenharia Elétrica. Área de Concentração: Sinais e sistemas.

Orientador: Prof. Dr. Victor Hugo Costa de Albuquerque.

FORTALEZA

2026

EUGENIO PEIXOTO JUNIOR

DIAGNÓSTICO MULTIMODAL DA DOENÇA DE PARKINSON COM ARQUITETURA  
AGENTIVA COLABORATIVA BASEADA EM INTERNET DE MODELOS DE  
LINGUAGEM MÉDICA

Tese apresentada ao Programa de Pós-Graduação em Engenharia Elétrica do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de doutor em Engenharia Elétrica. Área de Concentração: Sinais e sistemas.

Aprovada em: 27 de Janeiro de 2026

BANCA EXAMINADORA

---

Prof. Dr. Victor Hugo Costa de  
Albuquerque (Orientador)  
Universidade Federal do Ceará (UFC)

---

Prof. Dr. Antonio Carlos da Silva Barros  
IEDS – Instituto de Engenharia e  
Desenvolvimento Sustentável (UNILAB)

---

Prof. Dr. Carlos Mauricio Jaborandy de Mattos  
Dourado Junior  
Instituto Federal de Educação, Ciência e  
Tecnologia do Ceará (IFCE)

---

Prof. Dr. Francisco Nauber Bernardo Gois  
Conferência das Grandes Escolas (CGE)

---

Prof. Dr. Bruno Ricelli dos Santos Silva  
Universidade Federal do Ceará (UFC)

---

Dr. Stephen Rathinaraj Benjamin  
Universidade Federal do Ceará (UFC)

Ao meu orientador Prof. Dr. Victor Hugo Costa  
de Albuquerque.

## AGRADECIMENTOS

À Universidade Federal do Ceará (UFC) e ao PPGEE por concretizar meu doutorado nesta renomada instituição de ensino público.

Principalmente, ao Professor Dr. Victor Hugo Costa de Albuquerque, meu orientador e chefe do *Biomedical Data Analytics Research Group (BioData)*, deixo meu mais profundo e sincero agradecimento. Sua orientação incansável, que me acompanhou desde a graduação, passando pelo mestrado até o doutorado, que foi essencial não apenas para o desenvolvimento desta tese, mas também para a construção da minha trajetória como pesquisador.

Obrigado por acreditar no meu potencial desde os primeiros passos, pelos desafios impostos para evolução acadêmica e por compartilhar, com generosidade, seu conhecimento, sua visão científica e seu exemplo de liderança. Sua influência ultrapassa os limites deste trabalho e deixa uma marca duradoura na minha formação pessoal e profissional.

Ao Prof. Dr. Fabrício Gonzalez Nogueira, líder do Grupo de Pesquisa em Automação e Robótica (GPAR), por me acolher de forma excepcional desde meu ingresso no PPGEE. Sua receptividade, ao me apresentar aos colegas e proporcionar um ambiente acolhedor nos primeiros momentos, foi essencial para minha adaptação. Agradeço também por sua constante disponibilidade para conversas, sugestões e reflexões sobre projetos, sempre com atenção e incentivo.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo apoio financeiro fundamental, que garantiu as condições necessárias para que eu pudesse me dedicar integralmente ao desenvolvimento desta tese.

"It's only after we've lost everything that we're  
free to do anything."  
(Tyler Durden)

## RESUMO

Nesta tese é proposta uma plataforma inteligente voltada ao apoio no diagnóstico da Doença de Parkinson (DP), integrando tecnologias avançadas como Internet das Coisas Médicas (IoMT), Recuperação Aumentada por Geração (RAG) multimodal, compressão eficiente de modelos e operações com modelos de linguagem médica, *LLMOps* e *VLMOps*. Diferentemente de abordagens que utilizam modelos isolados ou focam em modalidades específicas, esta proposta define um ecossistema clínico agentivo, formado por agentes autônomos e colaborativos, baseados em LLMs ou SLMs, com suporte a *reinforcement learning* (RL). A arquitetura do sistema é adaptável, explicável e centrada no paciente, com foco em confiabilidade, tempo de resposta e aplicabilidade em ambientes clínicos reais. A fusão de dados heterogêneos, incluindo texto clínico, exames de imagem, sinais sensoriais e gravações de voz, é realizada por meio de variantes de RAG, como Multimodal RAG, *GraphRAG* e *Agentic RAG*, permitindo inferência contextualizada sem necessidade de re-treinamentos intensivos. A metodologia inclui a avaliação da eficiência operacional e do custo computacional em cenários simulados de produção, utilizando métricas como uso de GPU, custo por chamada de LLMs e escalabilidade. Técnicas de compressão, como quantização 4-bit e *QLoRA*, foram aplicadas para viabilizar o uso de modelos robustos em dispositivos *edge*. Por fim, agentes inteligentes refinam suas estratégias de inferência por meio de pós-treinamento com aprendizado por reforço, para otimizar tanto a precisão quanto a aplicabilidade prática do sistema. Os experimentos demonstraram desempenho competitivo entre diferentes classificadores, com destaque para o *Gradient Boosting* com acurácia de 83,85%, *F1-score* de 88,74%, e *HistGradient Boosting* com acurácia de 82,92%, *F1-score* de 88,12%, *AUC-ROC* de 0.909, superando abordagens tradicionais como SVM e KNN em robustez multimodal. Modelos como SVM e KNN demonstraram, em alguns casos, maior robustez em termos de sensibilidade, ao identificar corretamente uma proporção elevada de instâncias positivas, indicando uma baixa taxa de falsos negativos. A proposta se destaca pela capacidade de integração com fluxos clínicos reais, com potencial para se tornar um modelo de referência para plataformas neurointeligentes aplicadas à saúde digital conectada a internet.

**Palavras-chave:** Doença de Parkinson; Multi-Agentes; Aprendizado de Máquina; Aprendizado Profundo; Redes Neurais Convolucionais; RAG Multimodal; Diagnóstico Médico; Fusão Multimodal.

## ABSTRACT

This thesis proposes an intelligent platform to support the diagnosis of Parkinson's Disease (PD), integrating advanced technologies such as the Internet of Medical Things (IoMT), Multimodal Retrieval-Augmented Generation (RAG), efficient model compression, and medical language model operations (LLMOps and VLMOps). Unlike previous approaches that rely on isolated models or single data modalities, this work defines a clinical agentic ecosystem composed of autonomous and collaborative agents powered by Large or Small Language Models (LLMs/SLMs), with reinforcement learning (RL). The system architecture is designed to be adaptable, explainable, and patient-centered, with an emphasis on reliability, fast response times, and practical deployment in real clinical environments. Heterogeneous data modalities, including clinical text, medical imaging, sensor signals, and voice recordings, are fused through advanced RAG variants such as Multimodal RAG, GraphRAG, and Agentic RAG, enabling contextualized inference without the need for extensive fine-tuning. The methodology includes evaluating computational cost and operational efficiency in simulated production environments, using metrics such as GPU utilization, LLM call cost, and infrastructure scalability. Compression techniques such as 4-bit quantization and QLoRA are employed to enable robust models to run on edge devices. Finally, intelligent agents refine their inference strategies through post-training with reinforcement learning, using rewards derived from clinical analyses to improve both the accuracy and practical applicability of the system. Experimental results showed competitive performance across classifiers, with Gradient Boosting achieving 83.85% accuracy and an F1-score of 88.74%, and HistGradient Boosting reaching 82.92% accuracy, an F1-score of 88.12%, and an AUC-ROC of 0.909. While models such as SVM and KNN demonstrated higher sensitivity in some cases, they also showed a greater rate of false positives, highlighting the superior balance of the proposed approach. This work stands out for its capacity to integrate with real clinical workflows and holds strong potential to serve as a reference model for neurointelligent platforms in connected digital healthcare.

**Keywords:** Parkinson's Disease; Multi-Agent Systems; Machine Learning; Deep Learning; Convolutional Neural Networks; Multimodal RAG; Medical Diagnosis; Multimodal Fusion.

## LISTA DE FIGURAS

Figura 1 – Arquitetura do modelo CBOW ( <i>Continuous Bag of Words</i> ) . . . . .	29
Figura 2 – Arquitetura do modelo <i>Skip-Grim</i> . . . . .	30
Figura 3 – Arquitetura RAG . . . . .	35
Figura 4 – Estrutura da entidade hierárquica . . . . .	38
Figura 5 – Visão geral dos serviços da plataforma IoMT. . . . .	58
Figura 6 – Arquitetura de DataOps da plataforma IoMT proposta. . . . .	62
Figura 7 – Tela de login da plataforma IoMT . . . . .	65
Figura 8 – Tela de home plataforma IoMT . . . . .	66
Figura 9 – O processo autônomo de tomada de decisão clínica que integra uma arquitetura RAG multimodal com um codificador e o decodificador do modelo de linguagem Qwen 2.5. . . . .	74
Figura 10 – Exemplo de uma classificação usando LLM-as-Judge . . . . .	80
Figura 11 – Espectrogramas do Áudio Pré-processado. A primeira imagem se refere ao paciente do grupo de controle e a segunda imagem é referente ao paciente com DP. . . . .	91
Figura 12 – Distribuição de Classes Antes e Depois do Balanceamento . . . . .	93
Figura 13 – Curva ROC AUC . . . . .	95
Figura 14 – Matriz de confusão - Sensor . . . . .	96
Figura 15 – Curva ROC AUC . . . . .	97
Figura 16 – Matriz de confusão - Áudio . . . . .	99
Figura 17 – Representação da matriz de confusão de todos os modelos usando ajuste fino. . . . .	100
Figura 18 – Representação da matriz de confusão de todos os modelos usando aprendizagem por transferência. . . . .	101
Figura 19 – Representação da matriz de confusão de todos os modelos usando aprendizagem por transferência. . . . .	102
Figura 20 – Representação do gráfico de acurácia por época do treinamento dos modelos usando ajuste fino. . . . .	102

## LISTA DE TABELAS

Tabela 1 – Características Científicas e Analíticas dos trabalhos . . . . .	52
Tabela 2 – Recursos e Aplicações Práticas dos trabalhos . . . . .	52
Tabela 3 – Aprendizado de máquina e Aprendizado profundo - Características Científicas e Analíticas dos Estudos . . . . .	53
Tabela 4 – Aprendizado de máquina e Aprendizado profundo - Recursos e Aplicações Práticas dos Estudos . . . . .	54
Tabela 5 – IoMT - Características Científicas e Analíticas dos Estudos . . . . .	55
Tabela 6 – IoMT - Recursos e Aplicações Práticas dos Estudos . . . . .	56
Tabela 7 – Tempo de aquisição de dados por tarefas . . . . .	84
Tabela 8 – Schema dos dados no banco vetorial . . . . .	85
Tabela 9 – Descrição dos campos do schema . . . . .	85
Tabela 10 – Distribuição dos sujeitos por grupo . . . . .	90
Tabela 11 – Distribuição de sexo por grupo e resultado do teste qui-quadrado . . . . .	90
Tabela 12 – Estatísticas descritivas de idade por grupo . . . . .	90
Tabela 13 – Resultados obtidos a partir da base de dados sensorial . . . . .	95
Tabela 14 – Resultados obtidos a partir da base de dados de áudios . . . . .	98
Tabela 15 – Resultados obtidos a partir da base de dados de imagens . . . . .	103

## LISTA DE ABREVIATURAS E SIGLAS

AGI	<i>Artificial General Intelligence</i>
AUC-ROC	<i>Area Under The Receiver Operating Characteristic curve</i>
AWS	<i>Amazon Web Services</i>
BART	<i>Bidirectional and Auto-Regressive Transformer</i>
BERT	<i>Bidirectional Encoder Representations from Transformers</i>
CBOW	<i>Continous Bag of Words</i>
CLIP	<i>Contrastive Language–Image Pre-training</i>
CNNs	Redes Neurais Convolucionais
COMA	<i>Counterfactual Multi-Agent Policy Gradients</i>
CTDE	<i>Centralized Training with Decentralized Execution</i>
DataOps	Operações nos Dados
DDPG	<i>Deep Deterministic Policy Gradient</i>
Dec-POMDP	Descentralizado e Parcialmente Observável
DICOM	<i>Digital Imaging and Communications in Medicine</i>
DP	Doença de Parkinson
DPO	<i>Direct Preference Optimization</i>
DPR	<i>Dense Passage Retrieve</i>
GPT	<i>Generative Pre-trained Transformer</i>
GPU	Unidade de Processamento Gráfico
GRPO	<i>Group Relative Policy Optimization</i>
HL7-FHIR	<i>Health Level Seven – Fast Healthcare Interoperability Resources</i>
IA	Inteligência Artificial
IoT	<i>Internet of Things</i>
KNN	<i>K-Nearest Neighbor</i>
LCS	<i>Longest Common Subsequence</i>
LLMOps	<i>Large Language Models Operations</i>
LLMs	Linguagens de Larga Escala
LSTMs	<i>Long Short-Term Memory</i>
MADDPG	<i>Multi-Agent Deep Deterministic Policy Gradient</i>
MARL	<i>Multi-Agent Reinforcement Learning</i>

MFCC	<i>Mel-Frequency Cepstral Coefficients</i>
MHA	<i>Multi-Head Attention</i>
MHSA	<i>Multi-Head Self-Attention</i>
ML	<i>Machine Learning</i>
MLOps	<i>Machine Learning Operations</i>
MMHSA	<i>Masked Multi-Head Self-Attention</i>
NLP	<i>Natural Language Processing</i>
NNLM	<i>Neural Network Language Model</i>
PNL	Processamento de Linguagem Natural
PPMI	<i>Parkinson's Progression Markers Initiative</i>
PPO	Proximal Policy Optimization
QLoRA	<i>Quantized Low-Rank Adaptation</i>
RL	<i>reinforcement learning</i>
RLHF	<i>Reinforcement Learning from Human Feedback</i>
RM	Ressonância Magnética
RMI	Ressonância Magnética por Imagem
RMT	<i>Reward Model Training</i>
RNNs	<i>Recurrent Neural Networks</i>
SDPA	<i>Scaled Dot-Product Attention</i>
SFT	<i>Supervised Fine-Tuning</i>
SLMs	Pequenos Modelos de Linguagem
SVM	<i>Support Vector Machine</i>
T5	<i>Text-to-Text Transfer Transformer</i>
TC	Tomografia Computadorizada
VDN	Value Decomposition Networks
ViT	<i>Vision Transformers</i>
VLMOps	<i>Vision Language Models Operations</i>
VLMS	Modelos de Linguagem de Visão
VQA	<i>Vision Question Answering</i>

## LISTA DE SÍMBOLOS

$\Theta$	Theta
$\pi$	pi
$\Gamma$	Gamma
$\lambda$	lambda
$\Omega$	Omega
$\varepsilon$	Epsilon
$\mu$	$10^{-6}$
$\gamma$	gamma
$\phi$	phi
$\theta$	theta
log	Logaritmo
$\in$	Relação de Inclusão
$\mathbb{E}$	Média Esperada
$\mathcal{L}$	Função de Perda
$\nabla$	Gradiente em Relação aos Parâmetros
$i$	i-ésimo elemento
$\tau$	Intervalo de Tempo Discreto
$a_t$	Ação no tempo $t$
$s_t$	Estado no Tempo $t$
$t$	Tempo
$\mathcal{Q}$	Valor Estimado
$s$	Estado Atual
$a$	Ação tomada no Estado
$\mathcal{D}$	Buffer de armazenamento
$\sim$	Distribuição de Probabilidade
$\mathcal{A}$	Espaço de ações possíveis do agente

$\Sigma$	Somatório
$r_t(\theta)$	Razão de Probabilidade
$G$	Ambiente Multiagente
$P$	Probabilidade Condicional entre Variáveis
$R$	Função de Recompensa
$H(p)$	Entropia da Distribuição de Probabilidade $p$
$N$	Número Total de Palavras

## SUMÁRIO

<b>1</b>	<b>CAPÍTULO I: INTRODUÇÃO</b>	<b>18</b>
<b>1.1</b>	<b>Contextualização</b>	<b>18</b>
<i>1.1.1</i>	<i>Motivação</i>	<i>22</i>
<i>1.1.2</i>	<i>Objetivos</i>	<i>22</i>
<i>1.1.3</i>	<i>Organização da Tese</i>	<i>23</i>
<b>2</b>	<b>FUNDAMENTAÇÃO TEÓRICA</b>	<b>25</b>
<b>2.1</b>	<b>Processamento de Linguagem Natural</b>	<b>25</b>
<b>2.2</b>	<b>Transformadores e mecanismos de atenção</b>	<b>30</b>
<i>2.2.1</i>	<i>O Mecanismo de Atenção: Scaled Dot-Product Attention</i>	<i>31</i>
<i>2.2.2</i>	<i>Multi-Head Attention: Múltiplas Perspectivas de Atenção</i>	<i>32</i>
<i>2.2.3</i>	<i>A Arquitetura Completa: Blocos de Encoder e Decoder</i>	<i>32</i>
<i>2.2.4</i>	<i>O Legado do Transformer</i>	<i>34</i>
<b>2.3</b>	<b>Retrieval-Augmented Generation – RAG</b>	<b>34</b>
<i>2.3.1</i>	<i>Arquitetura</i>	<i>35</i>
<i>2.3.2</i>	<i>Inferência com Top-k Documentos</i>	<i>35</i>
<i>2.3.3</i>	<i>Treinamento</i>	<i>36</i>
<i>2.3.4</i>	<i>RAG Multimodal</i>	<i>36</i>
<i>2.3.5</i>	<i>Vantagens e Desafios</i>	<i>37</i>
<b>2.4</b>	<b>Sistemas Multi Agentes - SMA e Princípios de Agentes Autônomos</b>	<b>37</b>
<b>2.5</b>	<b>Aprendizado por Reforço Multiagente (<i>Multi-Agent Reinforcement Learning</i> – MARL)</b>	<b>40</b>
<i>2.5.1</i>	<i>Formalização: O Processo de Decisão de Markov Descentralizado e Parcialmente Observável (Dec-POMDP)</i>	<i>41</i>
<i>2.5.2</i>	<i>Desafios Fundamentais em MARL</i>	<i>42</i>
<b>2.6</b>	<b>Multi-Agent Deep Deterministic Policy Gradient - MADDPG</b>	<b>43</b>
<i>2.6.1</i>	<i>Arquitetura: Atores Descentralizados e Críticos Centralizados</i>	<i>43</i>
<i>2.6.2</i>	<i>Treinamento Centralizado</i>	<i>44</i>
<i>2.6.3</i>	<i>Estabilidade do Treinamento com Redes-Alvo</i>	<i>45</i>
<b>2.7</b>	<b>Proximal Policy Optimization — PPO</b>	<b>46</b>
<i>2.7.1</i>	<i>Objetivo de clipping</i>	<i>46</i>

2.7.2	<i>Extensões multi-agentes</i> . . . . .	46
2.7.3	<i>Considerações práticas</i> . . . . .	46
2.8	<b>Direct Preference Optimization - DPO</b> . . . . .	47
2.8.1	<i>Limitações do Pipeline Clássico de RLHF</i> . . . . .	47
2.8.2	<i>A Formulação do DPO: De RL para Otimização Direta</i> . . . . .	48
2.8.3	<i>O Objetivo de Otimização do DPO</i> . . . . .	48
2.9	<b>Group Relative Policy Optimization - GRPO</b> . . . . .	49
2.9.1	<i>A Limitação dos Dados Pares</i> . . . . .	49
2.9.2	<i>Generalizando Preferências com o Modelo Plackett-Luce</i> . . . . .	49
2.9.3	<i>O Objetivo de Otimização do GRPO</i> . . . . .	50
3	<b>TRABALHOS RELACIONADOS</b> . . . . .	51
3.1	<b>Diagnóstico precoce de DP usando sistemas multiagentes autônomos baseados em modelos de linguagem</b> . . . . .	52
3.2	<b>Abordagens de Aprendizado de Máquina e Aprendizado Profundo para o diagnóstico DP</b> . . . . .	53
3.3	<b>Plataforma de Internet das Coisas Médicas</b> . . . . .	55
4	<b>METODOLOGIA</b> . . . . .	57
4.1	<b>Arquitetura</b> . . . . .	57
4.1.1	<i>Serviços</i> . . . . .	57
4.1.1.1	<i>Ingestão de Dados e Gateway de Entrada</i> . . . . .	59
4.1.1.2	<i>Pipeline de Processamento e Armazenamento (DataOps)</i> . . . . .	59
4.1.1.3	<i>Ciclo de Vida de Modelos de IA (LLMOps/MLOps/VLMOps)</i> . . . . .	60
4.1.2	<i>DataOps</i> . . . . .	61
4.1.3	<i>MLOps/LLMOps/VLMOps</i> . . . . .	61
4.1.3.1	<i>Latência Clínica e Eficiência Computacional</i> . . . . .	63
4.1.4	<i>Interface gráfica</i> . . . . .	64
4.1.4.1	<i>Frontend</i> . . . . .	64
4.1.4.2	<i>Backend</i> . . . . .	67
4.2	<b>Retrieval Augmented Generation (RAG)</b> . . . . .	67
4.2.1	<i>RAG Multimodal</i> . . . . .	68
4.2.1.1	<i>Embedding Multimodal</i> . . . . .	68
4.2.1.2	<i>Busca de Resumos</i> . . . . .	68

4.2.1.3	<i>Recuperação Híbrida</i>	68
4.2.2	<b>Graph RAG</b>	69
4.2.3	<b>AgenticRAG</b>	69
4.3	<b>Multi-Collaborative Agent Systems (MCAS)</b>	71
4.4	<b>Tomada de decisão clínica</b>	72
4.4.1	<i>Ação</i>	74
4.4.2	<i>Responsabilidades dos agentes</i>	76
4.5	<b>Métricas de avaliação</b>	76
4.5.1	<i>Avaliação do Classificador</i>	76
4.5.2	<i>Avaliação do Retrieval-Augmented Generation (RAG)</i>	77
4.5.2.1	<i>Avaliação da Fase de Recuperação no Retrieval-Augmented Generation (RAG)</i>	78
4.5.2.2	<i>Métricas de Geração</i>	78
4.5.2.3	<i>LLM-as-a-Judge</i>	80
4.5.3	<i>Avaliação do Sistema</i>	81
4.6	<b>Descrição do dataset</b>	82
4.6.1	<i>Pré-processamento</i>	87
4.6.2	<i>Processamento</i>	88
5	<b>RESULTADOS E DISCUSSÃO</b>	90
5.1	<b>Análise Estatística</b>	90
5.1.1	<i>audio</i>	90
5.1.2	<i>imagem</i>	92
5.2	<b>Análise do desempenho dos Agentes</b>	93
5.2.1	<i>individualmente</i>	93
5.2.1.1	<i>Sensorial</i>	94
5.2.1.2	<i>Áudio</i>	97
5.2.1.3	<i>Imagens</i>	100
5.2.2	<i>Análise Comparativa das Abordagens de Treinamento: Ajuste Fino vs. Aprendizado por Transferência</i>	104
6	<b>CONCLUSÃO</b>	106
6.0.1	<i>Discussão, Oportunidades e Questões em aberto</i>	107
6.1	<b>Trabalhos Futuros</b>	108
	<b>REFERÊNCIAS</b>	109

# 1 CAPÍTULO I: INTRODUÇÃO

Nesse capítulo introdutório, serão apresentadas as informações, como, a sua contextualização, justificativa e problema de pesquisa incorporando diretamente sua abordagem metodológica, que servirão para entender todo o contexto e a abordagem do trabalho. O estudo diante de um contexto geral, tem grande importância e relevância na área médica, precisamente em uma análise multivariada para diagnóstico precoce da Doença de Parkinson (DP). Além disso, o trabalho lista objetivos e contribuições a serem contemplados, explicando claramente o propósito do estudo.

## 1.1 Contextualização

As doenças neurológicas degenerativas representam um dos grandes desafios atualmente à saúde pública [Lima *et al.* 2024]. Caracterizam-se pela perda progressiva e irreversível da função neuronal, resultando em comprometimentos físicos e cognitivos significativos, com impacto expressivo na qualidade de vida dos indivíduos com essas enfermidades neurológicas [Duque *et al.* 2023]. Entre as principais enfermidades neurológicas, mundialmente, destacam-se a DP, a Doença de Alzheimer (DA) e a Esclerose Múltipla (EM), que afetam milhões de pessoas no mundo todo [Zhang *et al.* 2023]. A prevalência dessas condições tende a crescer de forma proporcional ao envelhecimento populacional, o que evidencia a necessidade premente de investigar seu impacto sobre o sistema de saúde brasileiro [Assis *et al.* 2024].

A DP é a segunda condição neurológica mais comum no mundo, ficando atrás apenas da DA [Dong-Chen *et al.* 2023]; [Reddy *et al.* 2024]. É estimado que, em 2025, mais de 8,5 milhões de pessoas sofram com a DP, em escala global, com uma projeção de aumento de 50% até 2040, impulsionado pelo envelhecimento populacional, segundo o estudo *Global Burden of Disease* [GBD Parkinson's Disease Collaborators 2024]. No Brasil, cerca de 200 mil pessoas vivem com a DP e estima-se que a população com mais de 50 anos mostrou uma prevalência de 0,8%, correspondendo a mais de 500.000 brasileiros nessa faixa etária com diagnóstico da DP; [BRASIL. Ministério da Saúde 2023].

A manifestação clínica da DP abrange tanto sintomas motores, como, por exemplo, lentidão de movimentos, como, bradicinesia, rigidez e tremor em repouso, como também, não motores, incluindo alterações cognitivas, distúrbios do sono, sintomas depressivos e disfunções autonômicas [Skaramagkas *et al.* 2023]; [Mai *et al.* 2024]. Os sintomas não motores, que podem

se manifestar muitos anos antes do diagnóstico clínico tradicional e esta indicação não-motor pode ser diagnosticada anos após o diagnóstico inicial, devido a sua natureza incerta e a falta de registros suficientes que dificultam bastante a identificação precoce da doença [Piloto *et al.* 2024].

Clinicamente, a DP é determinada pela presença de bradicinesia e pelo menos um sintoma motor cardinal, como, rigidez ou tremor em repouso, em combinação com outras características de suporte e exclusão [Hendriks *et al.* 2024]. Expressão facial reduzida, diminuição da caligrafia (micrografia), problemas na fala e na voz, bem como dificuldade de engolir, representam aspectos motores adicionais da DP. Estes sintomas motores interferem significativamente nas atividades diárias dos pacientes, prejudicando sua qualidade de vida [Lucksch *et al.* 2023]. A progressão da disfunção motora pode aumentar anormalidades como na marcha e com o aumento das anormalidades na marcha e afetando a instabilidade postural, que interferem ainda mais a autonomia e a segurança dos pacientes.

incluem alterações de humor e de estado mental, como depressão, apatia (ou falta de motivação), além de comprometimento progressivo da função cognitiva

Embora os sintomas motores definam a síndrome clínica, a maioria dos pacientes com DP reclamam de outras sequelas que foram classificadas como não motoras, provavelmente relacionadas a vias não-dopaminérgicas. Estas incluem alterações de humor e de estado mental, como depressão, apatia, ou falta de motivação, além de comprometimento progressivo da função cognitiva [Kinger *et al.* 2023]. Fadiga, distúrbios do sono, disfunções autonômicas, bem como queixas sensoriais são componentes comuns do espectro clínico da DP [Lauretani *et al.* 2023]. Sintomas não motores estão presentes nos estágios iniciais da doença, enquanto alguns antecedem o início dos principais sintomas motores por anos ou até décadas [Li *et al.* 2024]. Porém, esses sintomas vão ficando mais comuns e acabam influenciando bastante na qualidade de vida, na piora da incapacidade geral e até na mortalidade, conforme a doença vai avançando [Filippi *et al.* 2023].

Atualmente, o diagnóstico da DP permanece fundamentalmente clínico e, por consequência, pode depender da opinião de cada profissional, variar entre especialistas e ter dificuldades para realizar exames adicionais. Portanto, pesquisas por soluções tecnológicas possibilitam diagnósticos mais sensíveis, precoces e automatizados tem se intensificado [Sun *et al.* 2024]. A Inteligência Artificial (IA) já transformou a área da saúde e ainda tem muito potencial para novos avanços [Xia *et al.* 2025]; [Liu *et al.* 2025]. Assim, a Engenharia Biomédica, especialmente na

convergência com a IA, emerge como uma aliada no processo ao propor sistemas diagnósticos baseados em dados reais, processados de maneira automática e que podem ser atualizados constantemente [Gao *et al.* 2025].

Neste contexto, esta tese propõe uma plataforma de apoio ao diagnóstico da DP baseada na integração de arquiteturas modernas, como Operações nos Dados (*DataOps*), Operações com Modelos de Linguagem (*LLMOps*) e, mais recentemente, *Vision Language Models Operations (VLMOps)*. Essa estrutura acompanha os avanços recentes dos modelos de linguagem de última geração que operam de forma multimodal, lidando não apenas com texto, mas também com imagens médicas, vídeos e sinais sensoriais. A metodologia proposta envolve o uso intensivo de Modelos de Linguagem de Larga Escala (LLMs), técnicas de compressão como quantização 4-bit e *Quantized Low-Rank Adaptation (QLoRA)* [Zhou *et al.* 2025], além de estratégias robustas de recuperação e geração de conhecimento multimodal [Wang *et al.* 2025].

O pipeline metodológico contempla desde a coleta de dados clínicos e sensoriais por meio de dispositivos conectados via *Internet of Medical Things (IoMT)*, passando pelo seu pré-processamento e validação contínua em pipelines de *DataOps*, até sua utilização por agentes inteligentes colaborativos, os *Collaborative Intelligent Agents (MCATs)*. Esses agentes operam com abordagens de *Retrieval Augmented Generation (RAG)*, incluindo variantes como *Multimodal RAG*, *GraphRAG* e *Agentic RAG* [Yang *et al.* 2025, Zhu *et al.* 2024], otimizando a recuperação de informações clínicas e a precisão diagnóstica.

Adicionalmente, técnicas de RL são aplicadas na etapa de pós-treinamento, possibilitando que os agentes ajustem suas estratégias de forma adaptativa, com base em recompensas relacionadas a métricas clínicas ou consenso entre múltiplos agentes [Al-Hamadani *et al.*, 2024].

Portanto a aplicação de modelos de compressão se torna inevitável para otimizar a eficiência desses sistemas. Técnicas como quantização 4-Bit e QLoRA [Dettmers *et al.* 2023, Zhou *et al.* 2025] foram empregadas no uso tornar os SLMs/LLMs ainda mais otimizados, para garantir que os sistemas possam ser implantados em ambientes com recursos computacionais limitados [Li *et al.* 2024]; [Goldbarg 2024], viabilizando sua adoção prática em contextos médicos reais. As decisões diagnósticas são tomadas de forma colaborativa por múltiplos agentes especializados, utilizando mecanismos de votação e discussão entre agentes (*MCAS*) para maximizar a confiabilidade e interpretabilidade das recomendações médicas [Kim *et al.* 2024].

Para integrar todas as técnicas, sistemas e modelos, foi proposto neste estudo o

desenvolvimento de uma plataforma médica baseada em Internet das Coisas, com comunicação segura. Essa plataforma integra pipelines de DataOps para coleta, transformação, armazenamento e validação de dados multimodais e LLMOps, para treinamento, deploy e monitoramento de modelos de linguagem, [Su *et al.* 2025], unidas a um sistema agêntico colaborativo baseado em RAG e pós-treinamento com algoritmos de ML. São suportadas diversas modalidades de dados tabulares, como sensores (CSV, JSON, XML), texto, áudio (MP3, WAV) e imagens médicas (DICOM, PNG, JPEG) [Ghayvat *et al.* 2024].

Em hospitais gerais e centros especializados em neurologia, a plataforma pode ser integrada aos fluxos de trabalho clínico existentes por meio da interoperabilidade com padrões como *Health Level Seven – Fast Healthcare Interoperability Resources* (HL7-FHIR) e *Digital Imaging and Communications in Medicine* (DICOM) [Bernardi 2024], permitindo que profissionais como neurologistas, geriatras, fisiatras e fonoaudiólogos acessem análises automatizadas e *dashboards* inteligentes diretamente a partir dos prontuários eletrônicos [Machorro-Cano *et al.* 2023]. Através de agentes multimodais treinados para lidar com dados clínicos, a plataforma pode auxiliar na triagem automatizada de pacientes, na avaliação de sintomas, e no apoio à decisão terapêutica, oferecendo recomendações baseadas em evidências e dados históricos. Além disso, sua arquitetura modular permite funcionamento híbrido em ambientes com ou sem conexão contínua à nuvem, facilitando a adoção em hospitais com infraestrutura heterogênea [Ehtesham *et al.* 2025].

Em cenários de monitoramento domiciliar, como programas de reabilitação ou cuidados continuados, a plataforma possibilita que cuidadores, fisioterapeutas e familiares coletem sinais de marcha, voz e atividades cotidianas por meio de sensores vestíveis e dispositivos móveis conectados via IoMT [Sarkar *et al.* 2024]. Essa abordagem oferece maior autonomia, acompanhamento personalizado e detecção precoce de agravamentos, mesmo fora do ambiente clínico. Ao promover um ciclo contínuo de aquisição de dados, inferência e recomendação explicável, a plataforma busca preencher lacunas críticas na gestão integrada e preventiva da DP.

As principais contribuições deste trabalho incluem:

- A identificação de limitações significativas na utilização de dados de áudio para o diagnóstico de DP com as abordagens testadas, fornecendo *insights* significantes para o direcionamento de pesquisas futuras nesta área.
- A estruturação de uma base para um sistema de apoio à decisão clínica que, ao integrar informações de múltiplas modalidades, visa aumentar a assertividade e a abrangência do

diagnóstico da DP.

### **1.1.1 Motivação**

A DP sendo a segunda condição neurológica mais prevalente no mundo, afeta milhões de indivíduos com impacto significativo na autonomia e qualidade de vida. Clinicamente, a DP manifesta-se principalmente por sintomas motores (como tremores, rigidez e bradicinesia), mas é cada vez mais reconhecida pelos sintomas não motores, incluindo distúrbios do humor, alterações cognitivas e disfunções autonômicas, que muitas vezes precedem os sintomas motores em anos ou até décadas.

Essa complexidade clínica e temporal impõe grandes desafios ao diagnóstico precoce e manejo efetivo da DP. Os métodos tradicionais, como exames neurológicos e de neuroimagem, enfrentam limitações práticas: são caros, demandam especialistas e falham em capturar alterações sutis nas fases iniciais. Tais barreiras reduzem o acesso, atrasam intervenções precoces e comprometem a qualidade da atenção ao paciente.

Nesse cenário, abordagens computacionais inteligentes tornam-se promissoras, sobretudo aquelas capazes de integrar sinais clínicos heterogêneos, incluindo dados textuais, como prontuários, imagens médicas, sinais de voz e sensores de movimento, sob uma perspectiva multimodal. No entanto, a maioria das soluções existentes ainda apresenta lacunas: não dialogam bem com os fluxos clínicos reais, carecem de interpretabilidade, ou não estão prontas para ambientes com restrições de conectividade, latência ou infraestrutura.

Este trabalho visa a contribuir com esse panorama, por meio do desenvolvimento de uma plataforma diagnóstica baseada em inteligência artificial multimodal, alicerçada em agentes colaborativos, modelos de linguagem de última geração e tecnologias IoMT. O objetivo é construir uma solução escalável, explicável, integrável a fluxos clínicos reais, como aqueles baseados em HL7-FHIR e DICOM, e adaptada a ambientes hospitalares ou domiciliares conectados. Ao unir inovação técnica com aplicabilidade clínica concreta, este estudo busca transformar a forma como a DP é detectada, monitorada e gerenciada.

### **1.1.2 Objetivos**

O objetivo principal deste trabalho é desenvolver e implementar uma abordagem inovadora baseada em uma multi colaboração de agentes para aprimorar a detecção e o diagnóstico precoce da DP, utilizando métodos avançados como Agentes Autônomos, LLMs, SLMs e RAG.

Os objetivos específicos deste estudo são:

- Realizar uma análise comparativa dos métodos RAG, destacando a eficácia de abordagens como: multimodal RAG para integrar dados de diversas naturezas e GraphRAG para explorar relações entre eles através de grafos;
- Desenvolver e avaliar o desempenho da nossa plataforma, construída em Vue.js para o frontend e FastApi para o backend, baseada em um sistema agentivo autônomo, capaz de interagir colaborativamente para a tomada de decisão clínica;
- Validar dados na etapa de pré-processamento e processamentos dos dados, para cada modalidade pelo menos um agente é responsável por compreender aquela modalidade de dado específica e gerar uma classificação entre paciente saudável ou com DP;
- Propor uma estratégia colaborativa customizada e gerenciada para DP, além de comparar e validar com as existentes, como: votação majoritária, unanime, discussão e consenso , entre outras;
- Analisar o impacto dos modelos unimodais, bimodais e multimodais no desempenho diagnóstico, identificando a configuração mais eficiente e precisa para a integração de dados médicos;
- Implementar e avaliar técnicas de compressão de modelos de linguagem (LLMs/SLMs), com ênfase em: quantização (8bits) para reduzir o consumo de recursos computacionais e viabilizar diagnósticos em tempo real, quantização e QLoRA (4bits), especialmente em ambientes com restrição de hardware;
- Integrar os métodos desenvolvidos na plataforma IoMT, de maneira, eficiente e segura, visando, garantir acompanhamento seguro e periódico do paciente, disponibilizando informações de paciente para facilitar a utilização em cenários clínicos reais.

### ***1.1.3 Organização da Tese***

Esta tese foi estruturada em seis capítulos. O Capítulo 1, sobre a introdução, tem a responsabilidade de contextualizar a pesquisa ao leitor, apresentando uma visão geral do assunto, sua importância e relevância. A introdução expõe claramente a problemática abordada e os objetivos da do trabalho, definindo claramente a estrutura da pesquisa.

O capítulo 2 trata sobre os fundamentos teóricos, estes essenciais para o desenvolvimento da plataforma IoMT baseados em pequenos modelos de linguagem propostos nesta tese. Nesta etapa, detalha-se o surgimento do uso de agentes, modelos de linguagem e abordagens

*onemodal, bimodal, trimodal e multimodal*. Na sequência, são abordados aspectos técnicos fundamentais para entendimento desta tese,

analisando métodos de diagnóstico precoce e prevenção da DP através dos dados mockados multimodais, que incorpora sintomas motores e não motores, exames médicos realizado por imagens, patologias hereditárias, entre outros.

O capítulo 3 propõe tratar sobre os trabalhos relacionados. Nessa seção, são discutidas inúmeras metodologias usadas em pesquisas que foram identificadas importantes para evidenciar a solução proposta nesse trabalho como única, eficaz e inovadora para o diagnóstico precoce preciso de DP, fundamental para o desenvolvimento desse trabalho.

O Capítulo 4 apresenta os métodos e procedimentos empregados nesta pesquisa. A seção descreve detalhadamente a abordagem adotada, as técnicas de coleta de dados, os instrumentos utilizados e as etapas de análise. A metodologia é exposta de maneira clara e objetiva, garantindo que o estudo possa ser replicado por outros pesquisadores.

No Capítulo 5, na seção dos resultados, são apresentados de forma estruturada os principais achados da pesquisa, analisando-os em relação aos objetivos estabelecidos. Tabelas, gráficos e outros recursos visuais são empregados para ilustrar e facilitar a interpretação dos dados coletados. Este capítulo destaca a relevância dos resultados obtidos, contextualizando-os no escopo da pesquisa e enfatizando sua contribuição para o tema estudado.

Finalmente, o Capítulo 6 trata da conclusão, apresentando todos os resultados obtidos e as conclusões da pesquisa. Analisam-se criticamente os frutos da pesquisa, considerando os objetivos apresentados e a unicidade inovadora do trabalho em detrimento das literaturas. A conclusão apresenta, por fim, as contribuições do estudo para o campo da pesquisa, discute possíveis limitações e sugere pesquisas adicionais.

## 2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, serão apresentados os principais fundamentos teóricos que embasam o desenvolvimento de um ecossistema IoT inteligente, orientado por agentes autônomos, para a análise da DP. A fundamentação contempla tópicos essenciais como Processamento de Linguagem Natural (PNL), mecanismos de atenção e transformadores, com foco em sua evolução até abordagens como RAG; princípios de agentes autônomos e sistemas multiagentes, incluindo técnicas de aprendizado por reforço em ambientes multiagentes (*Multi-Agent Reinforcement Learning* (MARL), *Multi-Agent Deep Deterministic Policy Gradient* (MADDPG), *Proximal Policy Optimization* (PPO)); bem como conceitos relacionados à IoMT. Essa base teórica é fundamental para contextualizar a proposta, alinhar os componentes técnicos ao domínio da saúde digital e justificar as decisões metodológicas adotadas ao longo do trabalho.

### 2.1 Processamento de Linguagem Natural

O PNL ou, do termo em inglês *Natural Language Processing* (NLP), marca a interseção entre IA e a linguística. Este marco exteriorizou publicamente ao mundo em 8 de janeiro de 1954 sob a manchete "Russian is turned into English by a fast electronic translator" na primeira página do jornal *New York Times* através do experimento de Georgetown-IBM [Hutchins 2004]. O experimento consistiu em uma translação automática de mais de sessenta frases russas para o inglês usando o computador IBM 701.

Ainda na década de 50, outro importante marco no campo que integra a inteligência artificial e a linguística foi a introdução do conceito de gramática gerativa, proposto por Noam Chomsky em sua obra pioneira "Syntactic Structures" [Chomsky 2002], a qual ajudou significativamente pesquisadores a terem uma base teórica para compreender a estrutura da linguagem natural e como poderiam avançar as pesquisas de tradução automática realizada por máquina. A ideia por trás da gramática gerativo-transformacional de Chomsky é propor uma teoria que pode ser fundamentalmente estruturada e explicada através dos principais pressupostos propalados: estruturas superficiais e profundas, gramática gerativa-transformacional, gramática universal, aquisição da língua e teoria dos princípios e parâmetros.

De acordo com Chomsky, a estrutura profunda é a forma subjacente, ou seja, abstrata, comum a todas as línguas humanas, que determina o significado base de uma frase, enquanto a estrutura superficial é a manifestação aparente, concreta e visível da língua que varia de uma

língua para outra.

A gramática gerativa-transformacional, ao contrário da gramática tradicional, tenta explicar como a língua é gerada na mente do ser humano. Parte-se de um número finito de regras e princípios para gerar um número finito de frases. As frases são geradas na estrutura profunda da língua e passam por um processo de transformação até chegar à superfície, e essas modificações incluem modificações na ordem das palavras, acréscimos e remoção de segmentos, ajustes na sintaxe da frase, entre outras modificações.

A língua é um conjunto de estruturas que está presente em todas as línguas humanas. São princípios linguísticos inatos compartilhados por todos os seres humanos, explicitando a gramática universal. Chomsky denomina a aquisição da língua como efeito Platão; isto quer dizer que a complexidade da língua vai além dos dados linguísticos aos quais o ser humano é exposto.

Em seu último pressuposto, a teoria dos princípios e parâmetros, ele afirma que os princípios compõem as regras compartilhadas por todas as línguas humanas, ou seja, a gramática universal, e se constituem a partir do dispositivo da aquisição da língua, recurso inato e presente em todos os falantes de qualquer língua. Já os parâmetros compõem aqueles elementos da gramática que podem variar de uma língua para outra, como as ordens das palavras em uma frase, concordância verbal, etc.

Com a euforia gerada por sistemas de tradução automática, na década de 60, o ALPAC, Automatic Language Processing Advisory Committee, comitê criado em 1966 pelo governo dos Estados Unidos para avaliar o progresso da linguística computacional, divulgou um relatório que recomendava a interrupção das pesquisas sobre tradução automática, o que teve um impacto significativo nas pesquisas sobre PNL e IA de forma mais ampla.

Em 1980 e 1990 foram os anos dos sistemas especialistas e a introdução de modelos estatísticos. Os sistemas especialistas utilizavam abordagens simbólicas, com gramáticas formais, como gramáticas livres de contexto, regras de produção codificadas por linguistas ou especialistas e ontologias codificadas, estruturas de conhecimento que armazenavam fatos, conceitos e relações entre eles, organizadas em domínios específicos, como por exemplo: tradução automática baseada em regras, análise sintática (parsing) com base em árvores gramaticais e representações semânticas lógicas, como a lógica de predicados. Porém, no final dos anos 1980 e no começo dos anos 1990, os modelos estatísticos pareceram ser uma alternativa viável para superar os problemas presentes na abordagem simbólica. Diferentemente da abordagem simbólica que

dependia da definição teórica manual das regras e de difícil escalabilidade, os modelos estatísticos surgem com o diferencial de serem capazes de “aprender sozinhos”, por meio do aprendizado de máquina, ou do termo em inglês *Machine Learning* (ML), as inúmeras regras codificadas dos sistemas especializados. Essa alternativa se popularizou, sobretudo, pelo aumento dos recursos computacionais.

Ao longo das últimas décadas, os recursos computacionais evoluíram significativamente, tornando viável o uso de grandes volumes de dados no treinamento de modelos estatísticos. Esses modelos, por sua vez, tornaram-se cada vez mais eficientes e populares, especialmente a partir dos anos 2000. Na década de 2000, as redes neurais começaram a ser amplamente utilizadas no contexto do PLN, inicialmente com o objetivo de modelar a distribuição de palavras em uma língua e prever termos subsequentes em um texto a partir do contexto anterior, tarefa conhecida como modelagem de linguagem.

Ainda nesse período, surgiu o conceito de incorporação de palavras (word embeddings), que consiste em representar palavras da língua como vetores densos em um espaço contínuo, de forma que termos semanticamente semelhantes sejam mapeados para regiões próximas nesse espaço vetorial. Entretanto, os primeiros algoritmos enfrentavam dificuldades para aprender essas representações de maneira eficiente, o que limitava a qualidade dos vetores gerados, devido à escassez de dados ou à baixa capacidade dos modelos.

Um dos primeiros sistemas de PLN com aplicação comercial bem-sucedida foi o Google Tradutor, lançado em 2006, que utilizava modelos estatísticos para realizar traduções automáticas entre diferentes idiomas. Um avanço importante ocorreu em 2013, com a publicação do algoritmo Word2Vec, descrito no artigo “Efficient Estimation of Word Representations in Vector Space” [Mikolov *et al.* 2013]. Esse algoritmo foi pioneiro na aprendizagem eficiente de incorporações de palavras, elevando substancialmente a qualidade dos vetores em comparação com abordagens anteriores. Observou-se, por exemplo, que os vetores aprendidos pelos dois modelos, que faziam a representação de palavras no espaço vetorial, propostos pelo autor, o *Continuous Bag of Words* (CBOW) ou saco de palavras contínuas e o Skip-Gram, possibilitavam operações semânticas por meio de álgebra vetorial, o que possibilitava a captura de relações semânticas latentes.

Segundo o autor da pesquisa, o modelo CBOW apresenta semelhanças estruturais com a arquitetura de rede neural proposta por [Bengio *et al.* 2003], conhecida como *Neural Network Language Model* (NNLM). A estrutura do CBOW consiste em três camadas principais:

uma camada de entrada, que recebe as palavras do contexto, ou seja, frases anteriores e posteriores à palavra-alvo, uma camada de projeção, que converte essas palavras em vetores que capturam seu significado em formato numérico e uma camada de saída, que prediz a palavra central com base nos vetores combinados do contexto.

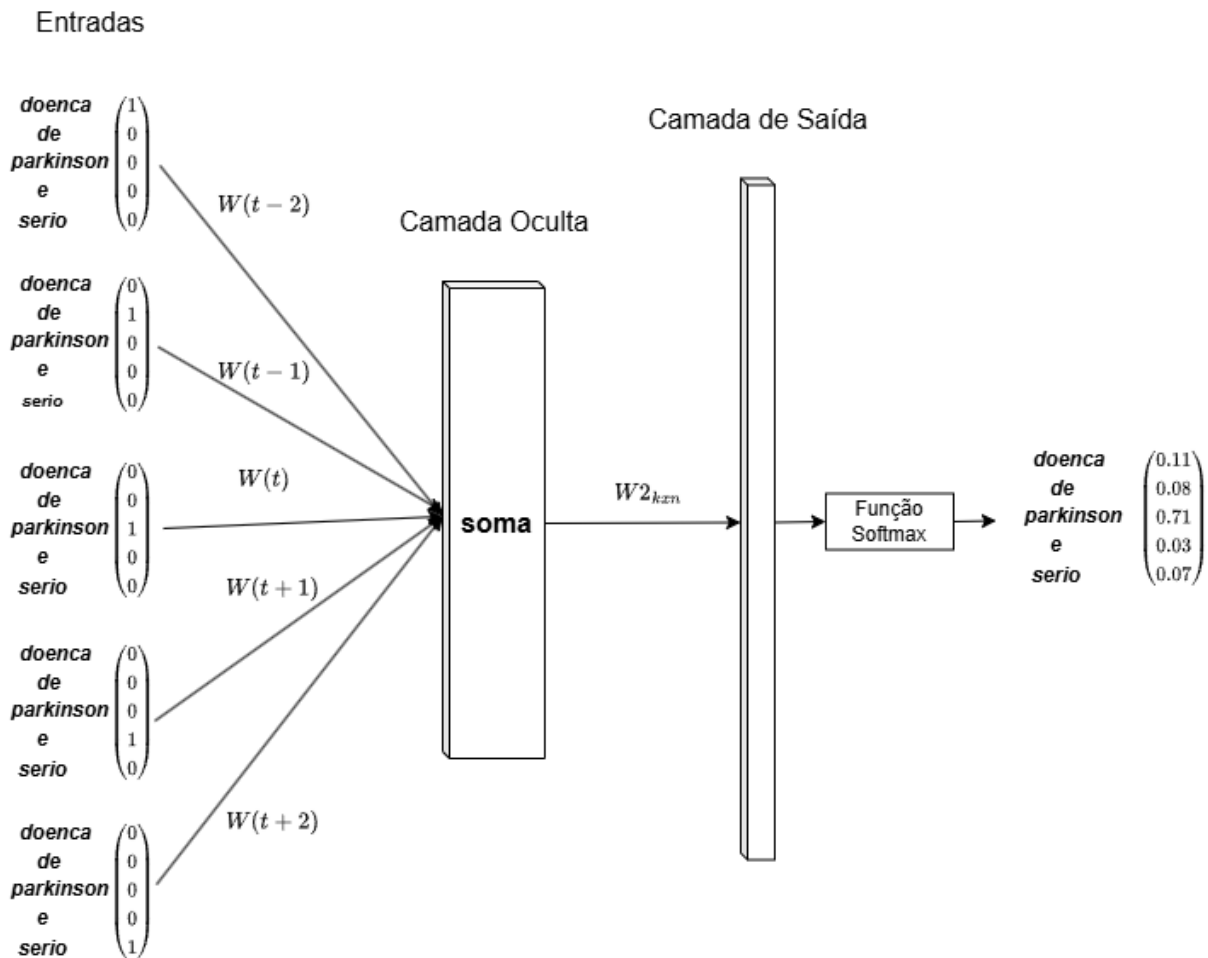
Diferentemente do NNLM, o CBOW remove a camada oculta não linear e compartilha os pesos da projeção entre todas as palavras do contexto, o que reduz significativamente o custo computacional. Os vetores de todas as palavras do contexto são agrupados, somados ou em média, projetando-os em uma mesma posição vetorial, como ilustrado na Figura 1, em que cada vetor de entrada passa por um mesmo conjunto de pesos  $W(t)$ , sendo somado antes de ser alimentado na camada de saída. Essa estrutura permite capturar de forma eficiente a representação distribuída do contexto da palavra-alvo, ainda que sem considerar a ordem das palavras [Mikolov *et al.* 2013].

No final, o modelo usa a Softmax para gerar uma lista de probabilidades e escolher a palavra mais provável para o meio da frase. Esse processo é comum em modelos que treinam vetores de palavras, como, o Word2Vec.

Além disso, o segundo modelo proposto pelo autor foi o Skip-gram, que também representa palavras em um espaço vetorial e sua arquitetura é bem semelhante à do CBOW, conforme mostra na figura 2, mas ele realiza o processo inverso. Em vez do modelo prever uma palavra baseada no contexto, ela tenta maximizar a classificação de uma palavra com base em outra palavra na mesma frase, ou seja, cada palavra atual é dada como entrada para um classificador com uma camada de projeção contínua e as palavras são previstas dentro de um intervalo, antes e depois da palavra atual, formando o contexto [Mikolov *et al.* 2013].

Popularmente esses dois modelos em seguida deram luz ao termo “embeddings”, que são as representações vetoriais no espaço de palavras da língua. Então os embeddings pré-treinados acabaram se tornando uma etapa fundamental em muitos sistemas de PLN, uma vez que seu uso demonstrou melhorar substancialmente o desempenho em diversas tarefas downstream. Desde então, esforços têm sido dedicados à obtenção de incorporações que representem de maneira mais precisa o significado de palavras e sentenças. Durante esse período, destacaram-se arquiteturas de redes neurais recorrentes, como as *Long Short-Term Memory* (LSTMs) e Redes Neurais Convolucionais (CNNs) aplicadas ao texto.

Em 2014, foi proposta uma nova formalização para tarefas de sequência para sequência, ou do inglês, sequence-to-sequence; como a tradução automática, por meio de uma arquite-

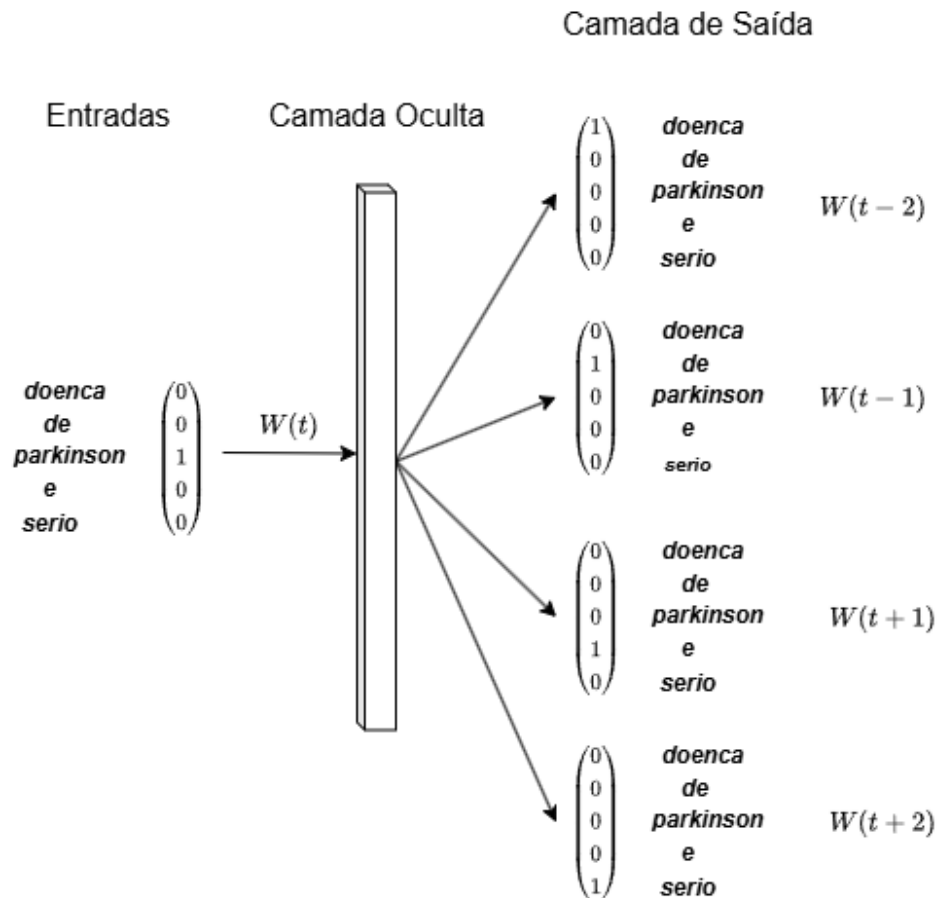
Figura 1 – Arquitetura do modelo CBOW (*Continuous Bag of Words*)

Fonte: Próprio autor, adaptado de [Mikolov *et al.* 2013]

tura baseada em dois módulos de redes neurais: o Codificador-Decodificador, Encoder-Decoder. O codificador tem como função transformar a sequência de entrada em uma representação vetorial densa, frequentemente denominada vetor de contexto, enquanto o decodificador é responsável por gerar a sequência de saída com base nessa representação.

Posteriormente, mecanismos de atenção foram incorporados a essa arquitetura, permitindo que o modelo aprendesse a focar seletivamente em partes relevantes da entrada durante a geração da saída. Essa modificação aumentou significativamente a capacidade dos modelos sequence-to-sequence, a ponto do Google Translate, em 2016, substituir completamente seus modelos estatísticos por redes neurais baseadas nessa abordagem.

Finalmente, em 2017, Vaswani introduziu a arquitetura de Transformadores, ou Transformers, com o influente artigo “Attention Is All You Need” [Vaswani *et al.* 2017]. A seguir serão discutidos com detalhes a arquitetura dos Transformers e os mecanismos de atenção que revolucionaram o modo que conhecíamos até então o processamento de linguagem natural.

Figura 2 – Arquitetura do modelo *Skip-Grim*

Fonte: Próprio autor, adaptado de [Mikolov *et al.* 2013]

## 2.2 Transformadores e mecanismos de atenção

A introdução da arquitetura Transformer, proposta por [Vaswani *et al.* 2017], representou uma mudança de paradigma no processamento de sequências, ao substituir estruturas recorrentes por mecanismos de atenção, promovendo ganhos expressivos em desempenho, paralelismo e escalabilidade. Sendo melhores fundamentalmente das *Recurrent Neural Networks* (RNNs) e CNNs que dominavam o campo do PLN até então. Modelos como LSTMs e GRUs, embora eficazes, processam dados sequencialmente, o que inerentemente limita a paralelização e cria desafios na captura de dependências de longo alcance devido ao caminho que a informação precisa percorrer ao longo da sequência [Hochreiter e Schmidhuber 1997]. O Transformer supera essas limitações ao dispensar a recorrência e adotar um mecanismo de atenção como seu principal bloco de construção.

Embora mantenha a estrutura de alto nível de um codificador-decodificador, encoder-decoder, sua operação interna é significativamente diferente, ao substituir os mecanismos recorrentes por atenção auto-regressiva, permitindo maior paralelismo e eficiência no processamento

de sequências. A seguir, detalharemos os componentes fundamentais que definem esta arquitetura.

### 2.2.1 O Mecanismo de Atenção: Scaled Dot-Product Attention

No núcleo do Transformer está o mecanismo de atenção, mais precisamente, o *Scaled Dot-Product Attention* (SDPA), que permite ao modelo focar nas partes mais relevantes da entrada. A intuição por trás da atenção é permitir que o modelo pondere a importância de diferentes palavras, ou tokens, na sequência de entrada ao processar uma palavra específica. Em vez de depender de um estado oculto que comprime toda a informação passada, a atenção permite acesso direto e ponderado a todas as posições da sequência.

A operação de atenção é baseada em três vetores principais: *Query* ( $Q$ ), *Key* ( $K$ ) e *Value* ( $V$ ), que representam, respectivamente, a busca, a referência e a informação a ser recuperada. Para cada *token* na sequência, geramos esses três vetores a partir de seu vetor de embedding de entrada.

- **Query (Q)**: Representa o token atual que está "buscando" informação.
- **Key (K)**: Representa todos os tokens na sequência que oferecem informação. A compatibilidade entre um *Query* e um *Key* determina o peso de atenção.
- **Value (V)**: Representa o conteúdo informacional de cada token.

O cálculo da atenção, ou o termo em inglês Scaled Dot-Product Attention, é realizado da seguinte forma:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2.1)$$

Para melhor compreensão, divide-se a equação mostrada acima em partes. Inicialmente,  $(QK^T)$  é o cálculo de similaridade, onde  $(K^T)$  é a matriz transposta de  $K$ , assim, o produto escalar entre o vetor *Query* de um token e os vetores *Key* de todos os outros tokens na sequência é calculado. O resultado é uma matriz de pontuações de similaridade. O escalonamento, dado por  $(\frac{1}{\sqrt{d_k}})$ , representa as pontuações que são escalonadas por um fator de  $\frac{1}{\sqrt{d_k}}$ , onde  $d_k$  é a dimensão dos vetores *Key* e *Query*. Este escalonamento é crucial para estabilizar o treinamento, evitando que os argumentos da função softmax se tornem excessivamente grandes, o que poderia levar a gradientes muito pequenos, saturação da softmax, [Vaswani *et al.* 2017]. A normalização, ou softmax, é a função aplicada às pontuações escalonadas para obter os pesos de atenção. Esses pesos são valores entre 0 e 1 que somam 1, funcionando como uma distribuição de

probabilidade que indica quanta atenção um token deve prestar a cada outro token. Por fim, o cálculo da saída ponderada ( $\dots V$ ), quando os vetores, “Value”, são multiplicados por esses pesos de atenção e somados, produzindo o vetor de saída para o token em questão. Este vetor de saída é uma representação do token que agora contém informações contextuais ponderadas de toda a sequência.

### 2.2.2 *Multi-Head Attention: Múltiplas Perspectivas de Atenção*

Em vez de calcular a atenção uma única vez, o Transformer aprimora esse processo por meio do mecanismo de *Multi-Head Attention* (MHA). Essa abordagem permite que o modelo aprenda diferentes tipos de relações contextuais simultaneamente, ao projetar os vetores *Query* ( $Q$ ), *Key* ( $K$ ) e *Value* ( $V$ ) em subespaços distintos. Cada "cabeça", ou seja, várias versões paralelas do cálculo de atenção, cada uma com seus próprios pesos de forma independente, e os resultados são posteriormente combinados. Com isso, o modelo é capaz de capturar múltiplas representações de dependência entre os tokens, enriquecendo sua capacidade de entender padrões complexos em sequências. Em vez de calcular a atenção uma única vez, o Transformer aprimora esse mecanismo com a MHA. Esta abordagem permite que o modelo aprenda diferentes tipos de relações contextuais simultaneamente. A ideia é projetar os vetores  $Q$ ,  $K$  e  $V$  em subespaços diferentes e aplicar o mecanismo de atenção em paralelo para cada "cabeça" (*head*).

O processo é o seguinte:

1. Os vetores  $Q$ ,  $K$  e  $V$  originais são linearmente projetados  $h$  vezes, onde  $h$  é o número de cabeças de atenção, com diferentes matrizes de pesos aprendidas.
2. O SDPA é aplicado em paralelo a cada uma dessas projeções, resultando em  $h$  vetores de saída.
3. Esses  $h$  vetores de saída são concatenados e passados por uma camada linear final para produzir o vetor de saída final do bloco de MHA. Isso permite que diferentes cabeças se especializem em capturar aspectos distintos da linguagem, como dependências sintáticas, relações de correferência ou associações semânticas.

### 2.2.3 *A Arquitetura Completa: Blocos de Encoder e Decoder*

O *Transformer* é composto por uma pilha de  $N$  codificadores idênticos e uma pilha de  $N$  decodificadores idênticos. Primeiramente, cada Bloco Codificador (*Encoder*) possui duas subcamadas principais:

**Multi-Head Self-Attention (MHSA):** O mecanismo de MHSA permite que o modelo processe múltiplas representações de atenção simultaneamente, onde os vetores Q, K e V vêm da mesma sequência de entrada, a saída da camada anterior do codificador. Isso permite que cada *token* na sequência de entrada atenda a todos os outros *token* na mesma sequência, construindo uma representação contextual rica.

**Feed-Forward Network (FFN):** A segunda camada é uma rede neural feed-forward totalmente conectada, aplicada a cada posição de forma idêntica e separada. Geralmente consiste em duas camadas lineares com uma função de ativação ReLU (ou GELU) entre elas. Além disso, cada uma dessas duas subcamadas possui uma conexão residual, residual connection, seguida por uma normalização de camada, layer normalization. As conexões residuais ajudam a evitar o problema do desaparecimento do gradiente em redes profundas, enquanto a normalização de camada estabiliza o treinamento.

Antes de entrar no primeiro codificador, os embeddings de entrada são somados a codificações posicionais, (positional encodings). Como o modelo não tem recorrência, ele não tem conhecimento inerente da ordem das palavras. As codificações posicionais injetam informações sobre a posição relativa ou absoluta dos *tokens* na sequência.

Em seguida, temos o Bloco Decodificador (Decoder) que é semelhante ao codificador, mas com uma terceira subcamada adicional:

1. **Masked Multi-Head Self-Attention (MMHSA):** O MMHSA é principalmente usado no decodificador do Transformer, garantindo que cada posição da sequência só possa acessar informações de posições anteriores ou iguais, preservando a lógica de geração passo a passo. A primeira camada é uma auto-atenção sobre a sequência de saída gerada até o momento. A principal diferença é a aplicação de uma máscara, que impede que as posições atendam a posições subsequentes. Isso garante que, durante o treinamento e a inferência, a previsão para a posição  $i$  dependa apenas das saídas conhecidas nas posições anteriores a  $i$ , preservando a propriedade auto-regressiva.
2. **Encoder-Decoder Attention:** Esta é a segunda camada, onde o decodificador foca na sequência de entrada codificada. Os vetores *Query* vêm da camada de auto-atenção do decodificador, enquanto os vetores *Key* e *Value* vêm da saída da pilha de codificadores. É neste ponto que a informação da sequência de entrada é efetivamente utilizada para guiar a geração da sequência de saída.
3. **Feed-Forward Network:** Idêntica à do codificador.

Assim como no codificador, cada subcamada no decodificador também é encapsulada por conexões residuais e normalização de camada.

#### 2.2.4 O Legado do Transformer

O modelo auto-regressivo descrito, onde os símbolos gerados anteriormente são consumidos como entrada para gerar o próximo, é uma característica fundamental da operação do decodificador durante a inferência. Essa capacidade, combinada com o poderoso mecanismo de atenção, permitiu avanços sem precedentes no campo. A flexibilidade e a eficiência da arquitetura Transformer são a base para a maioria dos modelos de LLMs modernos, incluindo as famílias de modelos *Bidirectional Encoder Representations from Transformers* (BERT), que utiliza apenas a arquitetura do codificador, *Generative Pre-trained Transformer* (GPT), que utiliza apenas a arquitetura do decodificador. Já os modelos *Text-to-Text Transfer Transformer* (T5)/*Bidirectional and Auto-Regressive Transformer* (BART), utilizam a arquitetura completa de codificador-decodificador, consolidando-se como a espinha dorsal dos sistemas avançados de linguagem natural.

### 2.3 Retrieval-Augmented Generation – RAG

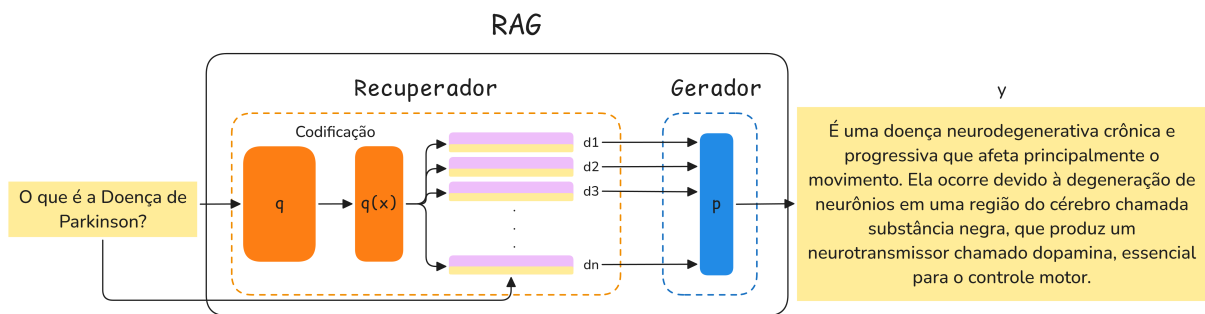
O RAG é uma abordagem que se destaca como uma alternativa promissora ao treinamento convencional de LLMs, cuja demanda computacional torna etapas como fine-tuning ou transfer learning dispendiosas e, muitas vezes, inviáveis em larga escala. com conhecimento externo, evitando limitações impostas por conhecimento fixo e desatualizado dos LLMs. Essa estratégia supera uma das principais limitações dos LLMs tradicionais, como a perda de informações mais antigas. Diferentemente dos modelos gerativos, o RAG consulta dinamicamente uma base de dados ou repositório de documentos e incorpora essas evidências na resposta gerada, o que o torna particularmente eficaz em tarefas que exigem precisão factual, como respostas a perguntas (question answering — QA), geração de conteúdo baseado em fontes e o desenvolvimento de agentes autônomos com comportamento mais informado e contextualizado [Lewis *et al.* 2020].

### 2.3.1 Arquitetura

Independente do paradigma dos sistemas RAG, os componentes principais, conforme figura 3, ainda é composto por duas etapas principais:

1. **Recuperador:** Dado um contexto ou pergunta  $x$ , o módulo de recuperação consulta uma base de documentos  $\mathcal{D}$  para retornar os  $k$  documentos mais relevantes  $\{d_1, d_2, \dots, d_k\}$ .
2. **Gerador:** Um modelo gerativo condicional (geralmente um Transformer) produz uma sequência  $y$  condicionada ao par  $(x, d_i)$  para cada documento.

Figura 3 – Arquitetura RAG



Fonte: Próprio autor, adaptado de [Lewis *et al.* 2020]

Matematicamente, a probabilidade de geração final modelada pelo RAG como:

$$P(y | x) = \sum_{i=1}^k P(y | x, d_i)P(d_i | x) \quad (2.2)$$

onde:

- $P(d_i | x)$  representa a probabilidade do documento  $d_i$  ser relevante para a entrada  $x$ .
- $P(y | x, d_i)$  é modelado por um LM como BART ou T5 [Abootorabi *et al.* 2025].

O módulo de recuperação pode ser baseado em embeddings semânticos com dense retrievers como *Dense Passage Retrieve* (DPR) [Karpukhin *et al.* 2020], que utilizam modelos como BERT para mapear perguntas e documentos em um espaço vetorial.

### 2.3.2 Inferência com Top-k Documentos

Durante a inferência, a geração é feita para cada um dos  $k$  documentos e agregada. Pode-se usar estratégias como:

$$y^* = \arg \max_y \sum_{i=1}^k P(d_i | x)P(y | x, d_i) \quad (2.3)$$

Em implementações como RAG-Sequence, a geração é feita para cada documento separadamente, e depois seleciona-se o mais provável. Já em RAG-Token, cada token gerado é condicionado em múltiplos documentos com atenção cruzada.

### 2.3.3 *Treinamento*

O treinamento é geralmente feito de forma end-to-end. A perda é calculada com base na log-verossimilhança marginalizada sobre os documentos:

$$\mathcal{L}_{\text{RAG}} = -\log \sum_{i=1}^k P(d_i | x) P(y | x, d_i) \quad (2.4)$$

Gradientes podem ser propagados para o gerador e, em versões avançadas, para o recuperador via métodos como aprendizado por reforço ou contrastivo.

### 2.3.4 *RAG Multimodal*

RAG Multimodal estende o paradigma original para lidar com múltiplas modalidades, como texto, imagem e áudio, ao emparelhar um sistema de recuperação multimodal com um modelo generativo capaz de interpretar entradas diversas [Radford *et al.* 2021, Wu *et al.* 2022]. A formulação geral se adapta para:

$$P(y | x_m) = \sum_{i=1}^k P(y | x_m, d_i^m) P(d_i^m | x_m) \quad (2.5)$$

onde  $x_m$  é a entrada multimodal (por exemplo, imagem + texto), e  $d_i^m$  representa documentos multimodais, armazenados como vetores em um espaço latente unificado.

Em muitos sistemas modernos, como Kosmos-1 [Peng *et al.* 2023] e Flamingo [Alayrac *et al.* 2022], utiliza-se uma combinação de *Vision Transformers* (ViT), *Contrastive Language–Image Pre-training* (CLIP) ou BLIP-2 como extratores de embeddings multimodais para indexação e recuperação. O modelo generativo pode ser um LLM com cross-attention para imagens.

Essa abordagem tem se mostrado eficaz em tarefas de *Vision Question Answering* (VQA), grounding de imagens, e RAG com suporte a PDFs escaneados e documentos estruturados.

### 2.3.5 *Vantagens e Desafios*

As vantagens do RAG incluem:

- Atualização de conhecimento sem re-treinamento do modelo gerativo;
- Geração mais factual e interpretável;
- Flexibilidade multimodal.

Entretanto, desafios incluem:

- Dependência da qualidade do módulo de recuperação;
- Latência em sistemas de tempo real;
- Alinhamento semântico entre documentos recuperados e conteúdo gerado.

## 2.4 **Sistemas Multi Agentes - SMA e Princípios de Agentes Autônomos**

Há muito tempo, os agentes autônomos são reconhecidos como uma abordagem promissora para alcançar a *Artificial General Intelligence* (AGI), que deve realizar tarefas por meio de planejamento e ações autogeridas. Em estudos anteriores, frequentemente se assume que os agentes agem com base em políticas simples e heurísticas, que aprendem em ambientes isolados e restritos. No entanto, essas suposições contrastam significativamente com o processo de aprendizado humano, que é altamente complexo, variado de indivíduo para indivíduo e ocorre em uma ampla variedade de contextos. Devido a essas lacunas, os agentes obtidos nos estudos anteriores geralmente estão longe de replicar os processos de tomada de decisão ao nível humano, especialmente em ambientes sem restrições e dinâmicos.

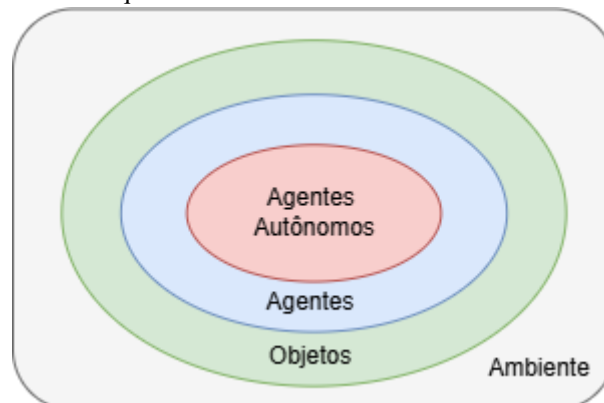
Em decorrência disso, apesar do uso substancial da IA e de todas suas amplas gamas de aplicações, muitas pesquisas se restringem apenas a um domínio, na solução de um problema, sem levar em conta variáveis externas reais que poderiam influenciar constantemente a solução e a noção de um agente com raciocínio adaptativo [Luck e d’Inverno 1995]. Ou seja, embora soluções providas pela IA fossem eficazes para problemas considerados impossíveis pela humanidade, sua aplicação era limitada devido à distância em relação a situações reais que são adaptáveis. Essa limitação motivou o crescente interesse no desenvolvimento de agentes capazes de exibir comportamentos inteligentes em contextos mais realistas, variados e adaptáveis [Luck e d’Inverno 1995].

Diversos trabalhos usam Agentes Autônomos em várias aplicações atualmente, como na medicina [Ghezloo *et al.* 2025, Pandey *et al.* 2024], segurança [Wang *et al.* 2025, Peigné

*et al.* 2025], controle [Devi *et al.* 2024, Kharchouf e Mohammed 2024], entre outras diversas aplicações, são desenvolvidas usando agentes autônomos. No entanto, apesar da aplicação dessa entidade, existe uma certa barreira para definir corretamente o que é um Agente, ou melhor, um Agente Autônomo, pois tem diversos trabalhos na área e a maioria tem seu próprio ponto de vista sobre um Agente, tornando a explicabilidade do conceito algo incerto. Russell em seu livro "Artificial Intelligence: A Modern Approach" [Russell e Norvig 2016] tem a seguinte definição de agente "Um agente é tudo aquilo que pode ser visto como percebendo seu ambiente por meio de sensores e atuando sobre esse ambiente por meio de atuadores.", enquanto que o Stan Franklin em sua obra [Franklin e Graesser 1996] propõe a seguinte definição para um Agente Autônomo "Um agente autônomo é um sistema situado dentro de um ambiente e parte dele, que percebe esse ambiente e age sobre ele, ao longo do tempo, em busca de seus próprios objetivos e de modo a influenciar o que percebe no futuro", ou Ali Dorri em seu trabalho "Uma entidade que é colocada em um ambiente e percebe diferentes parâmetros que são usados para tomar uma decisão com base em seu objetivo. A entidade então executa a ação necessária no ambiente com base nessa decisão". [Dorri *et al.* 2018].

Outro trabalho de destaque na área de IA é o de Mark d’Inverno e Michael Luck, apresentado no artigo "A Formal Framework for Agency and Autonomy". Nesse estudo, os autores não apenas propõem uma definição formal para o conceito de agente, como também introduzem a noção de agência (agency), entendida como a capacidade mental atribuída aos agentes. Diferentemente de outras abordagens da literatura, sua definição é abrangente e estruturada de forma hierárquica conforme a figura 4, permitindo uma categorização progressiva de entidades, de objetos que estão em um ambiente, a agentes, até agentes autônomos, com base em níveis crescentes de complexidade e capacidade [Luck e d’Inverno 1995].

Figura 4 – Estrutura da entidade hierárquica



Fonte: Próprio autor, adaptado de [Luck e d’Inverno 1995]

Apesar das inúmeras definições apresentadas, todas são claramente diferentes, mas convergem para uma explicação comum semelhante de alto nível para essas entidades, que agentes são mecanismos com capacidades de interagir com o ambiente e, baseados nele, tomar decisões.

Apesar dos inúmeros conceitos apresentados, os únicos que produzem categorias nítidas são os conceitos matemáticos, e eles são absolutos apenas porque não têm conteúdo. Os agentes “vivem” no mundo real, ou em algum mundo, e os conceitos do mundo real geram diferentes conceituações difusas sobre o assunto. Então, Stan Franklin propõe uma definição matemática de um agente autônomo. A definição deste, tenta capturar a essência de ser um agente autônomo e definir a classe mais ampla destes. Outras restrições podem ser adicionadas para definir classes mais específicas de agentes. Idealmente, esse esforço produziria uma nomenclatura de agentes que poderia ser usada de forma relativamente inequívoca por pesquisadores da área, resultando em comunicações mais claras [Franklin e Graesser 1996].

Inicialmente, para descrever matematicamente o funcionamento de um agente autônomo, primeiro precisamos recorrer a definições de ambiente, capacidades sensoras, ações, condutores e seleção da arquitetura da ação.

Sabemos que um ambiente pode ser compreendido como um sistema dinâmico em alto nível, que inicialmente é um estado estático, mas que tem uma dinamização sob alguma interação, portanto, modelaremos como um sistema dinâmico. Um ambiente,  $E$ , muda ao longo do tempo e, portanto, pode ser descrito como um sistema dinâmico  $T : X \rightarrow X$ , em que  $X$  é o espaço de todos os possíveis estados globais de  $E$  e  $T$  é a dinâmica global que atualiza o estado atual de  $E$ . O espaço de estado  $X$  é geralmente um espaço vetorial, mas pode assumir outras formas, como uma árvore com uma lista em cada nó [Song *et al.* 1996].  $T$  pode ser considerado uma implementação da física de  $E$  e pode ser atualizado em tempo discreto ou em tempo contínuo. A atualização discreta pode ser descrita por equações de diferença, e a atualização contínua por equações diferenciais.  $T$  pode ser determinístico ou estocástico (assim como nosso ambiente baseado na física quântica). A atualização de  $E$  pode ser influenciada pelas ações de agentes autônomos que “vivem” no ambiente. Como cada agente que vive em  $E$  é, por si só, uma parte de  $E$ , suas ações podem ser consideradas como influenciadoras da ação de atualização de  $T$ . Um ambiente inicial  $E(0)$  deve ser fornecido.  $E(t)$  denota o estado do ambiente no momento  $t$ . Cada  $E(t)$  é um elemento do espaço de estado  $X$  [Franklin e Graesser 1996].

Se  $E$  for um sistema de agente único, deixe  $A$  denotar esse agente. Por abuso de linguagem, deixe  $A : X \rightarrow X$  também denotar a atualização de  $E$  pelas ações do agente  $A$ . Deixe  $P : X \rightarrow X$  denotar a atualização de  $E$  que não seja pelas ações de  $A$ . Juntos,  $A$  e  $P$  constituem a dinâmica global  $T$  do sistema.

Em um sistema multiagente, permita um número finito de agentes  $A_j$ , cada um representando suas ações, e  $P$  a atualização que não seja dessas ações.

Os recursos de detecção de um agente autônomo  $A$  são descritos em termos de seus sensores. Cada sensor retorna para  $A$  alguma parte específica do ambiente atual  $E(t)$ . Em geral,  $E(t)$  é um vetor grande com cada atributo de  $E$  fornecendo uma dimensão. Assim, o sensoriamento por  $A$ , no tempo  $t$ , é dado por  $S(A, t) = P(E(t))$ , em que  $P$  é uma projeção em um subespaço que depende dos recursos de sensoriamento de  $A$ . Os sensores são incorporados ao agente, embora possam se degradar com o tempo (como a visão e a audição em humanos). Os sensores podem ser aumentados com o uso de instrumentos (telescópios, laços de joalheiro, receptores de TV). O sensoriamento pode ser ativo, exigindo ação por parte do agente, ou passivo. (Todo o sensoriamento do Sumpy é ativo, exigindo a emissão de comandos do UNIX. A audição humana é principalmente passiva. O sensoriamento pode ser interno ou externo ao agente  $A$ . (A propriocepção é interna, assim como o fato de um ser humano estar ciente do ato de se programar.)

Uma ação do agente  $A$  produz alguma alteração no estado atual de seu ambiente. Ou seja, a execução da ação influencia a ação de atualização de  $T$ . Descreva a ação descrevendo essa influência. Cada agente vem com um conjunto de ações primárias incorporadas. As sequências dessas ações podem produzir ações de nível superior. Em alguns agentes, as ações podem ser executadas em paralelo. O efeito de realizar uma determinada ação no momento  $t$  depende do estado do ambiente naquele momento.

## 2.5 Aprendizado por Reforço Multiagente (*Multi-Agent Reinforcement Learning – MARL*)

O Aprendizado por Reforço Multiagente (MARL) estende o paradigma de RL de um único agente para cenários que envolvem múltiplos agentes de aprendizado autônomos. Estes agentes coexistem em um ambiente compartilhado, onde a ação de um agente pode afetar o estado do ambiente e, conseqüentemente, as observações e recompensas recebidas pelos outros agentes. Essa interdependência transforma o problema de otimização de uma única política em um complexo jogo dinâmico, introduzindo desafios que não existem no cenário de agente único.

### 2.5.1 Formalização: O Processo de Decisão de Markov Descentralizado e Parcialmente Observável (Dec-POMDP)

A estrutura mais geral para problemas MARL é o Processo de Decisão de Markov Descentralizado e Parcialmente Observável (Dec-POMDP). Ele captura a essência dos desafios multiagente: execução descentralizada e percepção limitada do mundo. Um Dec-POMDP é formalizado pela tupla:

$$G = \langle \mathcal{S}, \{\mathcal{A}_i\}_{i=1}^N, P, \{R_i\}_{i=1}^N, \Omega, \{O_i\}_{i=1}^N, \gamma \rangle \quad (2.6)$$

onde  $N$  é o número de agentes;  $\mathcal{S}$  é o espaço de estados globais e o estado  $s \in \mathcal{S}$  contém toda a informação sobre o sistema, mas geralmente é inacessível para os agentes individualmente;  $\mathcal{A}_i$  é o conjunto de espaços de ações do agente  $i$ ;  $P(S'|s, a)$  é a função de probabilidade de transição de estado. Ela define a dinâmica do ambiente, que depende do estado atual  $s$  e da ação conjunta  $a$  de todos os agentes;  $R_i(s, a)$  é o conjunto de funções de recompensa, onde  $R_i(s, a)$  é a recompensa recebida pelo agente  $i$ . Note que a recompensa pode depender do estado global e da ação de todos;  $\Omega$  é o espaço de observações conjuntas e  $\{O_i\}_{i=1}^N$  é a função de observação.  $O_i(o_i|s)$  ornece a probabilidade do agente  $i$  receber a observação local  $o_i \in \Omega_i$  dado que o ambiente transitou para o estado  $s'$  após a ação conjunta  $a$ . A observação parcial ( $o_i \neq s$ ) é uma característica central do MARL;  $\gamma \in [0, 1]$  é o fator de desconto, que pondera a importância de recompensas futuras.

Neste cenário, cada agente  $i$  busca aprender uma política estocástica  $\pi_\theta(a_i|h_i)$ , que mapeia seu histórico local de observação-ação  $h_i^t = (o_i^0, a_i^0, \dots, o_i^t)$  para uma distribuição de probabilidade sobre suas ações. O objetivo de cada agente é maximizar seu próprio retorno esperado, que é o valor descontado acumulado de suas futuras recompensas:

$$J_i(\theta_i) = \mathbb{E}_{\pi_{\theta_1}, \dots, \pi_{\theta_N}} \left[ \sum_{t=0}^{\infty} \gamma^t R_i(s_t, \mathbf{a}_t) \right] \quad (2.7)$$

A expectativa  $E[\cdot]$  é tomada sobre as trajetórias de estado-ação geradas pelas políticas de todos os agentes. Essa interdependência nas políticas é a fonte dos maiores desafios em MARL. Dependendo da estrutura das funções de recompensa, os sistemas MARL podem ser classificados como:

- **Totalmente Cooperativos:** Todos os agentes compartilham a mesma função de recompensa ( $R_1 = \dots = R_N$ ).

- **Totalmente Competitivos:** O ganho de um agente é a perda de outro (soma-zero, por exemplo,  $\sum_i R_i = 0$ ).
- **Soma-Geral (ou Misto):** Não há uma relação fixa entre as recompensas dos agentes, sendo o caso mais geral e complexo.

### 2.5.2 *Desafios Fundamentais em MARL*

A otimização simultânea de  $N$  políticas interdependentes dá origem a dois desafios principais que distinguem MARL do RL de agente único.

#### **Não-Estacionariedade do Ambiente Percebido**

Do ponto de vista de um único agente  $i$ , o ambiente é não-estacionário. Enquanto o agente  $i$  atualiza sua política  $\pi_{\theta_i}$  para melhor se adaptar ao ambiente, os outros agentes, o ambiente do ponto de vista de  $i$ , também estão mudando suas políticas  $\pi_{\theta_{-i}}$ . Isso significa que a dinâmica de transição percebida pelo agente  $i$ ,  $P(s'|s, a_i)$ , não é fixa, pois ela depende implicitamente das políticas dos outros agentes. Essa não-estacionariedade viola a premissa da Propriedade de *Markov*, que é fundamental para a convergência de muitos algoritmos de RL tradicionais, como o *Q-learning*.

Uma abordagem dominante para mitigar esse problema é o paradigma de Treinamento Centralizado com Execução Descentralizada, ou do inglês, **Centralized Training with Decentralized Execution (CTDE)** [Oliehoek *et al.* 2008, Kraemer e Banerjee 2016]. Durante o treinamento, um "crítico" centralizado pode acessar informações globais (como os estados  $s$  e as ações  $a$  de todos os agentes) para aprender uma função de valor e fornecer um sinal de gradiente estável. Na fase de execução, no entanto, cada agente seleciona suas ações usando apenas sua política local (ator), que depende exclusivamente de suas observações locais, preservando a descentralização.

#### **O Problema da Atribuição de Crédito (Credit Assignment)**

Em cenários cooperativos, os agentes frequentemente recebem uma recompensa de equipe compartilhada (e.g.,  $R_i = R_{global}$ ). Se a equipe tem sucesso (recompensa positiva) ou falha (recompensa negativa), não tem como determinar a contribuição individual de cada agente para esse resultado e **este é o problema da atribuição de crédito multiagente**. Sem uma solução para isso, um agente que tomou uma ação útil pode ser "punido" por uma recompensa negativa causada por outros, e vice-versa. Isso leva a um sinal de aprendizado com alta variância e dificulta a convergência.

Técnicas como as **recompensas de diferença** (*difference rewards*) visam fornecer um sinal de aprendizado mais focado para cada agente. A recompensa de diferença para o agente  $i$  é calculada como:

$$D_i(s, \mathbf{a}) = R_i(s, \mathbf{a}) - R_i(s, \mathbf{a}_{-i}, a_i^{\text{baseline}}) \quad (2.8)$$

onde  $R(s, a)$  é a recompensa global da equipe e  $R(s, (a_{-i}, c_i))$  é uma recompensa contrafactual, que simula o que teria acontecido se o agente  $i$  tivesse tomado uma ação padrão (ou de linha de base)  $c_i$ , enquanto os outros agentes mantinham suas ações  $a_i$ . Intuitivamente,  $D_i$  isola a contribuição marginal do agente  $i$  para o resultado da equipe. Algoritmos como o *Counterfactual Multi-Agent Policy Gradients* (COMA), formalizam essa ideia usando um crítico centralizado para calcular essa vantagem contrafactual [Foerster *et al.* 2018]. Outra linha de pesquisa, como Value Decomposition Networks (VDN) e QMIX, aborda o problema assumindo que a função de valor conjunta pode ser decomposta (monotonicamente, no caso do QMIX) a partir das funções de valor individuais dos agentes [Sunehag *et al.* 2017, Rashid *et al.* 2020].

## 2.6 Multi-Agent Deep Deterministic Policy Gradient - MADDPG

O algoritmo *Multi-Agent Deep Deterministic Policy Gradient* (MADDPG), proposto por Lowe *et al.* (2017), é uma extensão direta do algoritmo *Deep Deterministic Policy Gradient* (DDPG) [Lillicrap *et al.* 2015] para o domínio multiagente. Ele foi projetado especificamente para lidar com a não-estacionariedade do ambiente que surge quando múltiplos agentes aprendem simultaneamente, aplicando o paradigma de treinamento centralizado com execução descentralizada *Centralized Training with Decentralized Execution* (CTDE).

A ideia central do MADDPG é que, se conhecermos as ações de todos os agentes, o ambiente se torna estacionário do ponto de vista de qualquer agente individual, mesmo que as políticas dos outros agentes estejam mudando. O MADDPG explora essa ideia durante o treinamento, mas permite que os agentes atuem de forma descentralizada durante a execução.

### 2.6.1 Arquitetura: Atores Descentralizados e Críticos Centralizados

O MADDPG mantém uma estrutura de Ator-Crítico para cada agente  $i$ , mas com uma modificação fundamental:

- **Ator (Política) Determinístico e Descentralizado:** Cada agente  $i$  possui uma rede de ator,  $\mu_{\theta_i}$ , que se aproxima de sua política. Esta política é descentralizada, pois mapeia apenas a **observação local** do agente  $o_i \in \mathcal{O}_i$  para uma **ação específica** (determinística)  $a_i \in \mathcal{A}_i$ .

$$\mu_{\theta_i}: \mathcal{O}_i \rightarrow \mathcal{A}_i \quad (2.9)$$

Essa descentralização é crucial para a execução, pois permite que o agente tome decisões no mundo real sem precisar de informações dos outros agentes.

- **Crítico (Função de Valor) Centralizado:** Cada agente  $i$  também possui uma rede de crítico,  $Q_{\phi_i}$ , que aprende a função de valor da ação,  $Q_i(s, a)$ . Diferente do ator, o crítico é centralizado: ele recebe como entrada não apenas o estado global  $s$  (ou uma concatenação das observações de todos os agentes), mas também a ação conjunta  $a = (a_1, \dots, a_N)$  de todos os agentes.

$$Q_{\phi_i}(s, \mathbf{a}) \quad (2.10)$$

Ao condicionar a função de valor às ações de todos os agentes, o crítico pode aprender um modelo estável do ambiente, contornando efetivamente o problema da não-estacionariedade durante a fase de treinamento.

### 2.6.2 Treinamento Centralizado

O treinamento ocorre de forma *off-policy*, utilizando amostras de um *replay buffer* compartilhado,  $\mathcal{D}$ , que armazena tuplas de transição  $(s, a, r, s')$ .

**Atualização do Crítico** O crítico de cada agente,  $Q_{\phi_i}$ , é atualizado para minimizar o erro de Bellman, de forma análoga ao DDPG. A função de perda é o erro quadrático médio (MSE) entre a predição atual do crítico e o valor-alvo (target)  $y_i$ :

$$\mathcal{L}(\phi_i) = \mathbb{E}_{(s, \mathbf{a}, r_i, s') \sim \mathcal{D}} \left[ (Q_{\phi_i}(s, \mathbf{a}) - y_i)^2 \right] \quad (2.11)$$

O valor-alvo  $y_i$  é calculado usando as redes-alvo, *target networks*, que são cópias das redes principais com atualização lenta. Isso estabiliza o treinamento, fornecendo um alvo de otimização mais estável.

$$y_i = r_i + \gamma Q_{\phi'_i}(s', \mathbf{a}') \quad (2.12)$$

Note que o cálculo do valor futuro  $\mathcal{Q}_{\phi_i}(s', a')$  utiliza as ações do próximo estado,  $a'$ , que são obtidas a partir das políticas-alvo de cada agente,  $\mu_{\phi_i}$ . O fato de o crítico centralizado ter acesso a essas políticas-alvo durante o treinamento é um pilar do *framework* CTDE.

**Atualização do Ator** O ator de cada agente,  $\mu_{\phi_i}$ , é atualizado para produzir ações que maximizem o valor esperado retornado pelo seu crítico. Isso é feito ascendendo ao longo do gradiente da política determinística. Como o crítico  $\mathcal{Q}_{\phi_i}$  é uma função diferenciável que considera as ações de todos os agentes, ele pode propagar um gradiente útil para o ator. O gradiente da política para o agente  $i$  é:

$$\nabla_{\theta_i} J_i = \mathbb{E}_{s \sim \mathcal{D}} [\nabla_{\theta_i} \mu_{\theta_i}(o_i) \nabla_{a_i} \mathcal{Q}_{\phi_i}(s, \mathbf{a})] \quad (2.13)$$

Intuitivamente,  $\nabla_{a_i} \mathcal{Q}_{\phi_i}(\cdot)$  indica a direção na qual a ação  $a_i$  deve ser alterada para aumentar o valor  $Q$ , e  $\nabla_{a_i} \mu_{\phi_i}(o_i)$  traduz essa direção para uma atualização nos parâmetros  $\theta_i$  do ator.

### 2.6.3 Estabilidade do Treinamento com Redes-Alvo

Para evitar oscilações e instabilidades no processo de aprendizado, o MADDPG, assim como o DDPG, utiliza redes-alvo (*target networks*) para o ator ( $\mu_{\theta'_i}$ ) e para o crítico ( $\mathcal{Q}_{\phi'_i}$ ). Os parâmetros dessas redes ( $\theta'_i, \phi'_i$ ) não são treinados diretamente por retropropagação. Em vez disso, eles são atualizados lentamente para rastrear os parâmetros das redes principais ( $\theta_i, \phi_i$ ) através de uma **atualização suave**, *soft update*:

$$\theta' \leftarrow \tau \theta_i + (1 - \tau) \theta'_i \quad (2.14)$$

$$\phi' \leftarrow \tau \phi_i + (1 - \tau) \phi'_i \quad (2.15)$$

onde  $\tau \ll 1$  é um hiperparâmetro que controla a velocidade da atualização. Isso garante que o valor-alvo  $y_i$  mude lentamente, proporcionando um alvo de regressão mais estável para o crítico e, por consequência, um gradiente mais estável para o ator.

Em resumo, o MADDPG oferece uma solução elegante e eficaz para o treinamento de múltiplos agentes em ambientes complexos, especialmente aqueles com espaços de ação contínuos. Sua principal contribuição é a forma como implementa o CTDE: usando críticos

centralizados para guiar o aprendizado de atores totalmente descentralizados, resolvendo a não-estacionariedade e permitindo a atribuição de crédito implícita por meio de uma função de valor informada globalmente.

## 2.7 Proximal Policy Optimization — PPO

### 2.7.1 *Objetivo de clipping*

Dada uma política estocástica  $\pi_\theta$  e sua versão antiga  $\pi_{\theta_{old}}$ , definimos

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \quad (2.16)$$

O objetivo de maximização é

$$L^{\text{clip}}(\theta) = \mathbb{E}_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \varepsilon, 1 + \varepsilon)\hat{A}_t)] \quad (2.17)$$

com  $\hat{A}_t$  vantagem generalizada (GAE- $\lambda$ ) e  $\varepsilon \in [0.1, 0.3]$ .

### 2.7.2 *Extensões multi-agentes*

Em cenários MARL, PPO costuma ser integrado com CTDE. Variantes notáveis:

- MAPPO: Um crítico centralizado global  $V_\phi(s)$  shared; políticas independentes  $\pi_{\theta_i}$
- IPPO: Cada agente aplica PPO de forma independente, tratando outros como parte do ambiente; estabilidade surpreendente em domínios cooperativos.

Gradientes: para cada agente  $i$

$$\nabla_{\theta_i} L_i^{\text{clip}} = \mathbb{E}[\nabla_{\theta_i} \log \pi_{\theta_i}(a_i|o_i) \hat{A}_i^{\text{GAE}} \mathbb{I}_{|r_i-1| \leq \varepsilon}] \quad (2.18)$$

### 2.7.3 *Considerações práticas*

- Normalização de recompensa/estado para evitar explosão nos raios
- *KL-penalty adaptation*: ajusta  $\varepsilon$  dinamicamente visando manter o desvio de política abaixo de um alvo predefinido.
- Para ambientes parcialmente observáveis, empregar recorrência (LSTM/GRU) dentro da política.

## 2.8 Direct Preference Optimization - DPO

O alinhamento de LLMs com preferências humanas é um passo crucial para garantir que sejam úteis e seguros. O paradigma dominante para esse alinhamento tem sido o (***Reinforcement Learning from Human Feedback (RLHF)***), ou do português, **Aprendizado por Reforço com Feedback Humano**. No entanto, o pipeline de RLHF é notavelmente complexo, envolvendo múltiplos estágios de treinamento que podem ser computacionalmente caros e instáveis. O *Direct Preference Optimization (DPO)* [Rafailov *et al.* 2023] surge como uma alternativa elegante e eficaz que simplifica drasticamente esse processo.

### 2.8.1 Limitações do Pipeline Clássico de RLHF

O pipeline canônico de RLHF consiste em três etapas principais:

1. ***Supervised Fine-Tuning (SFT)***, ou do português, **Ajuste Fino Supervisionado**: Um modelo de linguagem pré-treinado é ajustado em um conjunto de dados de alta qualidade de exemplos de prompt-resposta, resultando em uma política de referência,  $\pi_{\text{ref}}$ .
2. ***Reward Model Training (RMT)***, ou do português, **Treinamento de um Modelo de Recompensa**: Pares de respostas para um mesmo prompt ( $y_w, y_l$ ) são coletados, onde  $y_w$  é a resposta preferida ("vencedora") e  $y_l$  a preterida ("perdedora"). Um modelo de recompensa,  $r_\phi(x, y)$ , é então treinado para prever a preferência humana. Geralmente, isso é feito otimizando um objetivo de classificação binária sobre os pares, com base no modelo de Bradley-Terry, que postula que a probabilidade de preferência é uma função sigmoideal da diferença das recompensas:

$$p(y_w \succ y_l | x) = \sigma(r_\phi(x, y_w) - r_\phi(x, y_l)) \quad (2.19)$$

3. **Otimização por RL**: A política  $\pi_{\text{ref}}$  é otimizada usando um algoritmo de RL (comumente o PPO), maximizando a recompensa fornecida pelo RMT treinado. Para evitar que a política se desvie excessivamente da distribuição original, um termo de penalidade baseado na divergência KL é adicionado:

$$\max_{\pi_\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(y|x)} \left[ r_\phi(x, y) - \beta \log \left( \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} \right) \right] \quad (2.20)$$

Este processo, embora eficaz, é complexo e frágil. Requer treinar três modelos distintos (SFT, RM e a política final), e a fase de RL pode ser instável e sensível a hiperparâmetros.

### 2.8.2 A Formulação do DPO: De RL para Otimização Direta

A principal intuição do DPO é que o objetivo de RL restrito por KL pode ser otimizado diretamente sobre os dados de preferência, sem a necessidade de treinar um modelo de recompensa explícito ou executar um *loop* de RL.

O ponto de partida é a solução analítica para o objetivo de RL mencionado acima. A política ótima,  $\pi^*$ , que maximiza esse objetivo, pode ser expressa em função do modelo de recompensa  $r_\phi$  e da política de referência  $\pi_{\text{ref}}$ :

$$\pi^*(y|x) = \frac{1}{Z(x)} \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r_\phi(x, y)\right) \quad (2.21)$$

onde  $Z(x)$  é uma função de partição para garantir que as probabilidades somem 1.

O DPO inverte essa relação para expressar a recompensa em termos das políticas:

$$r_\phi(x, y) = \beta \left( \log\left(\frac{\pi^*(y|x)}{\pi_{\text{ref}}(y|x)}\right) + \log Z(x) \right) \quad (2.22)$$

Substituindo esta definição de recompensa na equação de preferência do modelo de Bradley-Terry, os termos da função de partição  $Z(x)$  se cancelam, resultando em uma expressão para a probabilidade de preferência que depende apenas das políticas:

$$p(y_w \succ y_l|x) = \sigma\left(\beta \log \frac{\pi^*(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi^*(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right) \quad (2.23)$$

Isso estabelece um mapeamento direto entre a otimização de uma política de linguagem e um objetivo de preferência humana.

### 2.8.3 O Objetivo de Otimização do DPO

Com essa nova formulação, o DPO treina a política  $\pi_\theta$  para satisfazer as preferências humanas diretamente. O objetivo é maximizar a verossimilhança dos dados de preferência humana, usando a política  $\pi_\theta$  no lugar da política ótima teórica  $\pi^*$ . Isso se traduz na minimização do seguinte objetivo de log-verossimilhança negativa:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[ \log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right) \right] \quad (2.24)$$

Vamos definir  $\hat{r}_\theta(y, x) = \beta \log(\pi_\theta(y|x)/\pi_{\text{ref}}(y|x))$  como a recompensa implícita aprendida pela política. O objetivo pode ser reescrito de forma mais clara:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} [\log \sigma(\hat{r}_\theta(y_w, x) - \hat{r}_\theta(y_l, x))] \quad (2.25)$$

Intuitivamente, este objetivo treina a política  $\pi_\theta$  para aumentar a probabilidade relativa da resposta "vencedora"  $y_w$  em comparação com a resposta "perdedora"  $y_l$ , em relação à política de referência. Em essência, ele classifica as respostas, utilizando um sinal de recompensa que é implicitamente definido pelas próprias probabilidades da política.

As vantagens do DPO são claras: ele colapsa o complexo pipeline de RLHF em uma única etapa de treinamento, análoga ao ajuste fino supervisionado, tornando o processo mais estável, eficiente e simples de implementar, ao mesmo tempo em que alcança desempenho competitivo ou superior aos métodos de RLHF.

## 2.9 Group Relative Policy Optimization - GRPO

Embora o DPO represente um avanço significativo sobre o RLHF, ele é inerentemente limitado a dados de preferência aos pares, pairwise. No entanto, os dados de feedback humano podem ser muito mais ricos, como classificações (por exemplo: notas de 1 a 5) ou *rankings* de múltiplas respostas. O *Group Relative Policy Optimization* (GRPO) [Shao *et al.* 2024] generaliza o DPO para alavancar eficientemente esses dados de preferência em nível de grupo.

### 2.9.1 A Limitação dos Dados Pares

Considere um cenário onde, para um *prompt*  $x$ , um anotador humano classifica um conjunto de  $K$  respostas,  $\mathcal{Y} = \{y_1, y_2, \dots, y_K\}$ , resultando em um ranking completo, por exemplo,  $y_1 \succ y_2 \succ \dots \succ y_K$ , ou seja, a resposta de  $y_1$  é preferida em relação a  $y_2$  que por sua vez é mais preferível que  $y_3$ , assim por diante. Para usar o DPO, seria necessário decompor esse ranking em  $\binom{K}{2}$  pares de preferências (por exemplo:  $(y_1, y_2), (y_1, y_3), \dots, (y_{K-1}, y_K)$ ). Essa abordagem é subótima por duas razões: i. Ineficiência Computacional, aumenta drasticamente o tamanho do conjunto de dados; ii. Perda de Informação, a estrutura holística do *ranking* é perdida, tratando cada par como uma observação independente. O GRPO aborda essa limitação ao modelar diretamente a probabilidade de todo o *ranking* do grupo.

### 2.9.2 Generalizando Preferências com o Modelo Plackett-Luce

O GRPO substitui o modelo de Bradley-Terry, usado para pares, pelo modelo de Plackett-Luce, que define uma distribuição de probabilidade sobre todos os *rankings* possíveis

de um conjunto de itens. A probabilidade de um *ranking* específico  $y_1 \succ \dots \succ y_K$  é dada por:

$$P(y_1 \succ \dots \succ y_K | x) = \prod_{i=1}^K \frac{\exp(r(x, y_i))}{\sum_{j=i}^K \exp(r(x, y_j))} \quad (2.26)$$

onde  $r(x, y_i)$  é o escore de utilidade, recompensa, latente da resposta  $y_i$ . Esta fórmula pode ser entendida sequencialmente: a probabilidade de  $y_1$  ser a melhor é seu escore sobre a soma de todos os escores. Dado que  $y_1$  foi escolhida, a probabilidade de  $y_2$  ser a próxima melhor é seu escore sobre a soma dos escores restantes, e assim por diante.

### 2.9.3 O Objetivo de Otimização do GRPO

Assim como o DPO, o GRPO dispensa um modelo de recompensa explícito e, em vez disso, utiliza a recompensa implícita definida pela política:  $\hat{r}_\theta(y, x) = \beta \log(\pi_\theta(y|x)/\pi_{\text{ref}}(y|x))$ . Ao substituir  $r(x, y)$  por  $\hat{r}_\theta(y, x)$  na fórmula de Plackett-Luce, obtemos a probabilidade de um *ranking* em função da política  $\pi_\theta$ .

O objetivo do GRPO é maximizar a log-verossimilhança de observar os rankings de grupo no conjunto de dados. A função de perda do GRPO, para um conjunto de respostas ranqueadas  $\mathcal{Y} = \{y_1, \dots, y_K\}$ , é:

$$\mathcal{L}_{\text{GRPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, \mathcal{Y}) \sim \mathcal{D}} \left[ \log \prod_{i=1}^K \frac{\exp(\hat{r}_\theta(y_i, x))}{\sum_{j=i}^K \exp(\hat{r}_\theta(y_j, x))} \right] \quad (2.27)$$

Expandindo o logaritmo, a perda se torna:

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}_{(x, \mathcal{Y}) \sim \mathcal{D}} \left[ \sum_{i=1}^K \left( \hat{r}_\theta(y_i, x) - \log \sum_{j=i}^K \exp(\hat{r}_\theta(y_j, x)) \right) \right] \quad (2.28)$$

Este objetivo incentiva a política  $\pi_\theta$  a atribuir escores de recompensa implícitos ( $\hat{r}_\theta$ ) que são consistentes com o ranking completo do grupo. Ele efetivamente maximiza a margem entre cada item e os itens classificados abaixo dele no *ranking*.

É importante notar que, quando o tamanho do grupo é  $K = 2$ , o modelo de Plackett-Luce se reduz ao modelo de Bradley-Terry, e a função de perda do GRPO se torna idêntica à do DPO. Isso demonstra que o GRPO é uma generalização natural e consistente do DPO.

Ao permitir o uso direto de dados de *ranking*, o GRPO oferece um método mais poderoso e eficiente em termos de dados para o alinhamento de LLMs, capturando nuances nas preferências humanas que seriam perdidas em uma formulação estritamente pareada.

### 3 TRABALHOS RELACIONADOS

O diagnóstico precoce da DP pode ser muito complexo devido à grande modalidade de dados que podem ser analisados para fazer uma análise da doença. Geralmente, as abordagens mais comuns envolvem analisar individualmente pelo menos um desses tipos de dados:

1. Tabulares
  - 1.1 Temporal
  - 1.2 Atemporal
2. Áudios
3. Vídeos
4. Imagens

Em virtude disso, a solução inovadora baseada em um ecossistema agêntico com análise multimodal online, ou quase tempo real, do termo em inglês *near-real-time*, usando IoT, para diagnóstico precoce da DP emerge como um grande desafio na literatura atual pelo número de fatores analisados, uso de dispositivo de borda, comunicação com a nuvem, fusão de dados (sensores, imagens, vídeos, áudios, tabulares e etc), para um resultado extremamente assertivo, rápido e otimizado. Para compreender o estado da arte do tema, realizou-se uma revisão bibliográfica entre os anos de 2023 e 2025, com foco nas contribuições recentes da literatura científica. Portanto, destacam-se os principais trabalhos relacionados identificados a este trabalho, conforme mostram as tabelas 1, 2, 5 e 6.

A pesquisa do estado da arte se dividiu em três partes, a primeira destacando análise multimodal usando um sistema multiagente para diagnóstico precoce de DP. A segunda, procurando analisar abordagens não multimodais que usam arquiteturas baseadas em transformadores, redes neurais convolucionais, redes neurais recorrentes ou *Machine Learning*, para o diagnóstico precoce de DP. Por fim, trabalhos que apresentam plataformas do começo ao fim, *end-to-end*, que oferecem um sistema de diagnóstico precoce mediante um *chat* conversacional não invasivo, com respostas em quase tempo real que, além de unificar dados dos pacientes usuários, informam respostas baseadas em seus dados.

Na primeira parte, o método utilizado de busca incorporou análises em sete periódicos, Spring, IEEE, ACM, Nature, PlosOne, Wiley e Science Direct, com a seguinte consulta “((*multi-agents*) AND (*parkinson*))”. Em seguida, utilizando uma consulta e analisando somente IEEE, Spring, Nature e Science Direct, buscou-se “((*parkinson*) AND (*transformers*)) OR (*cnn*) OR (*lstm*) OR (*machine learning*)”.

### 3.1 Diagnóstico precoce de DP usando sistemas multiagentes autônomos baseados em modelos de linguagem

Tabela 1 – Características Científicas e Analíticas dos trabalhos

Autores	Objetivo do Estudo	Metodologia	Análise Estatística	Principais Resultados	Contribuições	Limitações
[Rabie e Akhloufi 2025]	Revisar técnicas de ML e DL para detectar Parkinson	Revisão sistemática	-	DL melhora acurácia de diagnóstico	Síntese abrangente de abordagens	Validação clínica pouco discutida
[Gelderen e Tejedor-García 2024]	Revisar DL baseado em fala para classificação de PD	Revisão sistemática	Comparações entre estudos	DL supera ML em robustez	Síntese de arquiteturas DL	Pouca padronização
[Akila e Nayahi 2024]	Classificar PD com NN + algoritmo de massa	Extração de voz + NN	Acurácia, curva ROC	Bons resultados (acurácia 97%)	Uso inovador de algoritmo de massa	Dataset pequeno
[Xavier <i>et al.</i> 2025]	Revisar análise de voz para PD	Revisão sistemática	Qualitativa	Voz útil na triagem de PD	Organização por técnica e dados	Foco exclusivo em voz

Fonte: Próprio autor.

A Tabela 1 reúne recentes trabalhos que exploram métodos baseados em ML e DL para a detecção da DP, com ênfase em abordagens aplicadas à fala. As revisões sistemáticas de [Rabie e Akhloufi 2025] e [Gelderen e Tejedor-García 2024] destacam o desempenho superior do DL em relação ao ML, embora apontem limitações como validação clínica insuficiente e falta de padronização metodológica. O estudo experimental de [Akila e Nayahi 2024] obteve alta acurácia (97%) combinando redes neurais e algoritmos de massa, mas seu conjunto de dados limitado compromete a generalização dos resultados. Já [Xavier *et al.* 2025] reforça o potencial da análise vocal como ferramenta de triagem, embora tenha abordado apenas esse tipo de dado, o que limita a abrangência. No conjunto, os trabalhos apontam avanços promissores, mas também evidenciam desafios quanto à validação prática e padronização das técnicas.

Tabela 2 – Recursos e Aplicações Práticas dos trabalhos

Área de Aplicação	Tipo de Dados	Dados Utilizados	Link para Código	Frameworks/ Tecnologias
Diagnóstico de Parkinson	Áudio, Imagem	Vários datasets revisados	-	TensorFlow, PyTorch
Diagnóstico por fala	Áudio	mPower, PC-GITA, UCI Voice.	-	CNN, LSTM, MFCC
Classificação por voz	Áudio	UCI Parkinson (prov.)	-	MATLAB, NN toolbox
Análise vocal para PD	Áudio	Diversos	-	PRAAT, OpenSMILE, WEKA

Fonte: Próprio autor.

Já na, Tabela 2, é apresentado um panorama dos recursos utilizados e das aplicações práticas nos estudos voltados à detecção da DP, com destaque para a análise de sinais de voz, destacado na tabela 1. A principal área de aplicação comum entre os trabalhos é o diagnóstico ou triagem da doença a partir de dados auditivos, com ênfase no uso de voz. Diversos conjuntos de dados foram empregados, incluindo bases amplamente reconhecidas como *mPower*, *PC – GITA*

e *UCIVoice*, sendo estas utilizadas em abordagens baseadas em CNN, redes recorrentes como LSTM e extração de características com *Mel-Frequency Cepstral Coefficients* (MFCC), ou em português, Coeficientes Cepstrais na Escala de Mel. Algumas pesquisas empregaram ferramentas como *MATLAB* e sua *Neural Network Toolbox* para classificação, enquanto outras utilizaram *frameworks* populares como *TensorFlow* e *PyTorch*, evidenciando a predominância de técnicas de DL. Embora os trabalhos façam uso de diferentes tecnologias e *datasets*, é notável a ausência de *links* diretos para os códigos implementados, o que limita a rigidez de novos estudos relacionados.

### 3.2 Abordagens de Aprendizado de Máquina e Aprendizado Profundo para o diagnóstico DP

Tabela 3 – Aprendizado de máquina e Aprendizado profundo - Características Científicas e Analíticas dos Estudos

Autores	Objetivo do Estudo	Metodologia	Análise Estatística	Principais Resultados	Contribuições	Limitações
[Chauhan e Ghosal 2024]	Prever precocemente a DP usando dados biométricos e redes neurais	Arquitetura híbrida CNN + BiLSTM	Acurácia, Precisão, Revocação	Acurácia significativa com dados sensoriais	Integração temporal e espacial eficiente para predição precoce	Dados limitados a tarefas motoras específicas
[Chen <i>et al.</i> 2024]	Segmentar núcleos cerebrais relacionados a DP em imagens de ressonância	CNN com codificador intercalado com Transformer	Dice coefficient, IoU	Segmentação mais eficiente que redes tradicionais	Novo encoder CNN-Transformer com fusão de características	Necessita validação em conjuntos clínicos maiores
[Khanom <i>et al.</i> 2025]	Propor abordagem de boosting com seleção de atributos	Modelo híbrido de Machine Learning (LightGBM + seleção)	Acurácia, precisão, recall, F1	Melhor desempenho que modelos padrão	Uso de explicabilidade (SHAP)	Dataset pequeno
[Zhao <i>et al.</i> 2025]	Métodos de IA de pesquisa para avaliação multimodal da DP	Revisão sistemática de literatura	Acurácia, Sensibilidade e Especificidade	Síntese de métricas de desempenho de múltiplos estudos, sem meta-análise. AUC de 89,5% indica superioridade do modelo proposto frente a métodos unimodais.	Taxonomia abrangente de técnicas multimodais	Heterogeneidade de datasets e falta de protocolos padronizados
[Sharma e Gupta 2024]	Desenvolver modelos de DL para previsão de disfunção cognitiva em DP	Redes neurais convolucionais	Validação cruzada e relatórios de AUC, F1-score e desvio padrão	Acurácia de 92% na detecção precoce usando fluxo óptico.	Arquitetura híbrida integrando dados clínicos e de imagem	Tamanho reduzido da amostra e ausência de validação externa
[Razzouki <i>et al.</i> 2024]	Deteção precoce de DP usando fluxo óptico e ViT	Processamento de vídeo com Vision Transformer	Precisão, recall e F1 em conjunto de testes separados.	Acurácia de 92% na detecção precoce usando fluxo óptico.	Primeira aplicação de ViT em dados de vídeo para PD	Elevado processamento e dependência de resolução elevada.
[Wang <i>et al.</i> 2024]	LSTM-CNN para detecção de DP por meio da escrita manual	Combinação LSTM e CNN aplicada a dados de escrita	Testes estatísticos (t-test) e análise de significância dos resultados	Acurácia de 88% em ambiente controlado de escrita.	Método híbrido para análise temporal de dinâmica de escrita	Cenário de laboratório não reflete escrita cotidiana.
[Khushbu <i>et al.</i> 2024]	Modelo híbrido CNN-Transformador para agnição de DP	Modelo híbrido CNN-Transformer	k-fold cross-validation om média e desvio padrão das métricas de acurácia	Acurácia média de 94.5%, com baixa variabilidade	Arquitetura híbrida demonstrou desempenho sólido na fusão multimodal.	Modelo volumoso, requer treinamento intensivo em GPU

Fonte: Próprio autor.

A Tabela 3 junta estudos recentes que aplicam técnicas de ML e DL para a detecção, segmentação e previsão de sintomas relacionados à DP, com destaque para abordagens híbridas e multimodais. Diversos trabalhos utilizaram arquiteturas combinadas, como *CNN – BiLSTM* [Chauhan e Ghosal 2024] e *CNN – Transformer* [Chen *et al.* 2024]; [Khushbu *et al.* 2024], evidenciando ganhos em acurácia, precisão e capacidade de captura de características temporais e espaciais, embora a limitação de dados clínicos amplos e amostras reduzidas ainda seja recorrente. Outras propostas, como o modelo baseado em *LightGBM* com seleção de atributos [Khanom *et al.* 2025], destacam-se pelo uso de técnicas de explicabilidade como *SHAP*, reforçando a transparência do processo decisório. Abordagens menos tradicionais, como o uso de *Vision*

*Transformers* em vídeo [Razzouki *et al.* 2024] e a análise de escrita manual com LSTM-CNN [Wang *et al.* 2024], demonstram acurácia superior a 88%, mas enfrentam desafios quanto à viabilidade prática em contextos cotidianos. Ainda, revisões sistemáticas como a de [Zhao *et al.* 2025] apontam a heterogeneidade dos *datasets* e a ausência de protocolos padronizados como fatores que dificultam a comparação entre estudos, enquanto [Sharma e Gupta 2024] reforça o potencial da integração de dados clínicos e de imagem na previsão de disfunções cognitivas, embora com limitações metodológicas relacionadas à validação externa. Em conjunto, os estudos revelam um avanço significativo na sofisticação dos modelos aplicados à DP, mas reforçam a necessidade de maior padronização, validação clínica e generalização dos resultados.

Tabela 4 – Aprendizado de máquina e Aprendizado profundo - Recursos e Aplicações Práticas dos Estudos

Área de Aplicação	Tipo de Dados	Dados Utilizados	Link para Código	Frameworks/ Tecnologias
Diagnóstico de Parkinson	Áudio, Imagem	Vários datasets revisados	-	TensorFlow, PyTorch
Neuroimagem médica	Ressonância magnética	Conjunto de imagens cerebrais	-	PyTorch
Detecção precoce de PD	Tabular (biomarcadores, voz)	Repositório de ML da UCI	-	LightGBM, SHAP
Avaliação de DP	Imagens médicas, voz, fala, batida de dedos e estabilidade postural marcha	Vários datasets revisados	-	-
Detecção de disfunção cognitiva	Imagens de ressonância e dados clínicos	Repositório de ML da UCI	-	-
Triagem em estágio inicial	Biomarcadores, vídeos MRI, registros audiovisuais	Vários datasets revisados	-	-
Análise de escrita à mão	Vídeo/IMU de escrita manual	DraWritePD (próprio), PaHaW	-	Pytorch
Reconhecimento de DP	Dados de ressonância magnética 3D	PPMI	-	MATLAB

Fonte: Próprio autor.

Na Tabela 4 se resumem as técnicas, recursos e aplicações práticas utilizados em estudos de aprendizado de máquina e profundo voltados à DP, referenciados na tabela 3. As abordagens cobrem uma ampla gama de dados, incluindo sinais de voz, imagens de RMI, biomarcadores tabulares, vídeos e registros de escrita manual. Conjuntos de dados públicos, como os da UCI e PPMI, bem como bases próprias, foram empregados em tarefas que variam do diagnóstico precoce à triagem cognitiva. Tecnologias como *TensorFlow*, *PyTorch*, *LightGBM*, *SHAP* e *MATLAB* foram aplicadas conforme o tipo de dado e a arquitetura proposta. Embora a diversidade de aplicações e fontes de dados evidencie a estudos na prática, a ausência de *links* de código e a limitação de *frameworks* descritos em alguns estudos ainda dificultam a reprodutibilidade e a adoção prática dos modelos.

### 3.3 Plataforma de Internet das Coisas Médicas

Tabela 5 –IoMT - Características Científicas e Analíticas dos Estudos

Autores	Objetivo do Estudo	Metodologia	Análise Estatística	Principais Resultados	Contribuições	Limitações
[Ayari <i>et al.</i> 2024]	Aplicações de e-health baseadas em IoMT para monitoramento remoto	Revisão bibliográfica e estudo de casos práticos IoMT	Qualitativa, comparativa.	Identificação dos principais sistemas IoMT, melhorias no monitoramento remoto	Síntese dos avanços e desafios da IoMT aplicada à e-health.	Ênfase qualitativa, com limitada análise quantitativa.
[He <i>et al.</i> 2025]	Propor um método para detecção precoce de DP usando fusão de múltiplos sensores.	Sensoriamento multimodal, processamento de sinais, fusão de sensores e ML.	Análise estatística descritiva e validação de modelos ML	AUC (0,891), sensibilidade (0,950) indicam vantagem sobre métodos unimodais.	Evidencia a eficácia da fusão multimodal diagnóstico precoce da DP.	Requer validação em larga escala e cenários clínicos reais.
[Goel e Nedunchelivan 2024]	Revisão de abordagens e limitações em IA e IoMT aplicadas à saúde.	Revisão sistemática e análise crítica de trabalhos existentes.	Análise qualitativa e quantitativa dos métodos revisados.	Mapeamento técnico da integração entre IA e IoMT.	Propõe caminhos para pesquisas futuras em saúde digital com IA.	Pouco foco em experimentação
[Huang <i>et al.</i> 2023]	Revisão sistemática sobre o estado da arte do IoMT	Revisão sistemática de literatura científica.	Estatísticas descritivas e qualitativas.	Revisão sobre avanços, aplicações e limitações do IoMT.	Síntese abrangente do estado da arte em IoMT.	Pouco foco nos dados próprios.
[Zhang <i>et al.</i> 2024]	Apresentar a plataforma Magnol.Ai para IoMT em saúde.	Desenvolvimento da plataforma e testes iniciais.	Avaliação qualitativa e quantitativa dos testes.	Plataforma funcional para coleta e análise de dados IoMT com AI.	Demonstração prática de integração IoMT e AI em saúde.	Testes limitados.

Fonte: Próprio autor.

A Tabela 5 engloba pesquisas que abordam a aplicação de IoMT na área da saúde, com ênfase em monitoramento remoto, detecção precoce e integração com IA. Trabalhos como os de [Ayari *et al.* 2024] e [Huang *et al.* 2023] realizaram revisões sistemáticas e qualitativas, sintetizando os avanços, aplicações e limitações da IoMT e o uso de tecnologias digitais e de informação na área da saúde (*eHealth*). Embora com pouca profundidade quantitativa. [He *et al.* 2025] apresentaram uma abordagem baseada na fusão de múltiplos sensores para a detecção precoce da DP, obtendo resultados promissores, como, (AUC 0,891; sensibilidade 0,950), mas ainda carecendo de validação em contextos clínicos reais, ou seja a dificuldade de validação clínica. Já [Goel e Nedunchelivan 2024] propuseram um mapeamento técnico da integração entre IA e IoMT, apontando lacunas e oportunidades futuras. Por fim, [Zhang *et al.* 2024] apresentaram a plataforma Magnol.Ai, demonstrando sua viabilidade na coleta e análise de dados IoMT com suporte de IA, embora os testes tenham sido limitados. No conjunto, os estudos destacam o potencial do IoMT na transformação digital da saúde, mas revelam a necessidade de validações robustas, maior uso de dados reais e foco em aplicações clínicas.

Já na Tabela 6 são apresentados os recursos e aplicações práticas adotados em estudos que exploram o uso da IoMT em contextos de saúde digital, informados na tabela 5. As áreas de aplicação incluem desde o monitoramento remoto e diagnóstico precoce da DP até o desenvolvimento de plataformas médicas inteligentes. Os estudos utilizaram predominantemente dados biométricos e sensoriais, como sinais vitais, movimentos, ECG, EEG, imagens e áudios, geralmente coletados por sensores vestíveis ou dispositivos IoMT. A maioria dos trabalhos empregou datasets próprios, o que evidencia um esforço de coleta personalizada, mas ao mesmo tempo limita a replicabilidade e a comparação entre estudos. Apenas um dos trabalhos mencionou

Tabela 6 – IoMT - Recursos e Aplicações Práticas dos Estudos

Área de Aplicação	Tipo de Dados	Dados Utilizados	Link para Código	Frameworks/ Tecnologias
Saúde eletrônica e monitoramento remoto	Dados biométricos e sinais vitais via IoMT	Dados biométricos, sinais vitais, sensores, imagens áudios	-	-
Diagnóstico precoce da DP	Sensores biomédicos (sinais vitais, movimentos)	Dataset próprio	-	-
Plataformas médicas inteligentes	Dados de ECG/EEG e outros dados de sensores	Dataset próprio	-	-
Revisão sistemática IoMT	Dados biomédicos e sensoriais	Dataset próprio	-	-
Plataforma IoMT para saúde	Dados biométricos coletados via sensores e vestíveis IoMT	Dataset próprio	-	Serviços AWS, Jupyter Notebook

Fonte: Próprio autor.

explicitamente o uso de *frameworks* ou serviços tecnológicos, no caso, a integração com *Amazon Web Services* (AWS) e o uso do *Jupyter Notebook*. De modo geral, a ausência de *links* para código-fonte e a escassez de informações detalhadas sobre as tecnologias empregadas reforçam a necessidade de maior transparência e padronização nas aplicações práticas de IoMT em saúde.

## 4 METODOLOGIA

Este capítulo apresenta um *framework* baseado em agentes autônomos que utilizam pequenos modelos de linguagem para prever a DP. A abordagem proposta distingue-se pela sua natureza interativa, onde grupos de agentes, cada qual com uma especialidade e potencializada por diferentes SLMs, colaboram para formar um diagnóstico e o gerenciamento da doença. Diferentemente dos processos decisórios convencionais, a exemplo da votação por maioria ou pontuação, o método clínico proposto é intrinsecamente customizado para se adaptar à heterogeneidade dos dados multimodais. O poder analítico do sistema é amplificado pelo uso de tecnologias de ponta em recuperação de informação, incluindo Multimodal RAG, *KnowledgeRAG*, *GraphRAG* e *AgenticRAG*, com o objetivo final de tornar a predição da doença mais acurada e confiável.

### 4.1 Arquitetura

A arquitetura proposta nesse trabalho será dividida em três grandes seções: *DataOps* - explora uma visão das operações que abrange todos os estágios dos dados (áudio, texto, imagem, tabular - temporal ou não temporal) da plataforma IoMT, como o pré-processamento e processamento de dados em tempo real, *MLOps/LLMOps/VLMOps* - explorando todo o contexto operacional, como o treinamento, deploy e inferência de modelos tradicionais de ML, mas também como os SLMs, LLMs e Modelos de Linguagem de Visão (VLMS). Ao final, uma visão geral sobre os serviços utilizados que sustentam toda a infraestrutura proposta.

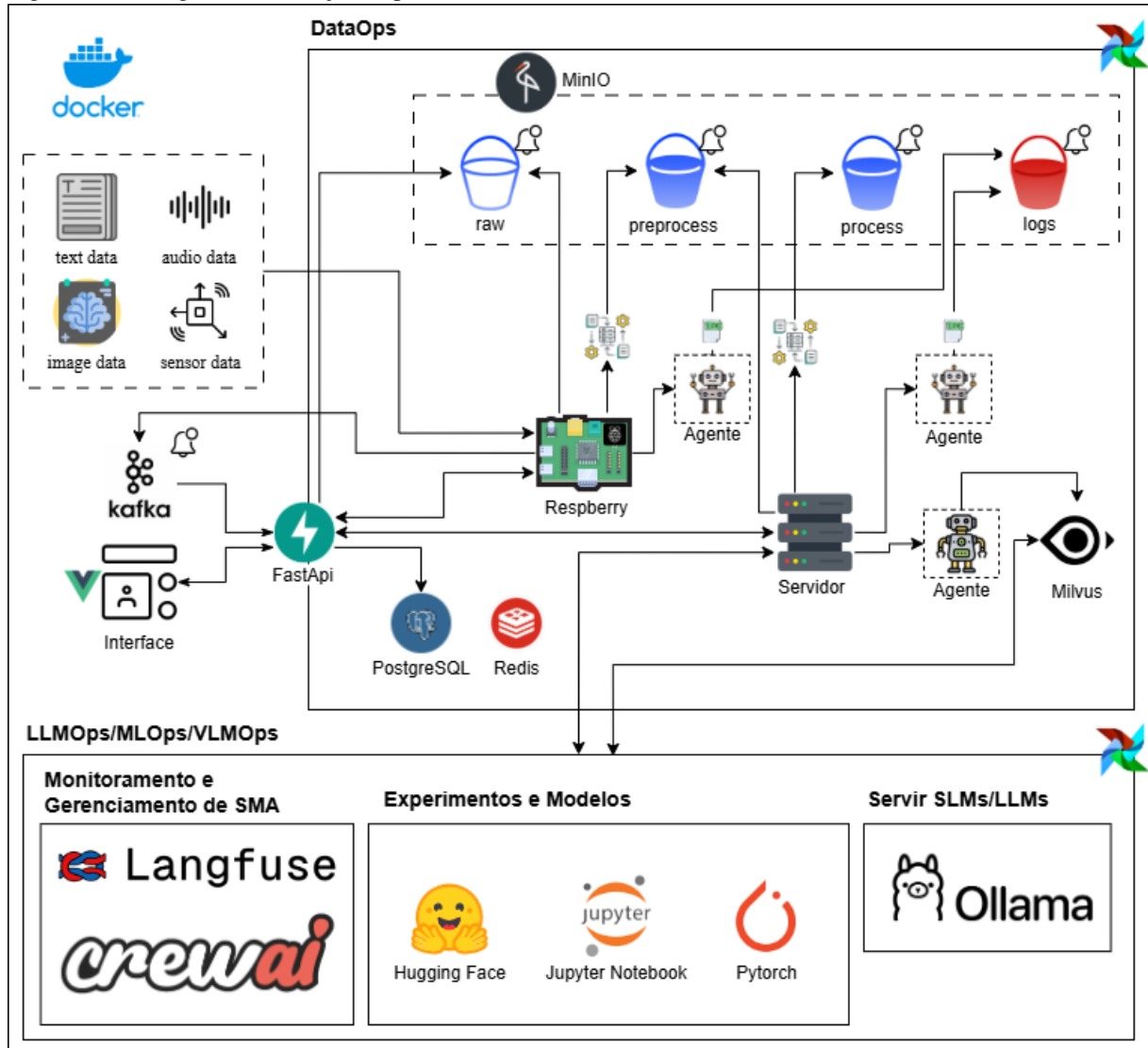
#### 4.1.1 Serviços

A concepção da nossa plataforma foi orientada para atender a dois requisitos operacionais fundamentais em um ecossistema de IoMT: o processamento de dados em tempo real e o acesso concorrente para múltiplos usuários. A arquitetura suporta um fluxo de dados sequencial, que parte dos dispositivos de aquisição de dados (sensores, vestíveis e sistemas externos) → dispositivo de borda (*Raspberry Pi 4 Model B*) → computação robusta e armazenamento na nuvem (máquina principal). Em paralelo, a plataforma deve permitir que múltiplos usuários, como, pesquisadores, profissionais de saúde e pacientes, interajam com os dados e serviços de forma simultânea.

A infraestrutura de serviços proposta para materializar esta visão está represen-

tada esquematicamente na Figura 6. A seguir, detalhamos a abordagem técnica adotada para implementar e orquestrar os componentes apresentados.

Figura 5 – Visão geral dos serviços da plataforma IoMT.



Fonte: Próprio autor.

Nossa plataforma foi projetada para ser um ecossistema completo e modular para análise multimodal e gerenciamento personalizado para a DP - podendo ser estendido a outras patologias - capaz de gerenciar o ciclo de vida dos dados desde a sua aquisição até a inferência por sistemas multi-agentes autônomos de IA. A estrutura é dividida em duas áreas funcionais principais, Operações nos Dados (DataOps) e *Machine Learning Operations* (MLOps)/*Large Language Models Operations* (LLMOps)/*Vision Language Models Operations* (VLMOps), todas operando sobre uma base containerizada com *Docker*, garantindo reprodutibilidade, modularidade e portabilidade.

#### 4.1.1.1 Ingestão de Dados e Gateway de Entrada

O ponto de partida da plataforma é a camada de aquisição de dados, que é heterogênea por natureza, refletindo a complexidade de fontes em um ambiente de IoMT:

1. Fontes de Dados: Incluem sensores sem fio, dispositivos vestíveis (como smartwatches), dados de entrada manual (representados pelo ícone da caneta) e formulários clínicos.
2. Interface de Usuário: Além das fontes automatizadas, uma interface permite a interação direta do usuário, seja para inserção de dados, seja para o envio de consultas ao sistema.
3. Apache Kafka: Todos esses fluxos de dados, contínuos ou discretos, são canalizados para o Apache Kafka, que atua como um barramento de mensagens, *message bus*, distribuído e tolerante a falhas. A utilização do Kafka é estratégica para desacoplar os produtores de dados dos consumidores, permitindo o processamento assíncrono e garantindo a escalabilidade e a resiliência da ingestão de dados em tempo real.
4. FastAPI: O componente FastAPI funciona como o principal gateway de API da plataforma. Ele consome as mensagens do Kafka e expõe endpoints para a interface de usuário. É responsável por receber todas as requisições externas, validá-las e orquestrar as chamadas para os serviços de backend apropriados no pipeline de DataOps.

#### 4.1.1.2 Pipeline de Processamento e Armazenamento (DataOps)

Uma vez que os dados são recebidos pela API, eles entram no pipeline de DataOps, onde são processados, transformados e armazenados de forma estruturada:

- MinIO (*Data Lake*): O MinIO serve como nosso *data lake*, um repositório de objetos compatível com o padrão S3. O fluxo de dados dentro do MinIO é organizado em estágios, representados por *buckets*:
  - *raw*: Armazena os dados brutos e inalterados, exatamente como foram adquiridos dos dispositivos de aquisição, considerados os dados do tipo bronze.
  - *preprocess*: Contém dados que passaram pela pipeline automática de pré-processamento realizado pelo dispositivo de borda (*Raspberry Pi 4 Model B*) e foram validados pelos agentes de pré-processamento, considerados dados do tipo prata.
  - *process*: Armazena os dados finais, pós-validação dos agentes de processamento, enriquecidos e prontos para serem consumidos pelos modelos de IA ou para análise.
- Agentes de (Pré)Processamento: Os "robôs" no diagrama representam agentes de pro-

cessamento automatizados. São agentes baseados em modelos de linguagem de visão, capazes de extrair informações gerais ou específicas a partir de uma imagem de *input*, como o pré-processamento é automatizado e heterogêneo, eles são responsáveis por extrair informações a partir do aspecto visual das imagens.

– Bancos de Dados:

- PostgreSQL: Atua como o banco de dados relacional do sistema. Ele armazena metadados essenciais: informações sobre os arquivos no MinIO, *logs* de execução dos agentes, resultados de experimentos, dados de usuários e outras informações estruturadas que governam a plataforma.
- Redis: Utilizado como um banco de dados em memória de alta velocidade. Sua função na arquitetura é servir como um cache para dados frequentemente acessados, além da necessidade de funcionamento com MinIO e airflow.
- Milvus: Utilizado como um banco de dados vetorial de código aberto de alto desempenho. Em nossa infraestrutura, ele atua para armazenar os dados dos pacientes e os dados da plataforma (como: áudio, imagens, tabulares (temporal e atemporal), textuais) para consultas rápidas usando RAG.

#### 4.1.1.3 Ciclo de Vida de Modelos de IA (LLMOps/MLOps/VLMOps)

Esta camada contém o conjunto de ferramentas para pesquisa, desenvolvimento, monitoramento e serviço dos modelos de Inteligência Artificial:

Experimentos e Modelos: o ambiente de desenvolvimento é composto por:

- *Jupyter Notebook*: Para experimentações interativas e prototipagem de modelos.
- *PyTorch*: O *framework* de DP utilizado para construir e treinar os modelos.
- *Hugging Face*: A plataforma e biblioteca que fornecem acesso a modelos pré-treinados, datasets e ferramentas essenciais para o ecossistema de *Transformers*.

Servir SLMs/LLMs (*Model Serving*): O componente *Ollama* é responsável por servir os modelos de linguagem (pequenos ou grandes). Ele encapsula os modelos treinados ou ajustados e os expõe através de uma API, permitindo que sejam facilmente consumidos para inferência pelo FastAPI ou pelos agentes de processamento.

Monitoramento e Gerenciamento de SMA (Sistemas Multiagentes): A subseção mais avançada, responsável pelo monitoramento e gerenciamento dos agentes inteligentes e orquestração deles para tomada de decisão inteligente e colaborativa.

- CrewAI: É um framework para orquestrar múltiplos agentes de IA autônomos, permitindo que eles colaborem para resolver tarefas complexas.
- Langfuse: É uma plataforma de observabilidade para aplicações baseadas em modelos de linguagem. Ela fornece ferramentas para rastrear, *tracking*, depurar e analisar as interações e o desempenho dos agentes e dos modelos de linguagem, o que é fundamental para garantir a confiabilidade e a qualidade das respostas geradas.

Portanto, a arquitetura proposta busca integrar o fluxo, desde a ingestão de dados em tempo real, um pipeline de processamento robusto até um ciclo de vida de *MLOps/LLMOps/VLMOps* completo. A separação de responsabilidades e a comunicação orquestrada através do *airflow* permitem um sistema escalável, manutenível e altamente capacitado para a pesquisa e aplicação de IA em cenários complexos de IoMT.

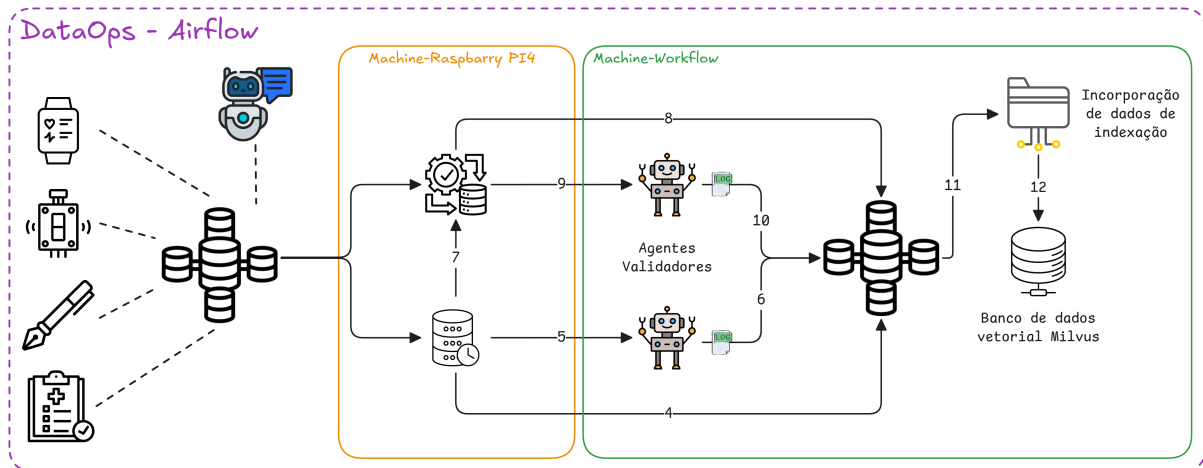
#### 4.1.2 *DataOps*

Segundo o autor Ereth *et al.*, o termo *DataOps* ou *Data Operations*, Operações nos Dados, apresentado em seu trabalho, oferece uma perspectiva integrada e orientada a processos sobre o tratamento automatizado de dados [Ereth 2018]. Todo esse processo do dado na plataforma proposta é feito baseado na aquisição do artefato, dado, em diversos dispositivos com comunicação e transferência de dados de baixa latência, configurando-se no nosso sistema IoT, bem como também o uso de informações baseadas em feedbacks diários informados por usuários com permissões de especialista, bem como fontes externas. A seção de dados da plataforma oferece suporte à subseção referente ao sistema com múltiplos agentes baseados em LLMs, SLMs e RAG, fornecendo acesso aos dados ouro, pós-validação realizada pela dupla de agentes validadores. Ao longo das etapas do dado, realiza-se uma série de transformações e, em cada etapa, o armazenamento e, por fim, validação autônoma com presença de *logs*. A Figura 6 apresenta em detalhes a arquitetura de *DataOps* proposta para a plataforma IoMT.

#### 4.1.3 *MLOps/LLMOps/VLMOps*

O termo *LLMOps* emergiu muito recentemente principalmente com a popularização da IA Gerativa e, por isso, existem poucas pesquisas e menos ainda as que analisam sistematicamente *LLMOps*. Assim, como *DataOps*, citado anteriormente, o novo termo consiste nas operações fundamentais usando LLMs [Diaz-De-Arcaya *et al.* 2024]. O *LLMOps* visa uma aplicação prática dos LLMs a fim de revolucionar a indústria de IA, garantindo o funcionamento

Figura 6 – Arquitetura de DataOps da plataforma IoMT proposta.



Fonte: Próprio autor.

e a atualização eficiente do ciclo de vidas dos LLMs, através da integração contínua/entrega contínua (CI/CD). Um estudo ainda mais recente realizado por [Pahune e Akhtar 2025], analisa possíveis semelhanças entre as operações usadas em *MLOps* com *LLMOps* e, com isso, mapeia todas as operações que possam surgir em decorrência da popularização do uso de LLMs em contextos de produção.

O ciclo de vida dos LLMs depende de conceitos fundamentais do *LLMOps* que se baseiam fundamentalmente na seleção de modelos, ajuste de modelos, implantação, engenharia de prompts e monitoramento. No trabalho de [Shan e Shan 2024], é definido os principais elementos do *LLMOps* organizados em fases, como: Descobrir, Destilar, Implantar e Entregar (4D). Descoberta se trata do estágio inicial, consiste no estudo da necessidade do uso de LLMs no projeto/área, analisando o consumo de recursos que será empregado para o desenvolvimento, preparação de base de dados e arquitetura da aplicação. Destilação consiste na etapa de transformação dos dados, visando refinar os dados para melhorar o desempenho ou reduzir o custo computacional do modelo através de técnicas de redução de uso de memória como *Quantization* e treinamento inicial do modelo. Implantação é a fase na qual o modelo será colocado a infraestrutura, tornando-se usável em larga-escala, mapeando informações como: versionamento do modelo, integrações com API, sistema(s) RAG e banco de dados vetorial usado, entre outros. Essa etapa deve ser analisada com veracidade os *softwares* e *hardwares* da máquina que garantirá que o modelo seja escalável, integrável e seguro. Por fim, a entrega, esse é o estágio final do ciclo de vida dos LLMs, focado no valor agregado gerado pelo LLM. Isso envolve monitorar seu desempenho em aplicações reais, promovendo melhorias contínuas com base em *feedbacks* e dados atualizados, além de assegurar que o modelo acompanhe as necessidades em constante

evolução da organização. Também é fundamental avaliar o impacto do LLM nos resultados de negócio e na experiência dos usuários, realizando os ajustes necessários para maximizar seu potencial e eficácia.

No entanto, a configuração de LLMs em produção ainda oferece grandes desafios significativos, principalmente na área da saúde. Tópicos como relacionados a escalabilidade de dados, devido a geração de milhares de dados multimodais gerados por dispositivos em tempo real, segurança de informações médicas que demonstra a importância da integração de *DevSecOps* e *LLMOps*, garantindo segurança compartilhada ao longo das camadas de desenvolvimento e uso da plataforma.

O uso de *LLMOps* na plataforma proposta inicia-se após a validação dos dados baseada em agentes. Uma vez que os dados são validados, são caracterizados como dados padrão ouro, são aqueles dados que passaram pela *pipeline* de *DataOps*, inicialmente coletados por sensores, pre-processados na borda e processados por uma máquina mais robusta.

Por fim, surgem também uma nova premissa de operações muito semelhante ao *LLMOps*, mas que diferem na tarefa final. O *LLMOps* surge com os novos modelos de linguagem do estado da arte, modelos estes baseados em aspectos da visão, como vídeo e imagens, como o *VLMOps*.

#### 4.1.3.1 Latência Clínica e Eficiência Computacional

A implantação de modelos inteligentes em ambientes clínicos conectados demanda atenção especial à latência de resposta e ao custo computacional dos sistemas [Ali-Eldin *et al.* 2021]. Em aplicações médicas, especialmente aquelas baseadas em monitoramento contínuo ou tomada de decisão em tempo real [Rancea *et al.* 2024], como é o caso da DP, atrasos de milissegundos podem impactar diretamente a usabilidade, confiabilidade e aceitação clínica das soluções [Liu *et al.* 2023].

Neste trabalho, estratégias de compressão e quantização, como *QLoRA* em 4 bits e o uso de SLMs, foram aplicadas com o objetivo de permitir a execução de agentes baseados em LLMs em dispositivos com restrições computacionais, como a *Raspberry Pi 4 Model B*, [Basit *et al.* 2024]. A compressão dos modelos reduz significativamente o tamanho e o custo de inferência, permitindo sua execução em ambientes de borda, *edge computing*, sem perdas significativas de desempenho, possibilitando maior autonomia computacional no ponto de aquisição dos dados [Shen *et al.* 2025]. Portanto, em ambientes com infraestrutura robusta,

como sistemas hospitalares, é possível manter maior complexidade local, incluindo modelos maiores parcialmente quantizados como LLMs ou VLMs [Zhang *et al.* 2024]; [?]

Para lidar com o *trade-off* entre complexidade e latência, foram aplicadas estratégias híbridas que distribuem a carga computacional entre o *edge*, dispositivo local, e a *cloud*, servidor remoto, [Rajagopal e Subramanian 2025]. Em cenários com restrições de conectividade ou recursos computacionais limitados, opta-se pela extração de características no *edge* e envio dos *embeddings* para processamento leve na nuvem [Gawade *et al.* 2025]. Por outro lado, em ambientes com infraestrutura mais robusta. Como hospitais, é possível realizar tanto a extração quanto a inferência local, reduzindo significativamente a latência total do sistema [Islam *et al.* 2025].

Essa arquitetura é fortemente influenciada pelos princípios da Internet da Saúde das Coisas (IoHT) [Khatun *et al.* 2023], incluindo tolerância a falhas, computação consciente de energia e resiliência a redes inconstantes [Kumar *et al.* 2025]. Mecanismos de *cache* local e retransmissão de pacotes foram incorporados para garantir que interrupções na conectividade não afetem a integridade nem a continuidade do fluxo clínico [Sathupadi *et al.* 2024].

#### **4.1.4 Interface gráfica**

A interface gráfica do usuário (GUI) do sistema desenvolvido foi concebida para proporcionar uma experiência intuitiva e eficiente, fundamental para a interação do usuário com as ferramentas de análise de dados biomédicos. A escolha das tecnologias e o design da interface foram guiados pelos princípios de usabilidade, responsividade e manutenibilidade.

##### **4.1.4.1 Frontend**

O frontend da aplicação foi construído utilizando Vue.js em conjunto com Vite. Esta combinação tecnológica foi selecionada devido a uma série de vantagens estratégicas para o desenvolvimento de sistemas complexos e de alto desempenho:

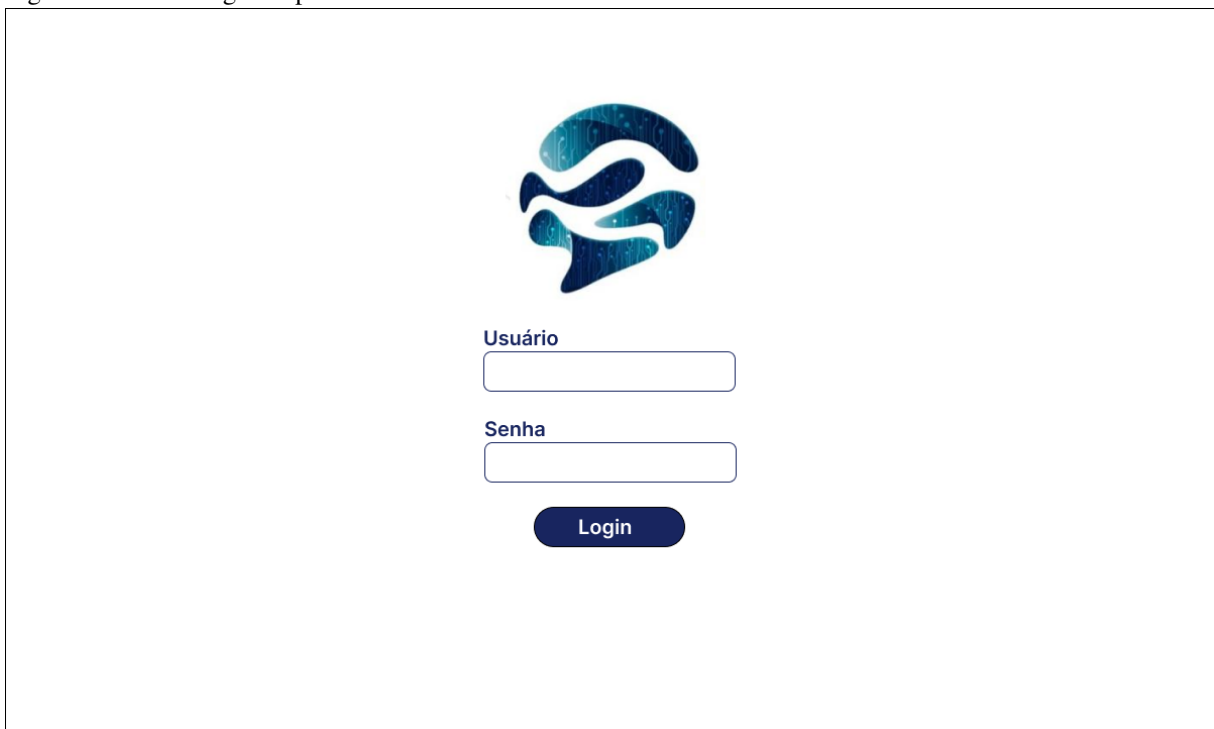
- Vue.js: A escolha do Vue.js como *framework JavaScript* principal foi motivada por sua curva de aprendizado suave e sua natureza progressivamente adaptável. Isso permitiu um desenvolvimento ágil e a construção de componentes reutilizáveis, essenciais para a modularidade e escalabilidade da interface. Sua reatividade granular otimiza a performance, garantindo que a interface se atualize eficientemente com as mudanças nos dados, um aspecto crítico em aplicações de análise.

- Vite: O Vite foi empregado como bundler e servidor de desenvolvimento devido à sua velocidade excepcional no desenvolvimento e na construção (hot module reloading - HMR). Essa característica agilizou significativamente o ciclo de desenvolvimento, permitindo iterações rápidas e reduzindo o tempo de espera do desenvolvedor. A eficiência do Vite na otimização de assets para produção também contribuiu para um carregamento rápido da aplicação final, melhorando a experiência do usuário.

A integração de Vue.js com Vite resultou em um ambiente de desenvolvimento robusto e performático, ideal para a criação de interfaces de usuário dinâmicas e responsivas.

A interface foi projetada com um foco na clareza e facilidade de uso, refletindo as necessidades dos pesquisadores e usuários que interagem com dados biomédicos. As telas apresentadas ilustram os pontos de entrada e a navegação principal do sistema.

Figura 7 – Tela de login da plataforma IoMT

A imagem mostra a tela de login da plataforma IoMT. No topo centralizado, há um logotipo abstrato em tons de azul e verde, com formas fluidas e orgânicas. Abaixo do logotipo, há dois campos de entrada de texto: o primeiro é rotulado "Usuário" e o segundo "Senha". Ambos os campos são retangulares com bordas arredondadas e uma borda azul. Abaixo dos campos, há um botão de "Login" com fundo azul escuro e o texto "Login" em branco, centralizado.

Fonte: Próprio autor.

Esta é a primeira interação do usuário com o sistema, apresentando um design limpo e direto. A simplicidade visual reduz a distração e direciona o foco para os campos de "Usuário" e "Senha", garantindo um acesso seguro e eficiente à aplicação. O logotipo distintivo da aplicação é proeminentemente exibido, reforçando a identidade visual.

Após o login, o usuário é direcionado para a tela principal, referenciada na imagem abaixo, que serve como o dashboard central.

Figura 8 – Tela de home plataforma IoMT



Fonte: Próprio autor.

A imagem de home conta com:

- Navegação Lateral ("Menu" e "Modos"): O ícone de menu (hambúrguer) e o item "Modos" sugerem uma estrutura de navegação clara, permitindo ao usuário alternar entre diferentes funcionalidades ou módulos da aplicação. Isso indica uma arquitetura modular, onde diferentes ferramentas podem ser acessadas de forma organizada.
- Identidade Visual ("MindAI" e Logotipo com "Biomedical Data Analytics Research Group"): O nome do sistema, "MindAI", e o logotipo com o texto "Biomedical Data Analytics Research Group" reforçam o propósito e a autoria do sistema, transmitindo profissionalismo e afiliação acadêmica.
- Área de Interação Principal ("Pergunte ao MindAI" e Ícone de Microfone): A presença de um campo de entrada com a *placeholder* "Pergunte ao MindAI" e um ícone de microfone sugere fortemente uma capacidade de interação conversacional ou de comando por voz. Esta funcionalidade, se implementada, representa um avanço significativo na usabilidade, permitindo aos usuários realizar consultas ou tarefas de forma mais natural e eficiente. O botão "+" ao lado do campo de entrada pode indicar a adição de novas consultas, configurações, ou a expansão das opções de interação.

#### 4.1.4.2 Backend

A comunicação entre o *frontend* (Vue.js) e o *backend* é estabelecida por meio de uma API RESTful, desenvolvida com *FastAPI*. A escolha do *FastAPI* foi estratégica por sua capacidade de oferecer:

1. **Alta Performance:** FastAPI é notável por seu desempenho excepcional, comparável ao Node.js e Go, o que é crucial para lidar com processamento de dados biomédicos que podem ser intensivos em recursos.
2. **Velocidade de Desenvolvimento:** A utilização de anotações de tipo Python e a geração automática de documentação OpenAPI (Swagger UI) e ReDoc aceleram o desenvolvimento da API, facilitando a comunicação entre o *frontend* e o *backend*.
3. **Validação de Dados:** A integração com Pydantic para validação de dados garante a integridade e segurança das informações trocadas entre cliente e servidor, minimizando erros e garantindo a robustez do sistema.

A combinação de Vue.js/Vite no frontend e FastAPI no backend resultou em uma arquitetura robusta, performática e escalável, capaz de suportar as demandas de um sistema de análise de dados biomédicos de alto nível.

## 4.2 Retrieval Augmented Generation (RAG)

Técnicas de Recuperação e Geração estão sendo amplamente utilizadas para tarefas de PNL. A técnica RAG inovou a precisão e acurácia das respostas dos LLMs com a recuperação de dados externos ao seu conhecimento base, evitando, dessa forma, *fine-tuning*, *transfer-learning* e outros métodos de treinamento que demandam um alto custo computacional [Lewis *et al.* 2020]. A arquitetura RAG é composta por dois componentes: recuperador e gerador. O recuperador se baseia no DPR [Karpukhin *et al.* 2020], que segue uma arquitetura *biencoder*, composta pelos componentes: *document encoder* e *query encoder*. Em ambos os componentes é usado um modelo para gerar representações vetoriais dos textos, para cada documento é calculado uma probabilidade que define a relevância dos  $n$  documentos dado uma consulta. Enquanto que o componente gerador é uma arquitetura *encoder-decoder*, no qual combina um *input*  $x$  com os documentos recuperados  $w$  e retorna o texto gerado.

### 4.2.1 RAG Multimodal

Novos desafios surgiram com a popularização do RAG tradicional ao tentar utilizar a técnica em dados não usuais, como dados médicos (Ressonância Magnética por Imagem (RMI), Tomografia Computadorizada (TC), ultrassom, bio, anamneses audiológicas), séries temporais, áudios, vídeos, entre outros. Para isso, diversos trabalhos propuseram uma abordagem envolvendo um RAG Multimodal [Chen *et al.* 2022, Zhao *et al.* 2023, Yasunaga *et al.* 2022].

#### 4.2.1.1 Embedding Multimodal

Assim como o RAG normal necessita de um modelo que gera representações vetoriais de textos, ou melhor, *embeddings* de textos, modelos multimodais geram representações vetoriais de dados de diversos tipos, como o modelo *CLIP* [Radford *et al.* 2021] da *OpenAI* que gera representações vetoriais a partir de imagens ou até o *Wav2CLIP* [Wu *et al.* 2022] que gera *embeddings* vetoriais de áudios/vídeos.

#### 4.2.1.2 Busca de Resumos

A abordagem de recuperação de busca por resumos ou, *summaries search*, em um contexto multimodal é simplesmente quando todos os seus dados não textuais são codificados do seu tipo natural para um tipo textual através de um LLM multimodal, gerando um metadado único para o seu dado. A partir dessa *caption* para o seu dado, cria-se uma representação vetorial deste metadado usando um modelo comum de *embedding* textual, como o *BERT* [Devlin 2018].

#### 4.2.1.3 Recuperação Híbrida

A abordagem de recuperação híbrida ou, *hybrid search*, é também uma importante abordagem eficaz usada para lidar com dados de diversas naturezas. A recuperação híbrida utilizam os mesmos mecanismos que a abordagem de um modelo de *embedding* multimodal, mas que combina as probabilidades de recuperação entre os *embeddings* de textos + geração de sumários de qualquer outro tipo de dados, áudio, vídeo, imagens, garantindo uma maior probabilidade de recuperar o dado correto, dado uma entrada. A geração de sumários são obtidos a partir de LLMs multimodais. Estes, basicamente, convertem um tipo dado para outro, como  $f(x) \rightarrow f(y)$ , onde  $f(x)$  é uma consulta, um texto do usuário, e  $f(y)$  é a saída do modelo, gerando um outro tipo, um vídeo, por exemplo. Por fim, a recuperação híbrida cria representações

vetoriais de todos esses dados, codificados em um tipo textual, juntamente com a sua forma bruta que são armazenados em um *Multi-Vector Retriever*, o seu texto como metadado de sua forma bruta e, por fim, usa-se um LLM multimodal que recebe como parâmetro tanto sua forma bruta quanto o sumário do dado.

#### **4.2.2 Graph RAG**

Com a constante evolução das tecnologias baseadas em recuperação para LLMs, uma nova abordagem de recuperação baseada em grafos surge, o *GraphRAG*. Segundo [Edge *et al.* 2024], permite a construção de sentido sobre a totalidade de um grande corpus de texto. Para tal, usa-se um LLM para construir um gráfico de conhecimento, onde os nós correspondem a entidades-chave no corpus e as bordas representam relacionamentos entre essas entidades. Em seguida, ele particiona o gráfico em uma hierarquia de comunidades de entidades intimamente relacionadas, antes de usar um LLM para gerar resumos no nível da comunidade. Esses resumos são gerados de baixo para cima, seguindo a estrutura hierárquica das comunidades extraídas, com resumos em níveis mais altos da hierarquia incorporando recursivamente resumos de nível inferior. Juntos, esses resumos da comunidade fornecem descrições globais e *insights* sobre o corpus. Finalmente, o *GraphRAG* responde a consultas por meio do processamento de redução de mapa de resumos da comunidade; na etapa Mapear, os resumos são usados para fornecer respostas parciais à consulta de forma independente e em paralelo, em seguida, na etapa Redução, as respostas parciais são combinadas e usadas para gerar uma resposta global final. Além disso, esse método é usado também no contexto médico, criando uma comunidade de grafos médicos para potencializar a recuperação das informações, como pontuado por [Wu *et al.* 2024] em seu trabalho.

#### **4.2.3 AgenticRAG**

O *Agentic Retrieval-Augmented Generation*, ou *Agentic RAG*, representa uma evolução significativa em relação aos paradigmas tradicionais de RAG ao incorporar agentes autônomos dotados de capacidade decisória dinâmica, adaptabilidade em tempo real e otimização contínua de fluxos de trabalho. Essa abordagem visa lidar com ambientes altamente dinâmicos e tarefas de complexidade variável por meio da modularização do processo de recuperação e geração, integrando ciclos iterativos de *feedback* e estratégias adaptativas de consulta. Segundo [Singh *et al.* 2025], em seu trabalho pontua as principais características desse novo

paradigma:

- Tomada de decisão autônoma: Agentes baseados em LLMs, ou SLMs avaliam a complexidade semântica da consulta e selecionam, de forma autônoma, estratégias de recuperação, podendo escolher entre múltiplos retrievers, pipelines e formas de condensação de contexto.
- Refinamento iterativo: O sistema opera em ciclos de *self-refinement*, onde respostas preliminares são avaliadas e aprimoradas com base em critérios semânticos, sintáticos ou contextuais, promovendo melhorias contínuas na relevância e precisão das respostas.
- Otimização de *workflow*: Os agentes são capazes de orquestrar dinamicamente tarefas como extração, recuperação, pré-processamento e geração, priorizando a eficiência computacional e a adaptabilidade em tempo real, especialmente em ambientes multi-domínio.

Esses elementos tornam o *Agentic RAG* particularmente promissor para aplicações em cenários abertos, como agentes de diálogo autônomos, sistemas de apoio à decisão médica e assistentes empresariais com múltiplas fontes de dados heterogêneas.

### *Desafios e limitações*

Apesar de seu potencial inovador, o paradigma *Agentic RAG* apresenta desafios técnicos substanciais:

- Alta complexidade de coordenação: A gestão das interações entre agentes exige mecanismos de orquestração sofisticados, frequentemente baseados em políticas hierárquicas ou em sistemas multiagentes com papéis diferenciados.
- Sobrecarga computacional: A execução simultânea de múltiplos agentes, cada um possivelmente envolvendo um LLM, impõe demandas significativas de recursos computacionais, especialmente em *workflows* com ciclos iterativos de decisão e recuperação.
- Limitações de escalabilidade: Embora teoricamente escalável, o sistema sofre com gargalos de desempenho em cenários de alta concorrência, dada a necessidade de constante reavaliação de estratégias e acesso simultâneo a bases de dados extensas.

Tais desafios indicam que o sucesso do *Agentic RAG* depende não apenas da evolução dos modelos base, mas também de avanços em técnicas de orquestração, compressão de contexto e eficiência em modelos de linguagem.

### 4.3 Multi-Collaborative Agent Systems (MCAS)

Os sistemas de (MCAS), ou no português, sistemas de múltiplos agentes colaborativos, representam uma evolução significativa na inteligência artificial, permitindo que múltiplos agentes, cada um potencializado por modelos de linguagens como Linguagens de Larga Escala (LLMs) Pequenos Modelos de Linguagem (SLMs), também por modelos de visão como VLMS e modelos comuns de DP/ML, pouco explorados devido à falta de capacidade de interagir através de linguagem natural entre um agente e outro. A principal funcionalidade desses sistemas é colaborar em conjunto, de alguma forma, para resolver tarefas reais complexas. A seguir, serão abordados os diferentes usos de Sistemas Multi-Agentes baseados em LLMs/SLMs, sistemas unimodal, bimodal e multimodal no diagnóstico e gerenciamento precoce da DP a partir de sistema *Internet of Things* (IoT) integrado à plataforma.

#### *Unimodal*

Na abordagem unimodal, os agentes operam exclusivamente em uma única modalidade de dados, como texto. Cada agente é especializado em uma sub tarefa específica, e a colaboração entre eles visa melhorar a eficiência e a precisão na resolução de problemas complexos.

- SMOA (*Sparse Mixture-of-Agents*): Este *framework* propõe a utilização de múltiplos agentes especializados, cada um treinado para tarefas específicas, colaborando de forma eficiente para melhorar a qualidade das respostas geradas por modelos de linguagem [Li *et al.* 2024].
- *Multi-Agent Collaboration Mechanisms*: Este estudo fornece uma análise abrangente dos mecanismos de colaboração em sistemas multiagentes baseados em LLMs, introduzindo um *framework* extensível para orientar pesquisas futuras. [Tran *et al.* 2025]

#### *Bimodal*

A abordagem bimodal envolve agentes que operam em duas modalidades de dados, como texto e imagem. Essa integração permite uma compreensão mais rica e contextualizada das informações, melhorando o desempenho em tarefas que requerem múltiplas fontes de dados.

- *Cooperative Sentiment Agents* (Co-SA): Este modelo propõe agentes de sentimento cooperativos que interagem entre si para analisar sentimentos em dados multimodais, como

texto e imagens, melhorando a precisão na análise de sentimentos. [Wang *et al.* 2024]

### **Multimodal**

Na abordagem multimodal, os agentes são capazes de processar e integrar informações de múltiplas modalidades, como texto, imagem, áudio e vídeo. Essa capacidade é crucial para aplicações que exigem uma compreensão holística de diferentes tipos de dados.

- *LMAgent*: Este estudo apresenta uma sociedade de agentes multimodais em larga escala, onde os agentes podem interagir em ambientes simulados, realizando tarefas como compras online e transmissões ao vivo, demonstrando comportamentos sociais complexos. [Liu *et al.* 2024]
- *SpeechAgents*: Este sistema multimodal simula a comunicação humana utilizando agentes baseados em LLMs que interagem por meio de sinais multimodais, como fala e texto, para gerar diálogos consistentes e emocionalmente ricos. [Zhang *et al.* 2024]
- *Beyond Text*: Este estudo propõe a implementação de sistemas multiagentes multimodais baseados em LLMs utilizando uma plataforma sem código, facilitando a adoção de IA em ambientes empresariais. [Jeong 2025]

#### **4.4 Tomada de decisão clínica**

A tomada de decisão em sistemas multiagentes é um fator determinante para o desempenho do método, dada a complexidade da interação entre os agentes. Um estudo recente conduziu um levantamento experimental envolvendo diferentes abordagens de decisão [Cho *et al.* 2024], entre as quais se destacam principalmente:

1. Unanimidade: Uma proposta só é aceita se receber o apoio de todos os agentes. Este é o critério mais rigoroso possível. Garante um consentimento total e pode ser desejável em cenários de alta criticidade. No entanto, sua principal desvantagem é a propensão a impasses, *gridlock*, onde nenhuma proposta consegue o apoio necessário, levando a uma baixa eficiência e performance, especialmente nas rodadas iniciais de colaboração.
2. Maioria: Uma proposta é aceita se receber votos de mais da metade dos agentes. É um mecanismo extremamente comum devido à sua simplicidade conceitual. Contudo, estudos mostram que este limiar rígido de 50% pode levar a uma colaboração ineficiente, descartando propostas que têm um apoio substancial, mas não majoritário.

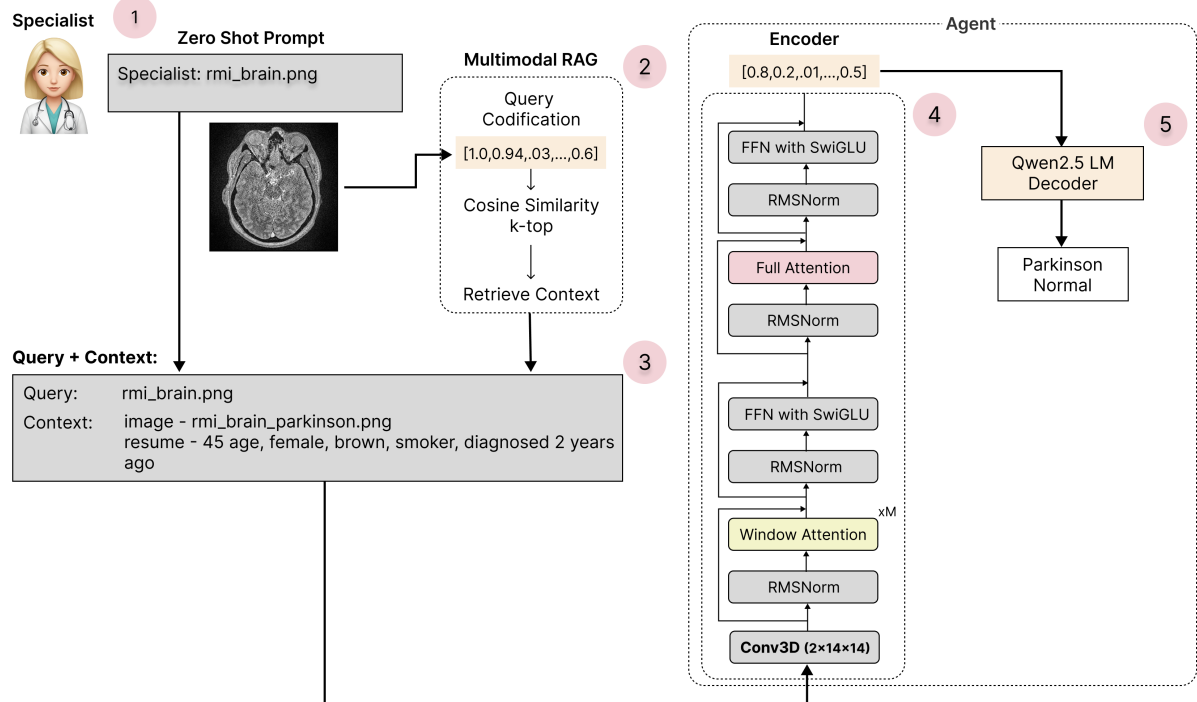
3. **Pluralidade:** A proposta que recebe o maior número de votos vence, mesmo que não atinja a maioria absoluta. É computacionalmente simples, mas pode levar a resultados paradoxais, onde a alternativa vencedora é, na verdade, a menos preferida pela maioria dos agentes quando comparada com outras opções.
4. **Votação por Pontuação:** Em vez de um único voto, os agentes atribuem uma pontuação a cada proposta (por exemplo, em uma escala de 1 a 5). A proposta com a maior soma de pontuações é a vencedora. Este método é mais expressivo, pois permite que os agentes indiquem a intensidade de suas preferências e não apenas sua primeira escolha.

Em termos operacionais, a camada de fusão adapta-se dinamicamente ao conjunto de modalidades disponíveis: se apenas um subconjunto, como áudio e dados tabulares, for fornecido, o sistema renormaliza os pesos ativos e prossegue com a inferência. Esse desenho garante robustez diante de dados ausentes e reflete cenários realistas de implantação, nos quais a disponibilidade de modalidades pode variar. O processo decisório combinado integra o raciocínio autônomo por meio do decodificador do modelo de linguagem de grande porte, *Qwen 2.5 LM*, com uma fusão de decisão estatisticamente fundamentada proveniente de classificadores específicos por modalidade, alinhando, assim, transparência e interpretabilidade aos fluxos de trabalho de tomada de decisão clínica.

Para ilustrar o fluxo de trabalho interno do *framework* agentivo proposto, a Figura 9 resume o pipeline de tomada de decisão clínica autônoma. Embora o exemplo represente o caminho visual usando dados de RM cerebral, o mesmo processo de raciocínio se estende de forma contínua a outras modalidades, incluindo sinais de áudio, sensores biométricos e texto clínico, quando estes estiverem disponíveis no prontuário do paciente.

Assim, a Figura 9 ilustra o pipeline ponta a ponta do *framework* proposto de tomada de decisão clínica autônoma. (1) Um especialista inicia um *prompt zero-shot*, definido como uma instrução única e não condicionada, que não fornece exemplos prévios nem modelos explícitos de tarefa ao modelo. Assim, o modelo de linguagem de grande porte deve derivar o raciocínio diagnóstico de forma autônoma a partir do próprio conteúdo do *prompt*, usando como único insumo contextual a RM cerebral do paciente (*rmi\_brain.png*) enviada à plataforma; (2) O *pipeline* RAG multimodal codifica a consulta, por meio de um modelo *image2vec* como o BLIP [Li *et al.* 2023], realiza uma busca por similaridade do cosseno para recuperar os *top-k embeddings* mais relevantes e reúne informações contextuais, como imagens semelhantes ou resumos de pacientes. (3) O sistema constrói um *prompt* interno combinando a consulta e o

Figura 9 – O processo autônomo de tomada de decisão clínica que integra uma arquitetura RAG multimodal com um codificador e o decodificador do modelo de linguagem Qwen 2.5.



Fonte: Próprio autor.

contexto recuperado. (4) Essa entrada multimodal é processada pelo codificador de visão, que extrai características tanto do texto quanto das imagens. (5) Por fim, o vetor de características é passado ao decodificador de modelo de linguagem *Qwen 2.5*, que executa raciocínio clínico autônomo e produz uma previsão diagnóstica (Normal ou Parkinson).

A seguir, é fornecido um pseudocódigo 4.4.1 que representa completamente como os nossos agentes são usados em nosso sistema.

#### 4.4.1 Ação

A ação da nossa arquitetura se dá através do uso de uma ferramenta criada para ser acionada de maneira autônoma baseada na escolha do agente. A ferramenta usada pela agente foi definida a partir do pseudocódigo 4.4.2 que exemplifica a interação entre um agente e a ferramenta customizada de uso de modelos especializados em classificação de dados. A seguir, a implementação conceitual da ferramenta de classificação de pacientes com DP e do agente que a utiliza.

---

```

1  CLASSE Agente:
2
3      // Atributos que definem o agente
4      PAPEL = "Médico Especialista"
5      OBJETIVO = "Analisar detalhadamente os dados do paciente e fornecer um
6      ↪ diagnóstico precoce de um paciente com DP ou normal."
7      FERRAMENTAS = [Modelo()]
8
9      // Método que descreve o fluxo de decisão do agente
10     FUNCAO pensar_e_agir(tarefa):
11         SE tarefa REQUER "classificar":
12             // O agente decide qual ferramenta usar
13             ferramenta_escolhida = SELECIONAR FERRAMENTA "Modelo"
14
15             // O agente usa a ferramenta, passando os dados da tarefa
16             resultado_da_ferramenta = ferramenta_escolhida.executar(tarefa.texto)
17
18             // O agente recebe o resultado e o usa para completar a tarefa
19             resposta_final = "O paciente foi diagnosticado com: " +
20             ↪ resultado_da_ferramenta
21
22     RETORNA resposta_final

```

---

Fonte: Próprio autor.

Listing 4.4.1 – Pseudo-código do Agente

---

```

1  CLASSE Modelo:
2
3      // Atributos da ferramenta, visíveis para o agente
4      NOME = "Modelo Classificador"
5      DESCRICAO = "Uma ferramenta que atua como wrapper de um modelo de machine
6      ↪ learning para classificação de pacientes com DP (1) ou Normal (0).".
7
8      // Método para inicializar o modelo de classificação
9      FUNCAO inicializar_modelo():
10         // Simula o carregamento de um modelo pré-treinado
11         RETORNA modelo_de_ia("classificador-zero-shot")
12
13         // O método principal que a ferramenta executa
14         // Ele recebe o dado como entrada
15         FUNCAO predict(dado):
16             // Define as classes possíveis
17             classes = ["DP", "Normal"]
18
19             resultado = self.modelo.predict(dado, classes)
20
21             // Retorna a classificação para o agente
22             RETORNA resultado

```

---

Fonte: Próprio autor.

Listing 4.4.2 – Pseudocódigo da Ferramenta de Classificação

#### 4.4.2 *Responsabilidades dos agentes*

Abaixo, nós propomos uma divisão clara e eficiente dos agentes usados na arquitetura proposta. A seguir, nossa arquitetura possui as seguintes definições de agentes:

**Validador** - O agente responsável por validar todo tipo de dado após a realização das etapas de pré-processamento e processamento, desde o dado não estruturado até o dado estruturado, independentemente da modalidade. Ao final, gera um relatório para controle e monitoramento de dados classificados como ouro - aqueles que passaram por todas as etapas e estão prontos para serem usados nos modelos.

**Recrutador** - O agente responsável por recrutar a equipe de especialistas após análise prévia de disponibilidade de dados para análise e gerenciamento da doença de DP.

**Especialista** - Esses agentes são especialistas baseados no contexto no qual estão inseridos. Esses agentes especialistas trabalham em um grupo com outros especialistas, discutem entre si, chegam a um consenso e, por fim, tomam a decisão médica referente à consulta do paciente, complementando com o fornecimento de relatórios técnicos sobre possíveis anomalias e patologias relacionadas à consulta do paciente.

**Moderador** - Esse agente modera discussões, resume informações e incentiva opiniões entre os agentes especialistas em um ambiente de equipe multidisciplinar. O moderador é o principal responsável em casos de discordância entre os especialistas, mediando discussões divididas em rodadas e turnos de fala. O moderador registra o relatório da reunião da MDT, destacando todas as informações importantes abordadas durante a reunião.

#### 4.5 *Métricas de avaliação*

A avaliação empírica dos componentes desenvolvidos nesta tese requer uma abordagem multifacetada, capaz de aferir o desempenho tanto de tarefas de classificação discriminativa quanto de sistemas de geração de linguagem natural. Para tal, esta seção detalha o conjunto de métricas empregadas, primeiramente para a avaliação do módulo classificador e, subsequentemente, para a análise rigorosa do sistema de RAG.

##### 4.5.1 *Avaliação do Classificador*

Para quantificar a eficácia dos modelos de classificação, adotamos um conjunto padrão de métricas derivadas da matriz de confusão. A matriz de confusão tabula as predições

do modelo em quatro categorias: Verdadeiros Positivos (VP), Falsos Positivos (FP), Verdadeiros Negativos (VN) e Falsos Negativos (FN). A partir dessas contagens, as seguintes métricas são calculadas:

- **Precisão (*Precision*):** Mede a proporção de instâncias classificadas como positivas que são, de fato, positivas. É uma métrica de exatidão das predições positivas, crucial em cenários onde falsos positivos têm um custo elevado.

$$\text{Precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}} \quad (4.1)$$

- VP = Verdadeiros Positivos
- FP = Falsos Positivos

- **Sensibilidade (*Recall* ou *Revocação*):** Mede a proporção de todas as instâncias positivas que foram corretamente identificadas pelo modelo. É uma métrica de completude, vital quando a não detecção de um caso positivo (Falso Negativo) é crítica.

$$\text{Sensibilidade} = \frac{\text{VP}}{\text{VP} + \text{FN}} \quad (4.2)$$

- FN = Falsos Negativos

- ***F1-Score*:** Corresponde à média harmônica entre Precisão e Sensibilidade. Essa métrica oferece um balanço entre as duas, sendo particularmente útil quando o custo de Falsos Positivos e Falsos Negativos é comparável ou quando há um desbalanceamento de classes.

$$F_1 = 2 \times \frac{\text{Precisão} \times \text{Sensibilidade}}{\text{Precisão} + \text{Sensibilidade}} \quad (4.3)$$

- ***Area Under The Receiver Operating Characteristic curve (AUC-ROC)*:** A curva ROC plota a Taxa de Verdadeiros Positivos (TPR, equivalente à Sensibilidade) dada por  $\text{TPR} = \text{TP}/(\text{TP} + \text{FN})$  em função da Taxa de Falsos Positivos (FPR), calculada como  $\text{FPR} = \text{FP}/(\text{FP} + \text{TN})$ , para diversos limiares de decisão. A área sob essa curva *AUC* quantifica a capacidade geral do modelo de discriminar entre as classes positiva e negativa. Um valor de *AUC* de 1,0 representa um classificador perfeito, enquanto 0,5 indica um desempenho não superior ao aleatório.

#### 4.5.2 Avaliação do Retrieval-Augmented Generation (RAG)

A avaliação de um sistema RAG é inerentemente modular, refletindo seu pipeline de operação. Portanto, nossa metodologia decompõe a análise em três eixos principais: a eficácia

da etapa de Recuperação de contexto [Es *et al.* 2023], a qualidade da etapa de Geração de texto [Es *et al.* 2023] e, por fim, uma avaliação holística do sistema por meio da abordagem *LLM-as-Judge* [Zheng *et al.* 2023].

#### 4.5.2.1 Avaliação da Fase de Recuperação no Retrieval-Augmented Generation (RAG)

A robustez de um sistema RAG depende fundamentalmente da sua capacidade de recuperar documentos relevantes e precisos da base de conhecimento. Para medir a eficácia desta etapa, utilizamos métricas baseadas em similaridade semântica, conforme popularizado por *frameworks* como RAGAs [Es *et al.* 2023].

**Similaridade Semântica Textual, ou do inglês, *Semantic Textual Similarity* - STS:** Esta abordagem quantifica a relevância entre a consulta do usuário, ou um texto de referência, e os documentos recuperados. O processo envolve a geração de representações vetoriais, *embeddings*, para os textos por meio de um modelo de linguagem pré-treinado. Subsequentemente, a similaridade entre os vetores é calculada, sendo a similaridade de cosseno a métrica mais comum para este fim [Rahutomo *et al.* 2012]. A similaridade de cosseno entre dois vetores  $v_1$  e  $v_2$ , que nesse caso são vetores de *embeddings*, é definida como:

$$\text{Similaridade de cosseno} = \frac{\vec{v}_1 \cdot \vec{v}_2}{\|\vec{v}_1\| \cdot \|\vec{v}_2\|} \quad (4.4)$$

Essa métrica é o alicerce para avaliar tanto a relevância do contexto (se os trechos recuperados são pertinentes à pergunta) quanto a revocação do contexto (se todos os trechos necessários para responder à pergunta foram recuperados).

#### 4.5.2.2 Métricas de Geração

Uma vez que o contexto é recuperado do banco de dados vetorial, a qualidade da resposta gerada pelo modelo de linguagem é avaliada. Esta segunda etapa de avaliação do RAG foca em aspectos como fidelidade factual, coerência e semelhança com uma resposta de referência, padrão-ouro.

**BLEU (*Bilingual Evaluation Understudy*):** Originária da tradução automática, a métrica BLEU avalia a qualidade de um texto gerado medindo a sobreposição de n-gramas (sequências de n palavras) com um ou mais textos de referência. Ela foca na **precisão** dos n-gramas e aplica uma "penalidade de brevidade", ou em inglês "*Brevity Penalty*", (*BP*) para evitar que respostas excessivamente curtas, mas com alta sobreposição, recebam pontuações inflacionadas [Papineni

et al. 2002]. A fórmula é:

$$\text{BLEU} = BP \cdot \exp \left( \sum_{n=1}^N w_n \log(p_n) \right) \quad (4.5)$$

onde  $p_n$  é a precisão modificada dos n-gramas e  $w_n$  são pesos (tipicamente uniformes).

**ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** Em contraste com o BLEU, ROUGE é um conjunto de métricas focado na revocação (*recall*). Ele mede quantos n-gramas do texto de referência estão presentes no texto gerado, tornando-o ideal para avaliar tarefas como sumarização e respostas a perguntas, onde a inclusão de informações essenciais é primordial. As métricas de ROUGE utilizadas neste trabalho são [Lin 2004]:

1. **ROUGE-N:** Mede a sobreposição de n-gramas. A fórmula para ROUGE-N é:

$$\text{ROUGE-N} = \frac{\sum_{\text{sentece} \in \text{reference}} \sum_{\text{n-gram} \in \text{sentece}} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{\text{sentece} \in \text{reference}} \sum_{\text{n-gram} \in \text{sentece}} \text{Count}(\text{n-gram})} \quad (4.6)$$

Onde  $\text{Count}_{\text{match}}$  é o número de n-gramas na referência que também aparecem na geração, e  $\text{Count}$  é o número total de n-gramas na referência.

2. **ROUGE-L:** difere do ROUGE-N, pois se baseia na mais longa subsequência comum, ou do inglês, *Longest Common Subsequence* (LCS) entre o resultado do modelo e a referência, ou seja, a mais longa sequência de palavras compartilhada entre ambos. Uma sequência compartilhada mais longa deve indicar maior semelhança entre as duas sequências [Lin 2004]. A notação formal dessa métrica é:

$$F_{lcs} = \frac{(1 + b^2)R_{lcs}P_{lcs}}{R_{lcs} + b^2P_{lcs}} \quad (4.7)$$

Nesse sentido,  $R_{lcs}$  é o *Recall* baseado na LCS,  $P_{lcs}$  é a *Precisão* baseada na LCS e, por fim,  $b$  é o parâmetro que pondera ambos.

**BERTScore:** Superando as limitações das métricas baseadas em sobreposição lexical, o BERTScore avalia a similaridade semântica entre a resposta gerada e a de referência [Zhang et al. 2019]. Ele utiliza *embeddings* contextuais de modelos como o BERT [Devlin 2018] para calcular a similaridade de cosseno entre os *tokens* de ambos os textos. Isso permite que a métrica reconheça sinônimos e paráfrases, proporcionando uma avaliação mais rica em termos semânticos. Reportamos os valores de precisão, *recall* e F1 do *BERTScore*.

**Perplexidade (Perplexity):** É uma métrica intrínseca que avalia a fluência e a confiança de um modelo de linguagem. Matematicamente, é o expoente da entropia cruzada e mede quão "perplexo" ou "surpreso" o modelo está ao prever a sequência de tokens de um texto de referência.

Figura 10 – Exemplo de uma classificação usando LLM-as-Judge

**Q:** In what ways does covid 19 increase the risk of domestic violence?

**A:** Why should I care about violence against women during COVID-19? Violence against women is a grave violation of women's human rights, no matter when, where, or how it takes place. There are many forms of violence against women, which have many potential negative health consequences for women and their children. Violence against women can result in injuries and serious physical, mental, sexual and reproductive health problems, including sexually transmitted infections, HIV, and unplanned pregnancies. In extreme cases, violence against women can result in death. The effects of violence are very often long-lived. Violence - in all its forms - can have an impact on a woman's well-being throughout the rest of her life. This is unacceptable, and the risks of violence that women and their children face during the current COVID-19 crisis cannot be ignored. When we are able to prevent violence, or to support women survivors of violence, we help to safeguard women's human rights, and promote physical and mental health and well-being for women throughout their lives. This also helps to alleviate pressure on already stretched essential public services, including the health system.

<p><b>Reviewer 1:</b></p> <p><b>Rating:</b> Bad</p> <p><b>Explanation:</b> Condemns violence against women but doesn't explain why the risk increases during the pandemic.</p>	<p><b>Reviewer 2:</b></p> <p><b>Rating:</b> Bad</p> <p><b>Explanation:</b> This answer does not provide any information on ways Covid-19 can increase the risk of domestic violence on women. Instead, it provides explanation and information on the impact and results of violence against women on their mental and physical health.</p>
--	---

**Mixtral-8x7B-Instruct:**

**Feedback:**

**Total rating:**  
Bad

**Evaluation:**  
The system's response is terrible: completely irrelevant to the question. The question asks about the ways in which COVID-19 increases the risk of domestic violence, but the system's response does not address this at all. Instead, it focuses on the general issue of violence against women, which is not specific to the current pandemic. The response is also excessively long and repetitive, making it difficult to extract any useful information."

Fonte: Próprio autor.

Uma perplexidade menor indica que a distribuição de probabilidade do modelo está bem alinhada com a do texto de referência, sugerindo maior qualidade e coerência na geração [Huang *et al.* 2025].

$$\text{Perplexidade}(P) = \exp \left( -\frac{1}{N} \sum_{i=1}^N \log_2 P(w_i | w_1, \dots, w_{i-1}) \right) \quad (4.8)$$

#### 4.5.2.3 LLM-as-a-Judge

Uma técnica não convencional mais comumente usada para avaliação, oferecendo dois benefícios principais: escalabilidade e explicabilidade [Zheng *et al.* 2023]. No contexto do *Big Data*, grandes volumes de dados são armazenados rotineiramente em lagos de dados, o que torna a anotação manual inviável nesses cenários. Portanto, surge a necessidade de treinar LLMs para atuarem como juízes e avaliarem automaticamente a exatidão e a precisão das respostas de outros LLMs. Além disso, o *LLM-as-a-Judge* fornece não apenas pontuações, mas também explicações, tornando seus resultados interpretáveis [Zheng *et al.* 2023], conforme a figura 9:

Como usado em outras etapas deste trabalho, empregamos essa técnica para analisar a resposta do nosso sistema baseado em RAG a fim de gerar análises sobre ele. Foi proposto o uso de um *LLM-as-a-Judge* para classificação de resposta única com um prompt motivacional de poucos disparos, o que significa que o juiz do LLM é solicitado a atribuir diretamente uma pontuação a uma única resposta. Os avisos de poucos disparos usam exemplos para ajudar a

melhorar a precisão da resposta de classificação, geralmente entre 20 e 30 exemplos para basear a análise e, em seguida, comparam os resultados fornecidos pelo aviso com os gerados pelo sistema para analisar se estão alinhados. Após a análise, ele retorna o *feedback*, pontuando de 1 a 4 a precisão da resposta quando comparada ao padrão-ouro. Além disso, ele calcula a correlação de *Pearson* entre a resposta gerada pelo LLM e o padrão-ouro. Os níveis de classificação gerados pelo *LLM Judge* são:

1. A resposta do sistema é péssima: completamente irrelevante para a pergunta feita ou altamente tendenciosa;
2. A resposta do sistema é, em sua maior parte, inútil: não aborda alguns aspectos importantes da pergunta;
3. A resposta do sistema é, em sua maior parte, útil: fornece suporte, mas ainda pode ser melhorada;
4. A resposta do sistema é excelente: relevante, direta, detalhada e aborda todas as questões levantadas na consulta do paciente.

#### 4.5.3 Avaliação do Sistema

No contexto de *Big Data*, a novidade e as incertezas inerentes a muitos recursos baseados em modelos de linguagem são inevitáveis. Portanto, é fundamental manter os padrões de privacidade de dados e a responsabilidade social. Atualmente, grandes plataformas e modelos de dados estão tentando aplicar LLMOps, um nicho de MLOps, sendo fortemente condicionados a manter essas plataformas em execução contínua e a monitorar seus ciclos de vida [Wang *et al.* 2024].

Assim, é essencial analisar e medir alguns aspectos do uso desses grandes sistemas focados em modelos de linguagem natural, tais como: monitorar a estrutura dos dados fornecidos para o treinamento desses modelos, a qualidade dos LLMs, SLMs, RAGs, agentes autônomos e todo o ciclo de vida dos modelos. Essas análises permitem que os desenvolvedores e as empresas obtenham *insights* valiosos sobre o uso dessas tecnologias no mundo real, garantindo a confiabilidade e a eficiência dos LLMs por meio de avaliações controladas e automatizadas.

Este trabalho, portanto, concentrou-se em avaliações on-line, que são realizadas em cenários de ambiente de produção expostos ao mundo real. Esse processo envolve avaliadores automáticos acionados por novas entradas de registro derivadas da produção ao vivo. A avaliação on-line é excelente para refletir as complexidades do uso no mundo real e integra um valioso

*feedback* do usuário, tornando-a ideal para o monitoramento contínuo do desempenho. Então, foram analisadas métricas de custo, as quais abordam fatores envolvendo o custo de chamada de LLMs/SLMs, uso da Unidade de Processamento Gráfico (GPU) e o custo de infraestrutura.

### **Métricas de custo**

- **Utilização de GPU** Essa métrica tem como objetivo medir o uso da GPU configurado no backend da plataforma com base no número total de tokens e no número de respostas recebidas.
- **LLM Calls Cost** Decidimos medir o preço de custo de uma API comercial, como, *Anthropic*, *OpenAI*) ou ao optar por comprar uma máquina com excelentes configurações para executar o servidor completo.
- **Custo da infraestrutura** Essa é uma métrica não convencional, mas muito importante para o contexto de implantação de serviços do LLM. O custo da infraestrutura do seu aplicativo torna-se um fator crucial em contextos de produção. O sistema abordado aqui não foi testado em produção com vários usuários usando a plataforma simultaneamente e continuamente, como deveria ser. Portanto, essa métrica tem como objetivo demonstrar um custo básico para manter uma estrutura de alto nível baseada em LLMOps em uma versão inicial de produção, 2/3 usuários simultâneos na plataforma, por um mês. Serão avaliados os custos de armazenamento de vetores, como o *pinecone*. Além disso, será considerada a infraestrutura de back-end necessária para garantir o desempenho adequado das aplicações. Essa análise incluirá tanto instâncias em nuvem, como a *AWS EC2*, modelo *t3.medium*, quanto a possibilidade de utilização de máquinas locais, armazenamento (registros, dados de pacientes, resultados, modelos LLM) e rede.

## **4.6 Descrição do dataset**

O presente estudo utiliza um conjunto de dados simulado, (*mock dataset*), gerado com base em diversas fontes, a fim de representar um cenário plausível para testes e validação do sistema. composto por várias fontes diferentes, incluindo a aquisição de dados em tempo real de vestíveis equipados com sensores, imagens de TC e Ressonância Magnética (RM), áudios e documentos textuais. O conjunto de dados adotado neste estudo está estruturado em informações relacionadas ao paciente e registros provenientes da plataforma utilizada.

### ***Dados da Plataforma***

Os dados da plataforma serão os dados que serão usados para treinar os grandes ou pequenos modelos de linguagem (LLMs/SLMs) semanalmente para que, posteriormente, sejam incorporados aos agentes do sistema. Estes dados são codificados usando criptografia para preservação do anonimato dos usuários. Além disso, esses dados também foram retirados de fontes de referência relacionadas a DP, como *Parkinson's Progression Markers Initiative* (PPMI), uma fonte gratuita de dados extremamente importante para a previsão, diagnóstico e tratamento de DP. O PPMI é um estudo observacional em larga escala que acompanha a evolução clínica da DP ao longo do tempo, iniciado no ano de 2010, pela Michael J. Fox Foundation, [THE MICHAEL J. FOX FOUNDATION FOR PARKINSON'S RESEARCH 2025], que tem como objetivo principal a identificação e validação de biomarcadores clínicos, genéticos, de imagem e de biosinais associados à progressão da DP acelerando avanços científicos e novos tratamentos. Devido à massiva variedade de tipos de dados relacionados a DP, optou-se por analisar minuciosamente as seguintes modalidades do portal de dados do PPMI:

- Histórico Familiar
- Prontuários médicos
- Biomarcadores
- Imagens
  - Tomografia Computadorizada
  - Ressonância Magnética

Ademais, uma outra importante fonte de dados são os adquiridos através de dispositivos, como vestíveis (caneta inteligente [Júnior *et al.* 2020]) *smartphones*, que se comunicam com o nosso servidor local, integrando nosso sistema IoT. A proposta da nossa infraestrutura é possuir baixa latência e envio seguro de dados adquiridos a partir de sensores/dispositivos para computação de borda, como usado em nosso trabalho, *Raspberry Pi 4 modelo B*, que realiza um pré-processamento simples e comum para cada modalidade do dado e realiza o envio para a plataforma, envolvendo o processo através da nuvem. Além dos vestíveis e *smartphones*, temos outros dispositivos que atuam na aquisição de dados para a plataforma, outros diretamente como os relatórios clínicos e, outros indiretamente como as imagens de TC e RM mencionados anteriormente, que requer que a unidade de saúde ou hospital tenha um sistema

É importante pontuar que diferentes dispositivos de aquisição possuem diferentes tempos de captação de dados e, dependendo da modalidade do dado, a inserção desses dados no

Tabela 7 – Tempo de aquisição de dados por tarefas

Tarefa	Dispositivo/Vestível Sistema	Tempo Médio (s)
Gravação de voz	Smartphone	10
Testes HandDP	Caneta Inteligente	1
Imagens de RM	Ressonância Magnética	300
Imagens de TC	Tomógrafo Computadorizado	60

Fonte: Próprio autor.

nosso servidor do *Data Lake* também, devido a espera do pré-processamento a partir do nosso dispositivo de borda. A seguir, tempo médio de envio de dados para a plataforma:

Contudo, ainda assim, a própria construção de infraestrutura proposta nesse trabalho, como mostrado anteriormente na figura 6, garante uma aquisição de dados de maneira online ou quase em tempo real.

Portanto, decorrentes da infraestrutura IoMT proposta, os seguintes dispositivos responsáveis pela aquisição de dados são:

- *Caneta Inteligente*: um dispositivo vestível proposto pelo autor Eugenio *et al.* em seu trabalho [Júnior *et al.* 2020]. O vestível conta com um sensor acelerômetro embutido no corpo da caneta que atua na captação de dados nos eixos X, Y e Z que reflete os movimentos motores dos membros superiores dos usuários do dispositivo.
- *Smartphone*: dispositivo móvel pessoal que é capaz de gravar áudios e vídeos.

Esses dados, conforme mencionado anteriormente na figura 6, seguem todo o fluxo de *DataOps* que, ao fim, são validados através de uma dupla de agentes validadores, na qual eles verificam baseados em um pré-processamento e processamento definidos previamente que varia de modalidade para modalidade do dado inserido no *Data Lake*. Em cada estágio do dado ele é armazenado, portanto, o dado bruto é armazenado em *raw-data* e então pre-processado pelo dispositivo de borda, novamente armazenado e, por fim, processado e validado automática pela dupla de agentes que, também é armazenado como dado ouro.

### ***Dados de Usuário***

Todos os usuários da plataforma possuem um catálogo de dados exclusivamente pessoal. Como é extremamente importante evitar o vazamento de dados sensíveis/pessoais dos usuários, esses dados são armazenados no *Data Lake* assim como os dados da plataforma, no entanto, esses dados são utilizados pra alimentar o sistema RAG usado nesse trabalho para aprimorar a resposta a partir de dados que fogem do seu escopo de aprendizado inicial. Dessa

forma, o sistema garante assertividade na resposta baseada em um histórico clínico do paciente. Esses dados, posteriormente, são armazenados em um banco vetorial com um identificador que varia de usuário para usuário, ou seja, a organização do banco de dados vetorial é exemplificada na Tabela 8:

Tabela 8 – Schema dos dados no banco vetorial

<b>Campo</b>	<b>Type</b>
uid	INT64
embedding	FLOAT_VECTOR(dim)
modalidade	VARCHAR(32)
classe	VARCHAR(32)
sumario	VARCHAR(150)
sumario_emb	FLOAT_VECTOR(dim)
raw_ref	VARCHAR(100)

Tabela 9 – Descrição dos campos do schema

<b>Descrição dos campos</b>
ID única do item
Vetor de embedding
Modalidade do dado
Classe (DP, Controle)
Resumo breve sobre o dado
Vetor de embeddings do resumo
Caminho do dado original

Fonte: Próprio autor.

Serão os dados que serão usados para treinar os grandes ou pequenos modelos de linguagem (LLMs/SLMs) semanalmente para que, posteriormente, sejam incorporados aos agentes do sistema. Estes dados são codificados usando criptografia para preservação do anonimato dos usuários. Além disso, esses dados também foram retirados de fontes de referência relacionadas a DP, como PPMI, uma fonte gratuita de dados extremamente importante para a previsão, diagnóstico e tratamento de DP. O PPMI é um estudo de referência que conta com a colaboração de parceiros do mundo todo através de informações de acesso aberto para acelerar os avanços científicos e novos tratamentos. Devido a massiva variedade de tipos de dados relacionados a DP, optou-se por analisar minuciosamente as seguintes modalidades do portal de dados do PPMI:

1. Imagens:

- Tomografia Computadorizada;
- Ressonância Magnética.

2. Histórico Médico:

- Clínico: Histórico de diagnóstico de PD, diagnóstico clínico, diagnóstico clínico primário, exame físico geral, ECG;
- Exame neurológico.

3. Características do sujeito:

- Histórico familiar;

- Status do paciente.

#### 4. Avaliações não motoras:

- Testes Autonômicos;
- Testes Neurocomportamentais: Escala de depressão geriátrica (versão curta), inventário de ansiedade traço-estado;
- Testes de distúrbios do sono: Escala de sonolência de Epworth, questionário de transtorno do comportamento do sono REM.

#### 5. Achados:

- RFQ Álcool
- RFQ Cafeína
- RFQ Altura e peso
- RFQ Atividade física
- RFQ Histórico de tabagismo

Adicionalmente a isso, temos um outro conjunto de dados que será usado para popularizar inicialmente os dados de usuários na plataforma. Os dados foram obtidos do site da Biblioteca Nacional de Medicina da Universidade da Índia. Esse conjunto de dados é dividido em duas modalidades, imagem e tabular. Conforme, anunciado anteriormente, o dado tabular se divide em dois, no qual, um representa informações sobre os metadados das imagens e, o segundo, representa informações de relatórios médicos de pacientes e suas respectivas doenças, ou seja, dados valiosos complementares para auxiliar no diagnóstico médico, como: imagens associadas ao paciente, indicações, comparações, achados e impressões.

A incorporação dos dois conjuntos ao catálogo de dados de usuários foi feita de maneira não pragmática, assim como na realidade, ou seja, sem um padrão definido, portanto, pode calhar de ter mais informações pessoais adicionais sobre um paciente específico do que outros, isso mostrará também que para pacientes com muitos dados na plataforma, a assertividade tende a ser melhor.

Este conjunto de dados é previamente submetido a um pré-processamento básico para otimizar o desempenho durante os processos realizados nele. Para as imagens, foram convertidas de DICOM para PNG, ajustando os valores de pixel para o intervalo de 0 a 255 e redimensionadas para um limite de 2048 pixels no lado mais curto.

Devido ao dinamismo dos dados e a falta de padronização decorrente dos diferentes propósitos dos dados, optou-se por analisar minuciosamente, de maneira individual, o que

representa cada característica de ambos os dados usados e, como isso, pode ajudar quando combinado ao sistema para geração de respostas.

#### **4.6.1 Pré-processamento**

A etapa de pré-processamento é crucial para a qualidade e robustez dos modelos subsequentes. Implementamos uma pipeline automatizada para cada modalidade de dados, garantindo a homogeneidade e a integridade das informações adquiridas. Esta rotina executa continuamente, processando os dados à medida que são coletados e armazenados em nossa infraestrutura de nuvem. Independentemente da modalidade, o fluxo padrão de pré-processamento segue a sequência:

Limpeza → Normalização → Codificação/Representação

A seguir, detalhamos as especificidades das etapas de pré-processamento para cada modalidade de dados em nossa plataforma DataOps.

##### *Sensorial*

- Limpeza: remoção de segmentos iniciais inativos e de tendências de corrente contínua (DC offset).
- Normalização: aplicação de filtro passa-banda Butterworth (0,5–20 Hz) e padronização Z-score para cada canal.
- Codificação: variáveis categóricas são transformadas via One-Hot Encoding. Os dados pré-processados são persistidos em formato CSV.

##### *Textual*

- Limpeza: eliminação de stopwords, emojis e caracteres especiais.
- Normalização: conversão de todo o texto para caixa baixa e remoção de espaços em excesso.
- Representação: tokenização do texto para posterior vetorização. A representação numérica é realizada via TF-IDF ou embeddings contextuais (e.g., BERT), dependendo da tarefa downstream.

## *Imagens*

- Limpeza: realiza uma verificação da imagem para checar se está corrompida.
- Normalização: decodificação da imagem de binário para um formato adequado, redimensionamento da imagem para  $224 \times 224$  pixels, aplicação do CLAHE para padronização da intensidade luminosa e filtro gaussiano para suavizar os pixels da imagem.
- Codificação: conversão em tensores e codificação para armazenamento em nuvem.

## *Áudio*

- Limpeza: carregamento do áudio e amostragem do sinal, cálculo do espectrograma, aplicando "noise gate" e "clipping" ao áudio, remoção de silêncio e aplicando filtro dual bandpass ao áudio.
- Normalização: reamostragem do sinal para 8 kHz, verificando duração mínima do áudio, normalizando loudness e aplicando o modelo VAD para identificação de fala para remoção de silêncio avançada.
- Codificação: codificação para formato original para armazenamento na nuvem.

### **4.6.2 *Processamento***

Após o pré-processamento, os dados passam por etapas de engenharia de características e/ou aumento de dados para otimizar o desempenho dos modelos. O fluxo padrão, quando aplicável, consiste em:

Engenharia de Características  $\longrightarrow$  Aumento de Dados

Em cada modalidade:

- Sensorial:
  1. Segmentação de Tarefas (teste HandPD): Os dados são segmentados de acordo com as tarefas do teste HandPD, incluindo: desenho de círculos (no formulário e no ar), desenho de espirais, desenho de meandros no formulário e diadococinesia para as mãos direita e esquerda.
  2. Sincronização Temporal: Criação de um timestamp unificado combinando hora e milissegundos para alinhamento preciso dos dados.

3. Engenharia de Características: Geração de novas características a partir de cada canal do acelerômetro (AcX, AcY, AcZ) e do giroscópio (GyX, GyY, GyZ). As características específicas dependem da análise a ser realizada (e.g., amplitude, frequência, variância, etc. — aqui seria bom ser mais específico se possível).
  4. Aumento de Dados: Aplicação da técnica SMOTE (Synthetic Minority Over-sampling Technique) na classe minoritária (grupo saudável) para balanceamento do conjunto de dados.
- Vetorização e Embeddings: criação de representações vetoriais densas e embeddings contextuais que capturam relações semânticas e sintáticas do texto, utilizando modelos pré-treinados ou treinados no próprio corpus.
- Imagens:
1. Aumento de Dados: Aplicação de transformações nas imagens originais para expandir o conjunto de treinamento e aumentar a robustez do modelo. As transformações empregadas incluem rotação, translação, espelhamento (flip) e recortes aleatórios (random crop). (É importante ser mais específico sobre quais flips e se random crop envolve redimensionamento.)
- Áudio:
1. Características de Fonação: extração de métricas fônicas tradicionais como Jitter, Shimmer, Razão Harmônico-Ruído (HNR) e Razão Não-Harmônico-Ruído (NHR).
  2. Índices Não Lineares: cálculo de métricas de complexidade e não linearidade, como o Parâmetro de Potência Relativa Discreta (RPDE), Análise de Flutuação Detrended (DFA) e Entropia de Permutação Ponderada (PPE).

## 5 RESULTADOS E DISCUSSÃO

### 5.1 Análise Estatística

A análise estatística desempenha um papel essencial na compreensão da estrutura da base de dados, especialmente por se tratar de um conjunto heterogêneo, composto por múltiplas fontes. Inicialmente, foi realizada uma análise estatística exploratória com o objetivo de identificar lacunas e características relevantes, como a presença de valores ausentes em dados tabulares e a necessidade de aplicar técnicas de aumento de dados, como a inserção de ruídos.

No caso dos dados de imagem, a análise focou-se na avaliação visual do desbalanceamento entre as classes controle e PD, além da aplicação de procedimentos de limpeza e normalização dos *pixels*. Para os dados de áudio, foram analisadas as distribuições por grupo, sexo e faixa etária com base nas informações demográficas, e propôs-se uma análise dos espectrogramas das vozes dos pacientes, tanto estáveis quanto com diagnóstico de DP.

#### 5.1.1 *audio*

Tabela 10 – Distribuição dos sujeitos por grupo

Label	Count	Percent (%)
HC	41	50.6
PwPD	40	49.4

Fonte: Próprio autor.

Tabela 11 – Distribuição de sexo por grupo e resultado do teste qui-quadrado

Label	Count		Percentagem (%)	
	F	M	F	M
HC	25	16	61.0	39.0
PwPD	19	21	47.5	52.5

Fonte: Próprio autor.

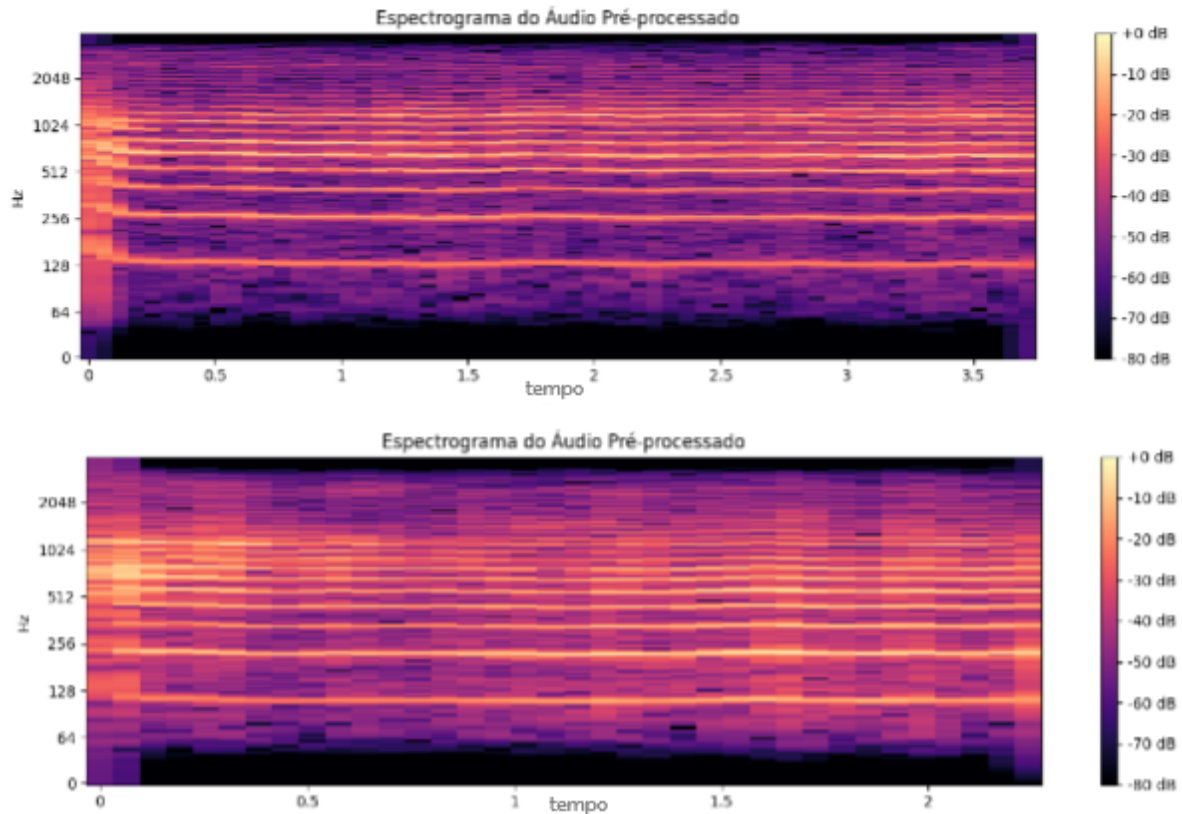
Teste qui-quadrado (sexo vs. grupo):  $\chi^2 = 0.988$ ,  $p = 0.320$  (não significativo)

Tabela 12 – Estatísticas descritivas de idade por grupo

Label	N	Média	Mediana	DP	Mínimo	Máximo	Assimetria
HC	41	47.73	45.0	14.33	18.00	79.00	0.29
PwPD	40	67.04	67.9	8.98	43.12	85.31	-0.56

Fonte: Próprio autor.

Figura 11 – Espectrogramas do Áudio Pré-processado. A primeira imagem se refere ao paciente do grupo de controle e a segunda imagem é referente ao paciente com DP.



Fonte: Próprio autor.

Os espectrogramas são representações visuais da variação do espectro de frequência de um sinal ao longo do tempo, sendo ferramentas poderosas para a análise de sinais de áudio. Eles permitem a observação de características como formantes, harmônicos e ruídos, que podem ser indicadores de patologias vocais. A Figura 11 apresentam exemplos de espectrogramas de áudios pré-processados de um participante do grupo de controle e de um participante com DP, respectivamente. Ambos os espectrogramas são plotados em uma escala logarítmica para a frequência (eixo Y, em Hz) e linear para o tempo (eixo X, em segundos), com a intensidade do sinal (amplitude em dB) representada pela escala de cores.

O espectrograma do paciente AH\_114S pertencente ao Grupo de Controle, segunda imagem da Figura 11, exibe padrões de energia espectral mais uniformes e bem definidos ao longo do tempo. As barras horizontais claras, especialmente nas frequências mais baixas (abaixo de 1024 Hz), indicam a presença de formantes vocais estáveis e harmônicos bem distintos, característicos de uma produção vocal saudável. A transição de intensidade ao longo da duração do áudio parece ser mais consistente, com menor variabilidade abrupta na distribuição da energia.

Isso sugere uma emissão vocal com menor flutuação de frequência e amplitude, refletindo uma coordenação neuromuscular típica.

Diferentemente, o espectrograma do paciente AH\_064F com DP, primeira imagem da Figura 11, pode apresentar características que refletem as alterações vocais frequentemente associadas à DP. Embora a estrutura geral de formantes ainda possa ser observada, é comum que pacientes com DP exibam sinais de instabilidade vocal, como:

- Irregularidades nos harmônicos: As linhas horizontais podem ser menos nítidas ou mais difusas, indicando uma perda da periodicidade vocal devido a tremores ou rigidez na laringe.
- Redução da amplitude ou "apagões" vocais: Áreas de menor intensidade (cores mais escuras na escala de dB) podem ser mais proeminentes ou ocorrerem de forma mais errática, refletindo a disfonia hipocinética, que se manifesta como volume reduzido (hipofonia) ou interrupções na voz.
- Maior variabilidade nas frequências: A consistência dos formantes pode ser comprometida, com flutuações mais visíveis, que podem ser indicativos de instabilidade de *pitch* ou *jitter/shimmer* elevados.

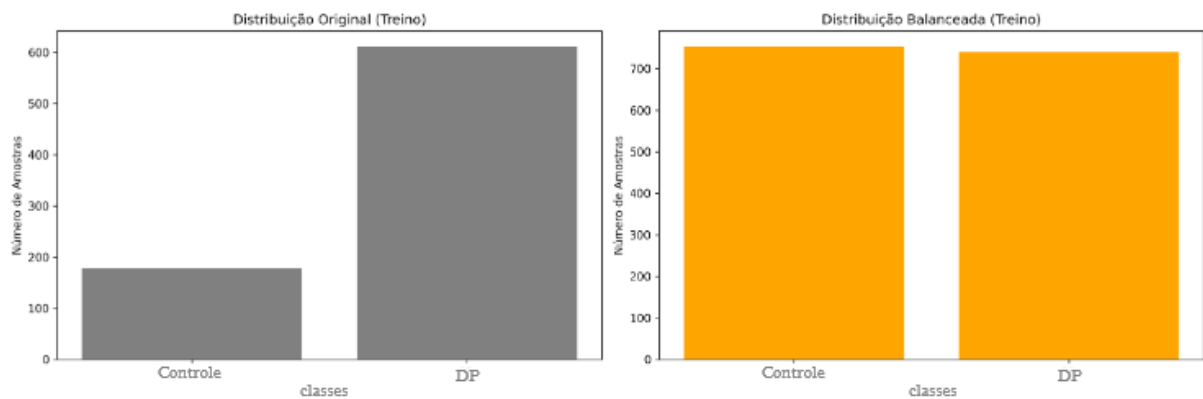
A análise visual dos espectrogramas é uma ferramenta valiosa para identificar essas diferenças qualitativas entre as vozes de indivíduos saudáveis e aqueles com DP, fornecendo *insights* importantes para o desenvolvimento de características, *features*, que podem ser utilizadas em modelos de classificação. A representação espectral serve como a base visual primária para a extração de *features* acústicas que, subsequentemente, alimentam os algoritmos de ML para a detecção da DP.

### 5.1.2 *imagem*

A distribuição das classes no conjunto de dados de treino é um aspecto crítico que pode impactar diretamente o desempenho de modelos de aprendizado de máquina, especialmente em cenários de classificação. A Figura 12 ilustra a distribuição comparativa das classes Controle e DP em duas configurações: a distribuição original do conjunto de treino e a distribuição após o balanceamento.

O gráfico à esquerda na Figura 12 revela uma assimetria significativa na distribuição das classes no conjunto de dados de treino original. A classe Controle possui aproximadamente 180 amostras, enquanto a classe DP apresenta um número substancialmente maior, ultrapassando

Figura 12 – Distribuição de Classes Antes e Depois do Balanceamento



Fonte: Próprio autor.

600 amostras. Essa discrepância caracteriza um desbalanceamento de classes, onde a classe DP é a majoritária e a classe Controle é a minoritária. Em modelos de classificação, o desbalanceamento pode levar a um viés, onde o modelo tende a favorecer a classe majoritária, resultando em menor sensibilidade para a classe minoritária e, conseqüentemente, em métricas de desempenho enganosas.

Em contrapartida, o gráfico à direita na Figura 12 demonstra a distribuição das classes após a aplicação de uma técnica de balanceamento. Observa-se que ambas as classes, Controle e DP, possuem um número equitativo de amostras, aproximando-se de 730 amostras por classe. Este balanceamento é crucial para mitigar o viés do modelo em relação à classe majoritária, permitindo que ele aprenda características representativas de ambas as classes de forma mais justa. A estratégia de balanceamento empregada visou criar um conjunto de treino onde o modelo pudesse otimizar seu aprendizado sem ser excessivamente influenciado pela dominância de uma classe, promovendo assim uma melhor generalização e desempenho na identificação de padrões de ambas as condições.

## 5.2 Análise do desempenho dos Agentes

### 5.2.1 *individualmente*

A priori, analisamos estatisticamente toda a conjuntura da base de dados referente a cada modalidade, seja áudio, texto, imagem, tabular, para que dessa forma possamos garantir que os dados serão devidamente utilizados com proficiência e qualidade. Como a análise e os cálculos estatísticos variam muito de acordo com a modalidade do dado, decidimos separar essa seção de análises para cada modalidade.

Os resultados da análise estatística obtida neste estudo foram gerados previamente, de maneira modular, ou seja, externo a plataforma. Após análises surgiram três grandes categorias distintas de agentes:

1. Agentes de Aprendizado de Máquina Tradicional: aplicam algoritmos clássicos (por exemplo, *Support Vector Machine (SVM)*, *Random Forest* e *K-Nearest Neighbor (KNN)*) para classificar pacientes em saudável ou com DP com base em dados sensoriais e de áudio (os arquivos .wav são convertidos em tabelas .csv após o pré-processamento).
2. Agentes de Aprendizado Profundo: utilizam redes neurais convolucionais ajustadas para processamento de imagens médicas (RM e TC), explorando arquiteturas como *ResNet* e *VGG* adaptadas ao nosso conjunto de dados.
3. Agentes Baseados em Modelos de Linguagem: empregam modelos pré-treinados, como *LLama 3.2*, *Deepseek R1*, para análise de dados textuais, histórico clínico, anotações médicas e questionários e extração de características semânticas.
4. Agentes Baseados em Modelos de Visão de Linguagem: são semelhantes aos modelos de linguagem tradicionais, mas se diferem pois são capazes de interpretar imagens (ou até vídeos). Foram usados VLMs pré-treinados, como *BLIP2*, *PaliGemma*, *LLaVA*, *Qwen* e as variantes: *QwenV2* e *QwenV2.5*, para análise de dados de imagens e textuais, capazes de gerar descrições híbridas ricas, a partir de MRI do cérebro e textos sobre os pacientes.

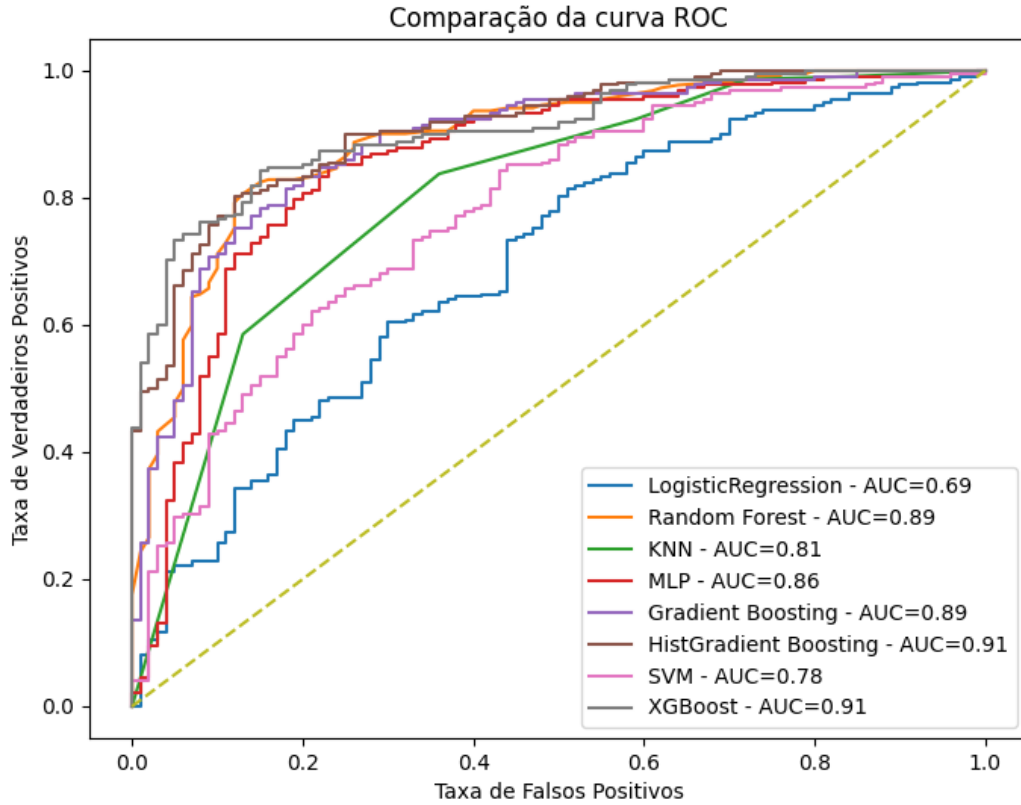
#### 5.2.1.1 Sensorial

Para cada categoria, avaliamos desempenho usando as métricas de acurácia, sensibilidade, precisão e *F1-score*, apresentadas na Tabela 13. Além disso, foi computado o gráfico que mostra a curva *ROC AUC* que relaciona a taxa de verdadeiros positivos e falsos positivos, conforme imagem 13.

A Figura 14 apresenta as matrizes de confusão, nas quais são comparados os resultados reais com as previsões feitas pelos modelos.

A Tabela 13 evidencia que os modelos de *ensemble* apresentaram o melhor desempenho geral na tarefa de classificação de pacientes com DP versus controles saudáveis. Em particular, o ***Gradient Boosting*** alcançou a acurácia mais alta (83,85 %) e o maior  $F_1$ -score (88,74 %), indicando um bom equilíbrio entre precisão e sensibilidade. Já o ***HistGradient Boosting*** obteve a maior área sob a curva ROC ( $AUC = 0,909$ ), sugerindo forte capacidade de discriminação entre as classes.

Figura 13 – Curva ROC AUC



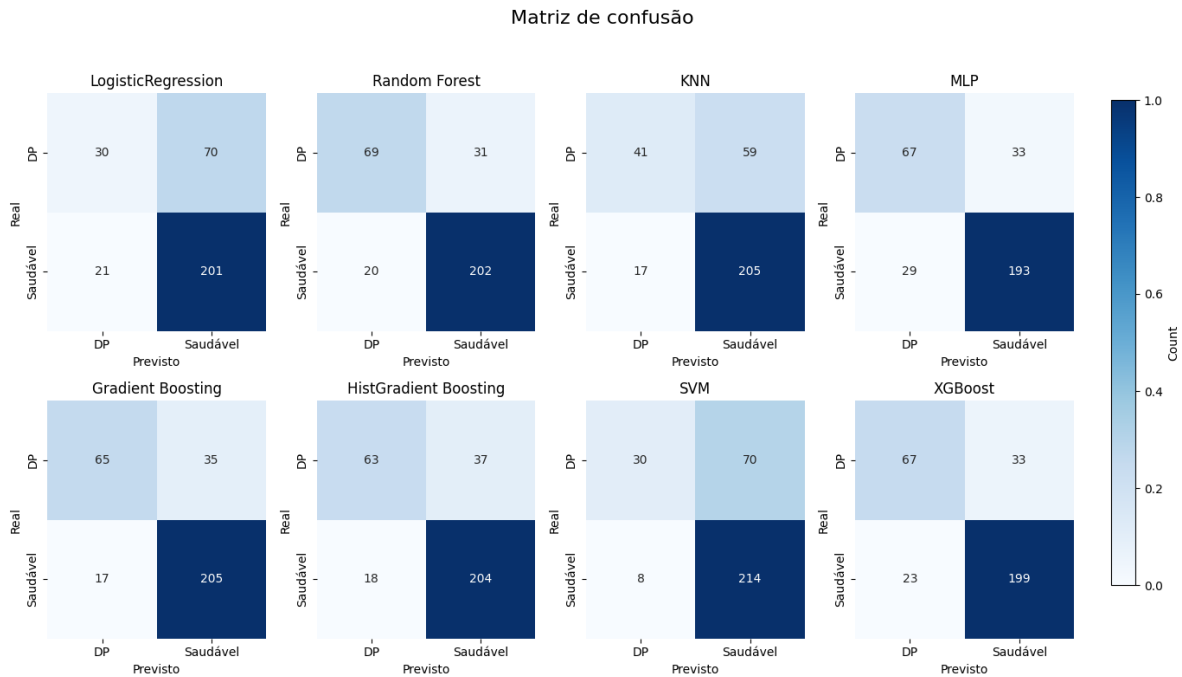
Fonte: Próprio autor.

Tabela 13 – Resultados obtidos a partir da base de dados sensorial

Modelo	Acurácia	Precisão	Sensibilidade	F1-Score	ROC AUC
Logistic Regression	0.7174	0.7417	0.9054	0.8154	0.6950
Random Forest	0.8261	0.8517	0.9054	0.8777	0.8923
KNN	0.7640	0.7765	0.9234	0.8436	0.8064
MLP	0.8168	0.8590	0.8784	0.8686	0.8573
Gradient Boosting	<b>0.8385</b>	0.8541	0.9234	<b>0.8874</b>	0.8880
HistGradient Boosting	0.8292	0.8465	0.9189	0.8812	<b>0.9093</b>
SVM	0.7578	0.7535	<b>0.9640</b>	0.8459	0.7756
XGBoost	0.8261	<b>0.8578</b>	0.8964	0.8767	0.9082

Fonte: Próprio autor.

Figura 14 – Matriz de confusão - Sensor



Fonte: Próprio autor.

Por outro lado, técnicas baseadas em distância KNN e margem SVM privilegiaram a *recall*, atingindo 92,34 % e 96,40 %, respectivamente, mas sacrificaram a precisão com (77,65 % no KNN e 75,35 % no SVM). Esse comportamento é esperado quando se busca minimizar falsos negativos, porém pode não ser o mais adequado em cenários clínicos onde o custo de um falso positivo também é relevante.

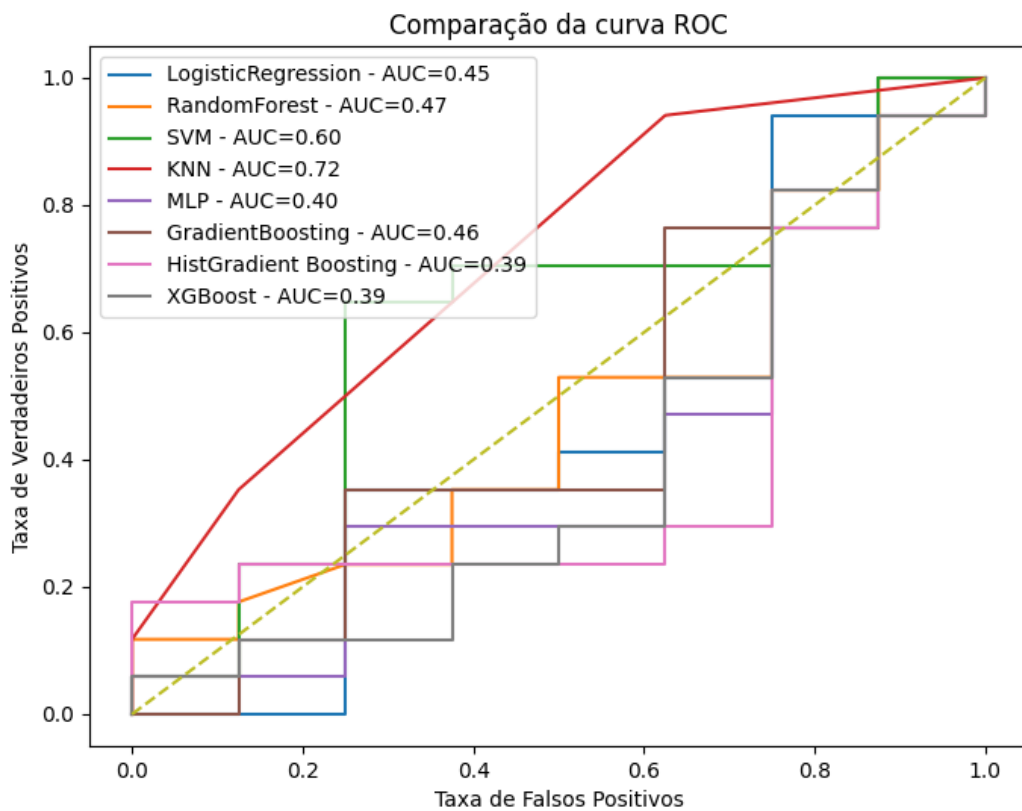
O **Random Forest** apresentou desempenho intermediário, com acurácia de 81,37 % e AUC de 0,892, confirmando sua robustez mesmo sem afinar tanto hiperparâmetros. Já o **XGBoost** com acurácia de 82,61 % e AUC de 0,908, mostrou-se competitivo com o **HistGradient Boosting**, reforçando o valor de métodos baseados em gradiente para conjuntos de dados multimodais.

Em síntese, os modelos de gradiente (*Gradient Boosting*, *HistGradient Boosting* e *XGBoost*) combinaram alta acurácia, bom F<sub>1</sub>-score e excelente AUC, destacando-se como as melhores opções para a classificação aqui proposta. As decisões entre eles podem se basear em fatores adicionais, como tempo de treinamento, interpretabilidade e sensibilidade a desbalanceamentos. Por fim, a consistência dos resultados sugere que a fusão de dados sensoriais, de áudio e de imagem fornece informação complementar e reforça a eficácia de abordagens *ensemble* em aplicações clínicas de apoio ao diagnóstico de DP.

### 5.2.1.2 Áudio

A análise do desempenho dos modelos na classificação da Doença de Parkinson (DP) utilizando dados de áudio revelou desafios específicos e uma performance geral menos robusta em comparação com os dados sensoriais. A Tabela 14 resume as métricas de acurácia, precisão, sensibilidade, *F1-Score* e *ROC AUC* para cada modelo testado, enquanto a Figura 15 ilustra as curvas *ROC AUC* e a Figura 15 apresenta as respectivas matrizes de confusão.

Figura 15 – Curva ROC AUC



Fonte: Próprio autor.

A Figura acima revela que a maioria dos modelos obteve valores de *ROC AUC* próximos ou até abaixo de 0.5. Valores de AUC próximos de 0.5 indicam que o modelo tem um poder de discriminação entre as classes não superior ao de um classificador aleatório. O modelo KNN novamente se destacou com o melhor *ROC AUC* de 0.7169, demonstrando uma capacidade discriminatória superior em comparação com os outros modelos. Ainda assim, um AUC de 0.7169, embora o melhor do grupo, é moderado e sugere que, mesmo para o melhor modelo, há espaço considerável para melhorias na separação das classes. Ao analisar, de maneira

geral, os resultados de *ROC AUC*, principalmente os valores baixos de *AUC* para a maioria dos modelos, especialmente para *HistGradient Boosting* e *XGBoost* com 0.3897, reforçam a hipótese de que as características extraídas dos dados de áudio podem não ser ideais ou que o conjunto de dados pode ser pequeno e/ou ruidoso, dificultando o aprendizado de padrões robustos para a distinção entre pacientes com e sem DP.

Tabela 14 – Resultados obtidos a partir da base de dados de áudios

Modelo	Acurácia	Precisão	Sensibilidade	F1-Score	ROC AUC
Logistic Regression	0.56	0.6667	0.7058	0.6857	0.4485
Random Forest	0.60	0.6842	0.7647	0.7222	0.4779
KNN	<b>0.76</b>	<b>0.7619</b>	<b>0.9412</b>	<b>0.8421</b>	<b>0.7169</b>
MLP	0.56	0.6667	0.7059	0.6857	0.4117
Gradient Boosting	0.56	0.6667	0.7059	0.6857	0.4044
HistGradient Boosting	0.60	0.6842	0.7647	0.7222	0.3897
SVM	0.68	0.6956	<b>0.9412</b>	0.8000	0.6029
XGBoost	0.56	0.6667	0.7059	0.6857	0.3897

Fonte: Próprio autor.

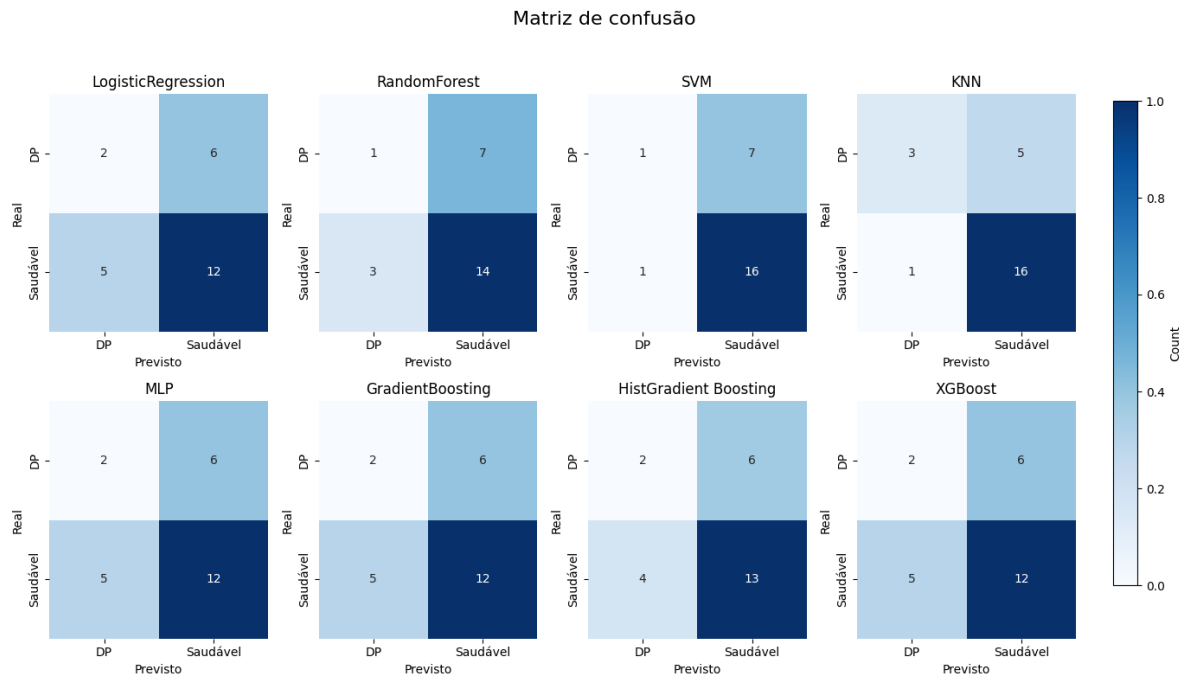
De modo geral, as acurácias dos modelos variaram entre 0.56, para *Logistic Regression*, *MLP*, *Gradient Boosting* e *XGBoost*, e 0.76, para o KNN. Este espectro de resultados é consideravelmente inferior ao observado na modalidade sensorial, indicando que a detecção de biomarcadores vocais da DP é uma tarefa mais complexa ou que as características extraídas (*Jitter*, *Shimmer*, *HNR*, *NHR*, *RPDE*, *DFA*, *PPE*) podem não ser suficientemente discriminativas para a base de dados utilizada. Com um destaque para o algoritmo KNN que apresentou a melhor acurácia, 0.76 e um *F1-Score* notável de 0.8421. Além disso, o KNN alcançou uma alta sensibilidade de 0.9412, sugerindo que ele é eficaz na identificação de pacientes com DP, verdadeiros positivos, o que é crucial em cenários de triagem onde a minimização de falsos negativos é prioritária. No entanto, sua precisão de 0.7619 é razoável, indicando que, embora identifique bem os positivos, ainda pode gerar uma quantidade moderada de falsos positivos.

Além disso, notou-se que os modelos baseados em *ensemble*, como, *Gradient Boosting*, *HistGradient Boosting*, *XGBoost*, diferentemente, do desempenho superior na modalidade sensorial, esses modelos apresentaram resultados mais modestos para dados de áudio, com acurácias em torno de 0.56 a 0.60 e *F1-Scores* de aproximadamente 0.68 a 0.72. Isso sugere que a combinação de múltiplos modelos pode não ser tão eficaz quando as características de entrada são inerentemente mais ruidosas ou menos informativas para a tarefa de classificação. Por fim, o

SVM obteve uma sensibilidade muito alta de 0.9412, similar ao KNN, mas com uma acurácia geral um pouco menor de 0.68, e a Regressão Logística e o *MLP* apresentaram os desempenhos mais baixos, com acurácia de 0.56.

Também foi feita a matriz de confusão, ilustrada na Figura 16, na qual são comparados os resultados reais com as previsões feitas pelos modelos.

Figura 16 – Matriz de confusão - Áudio



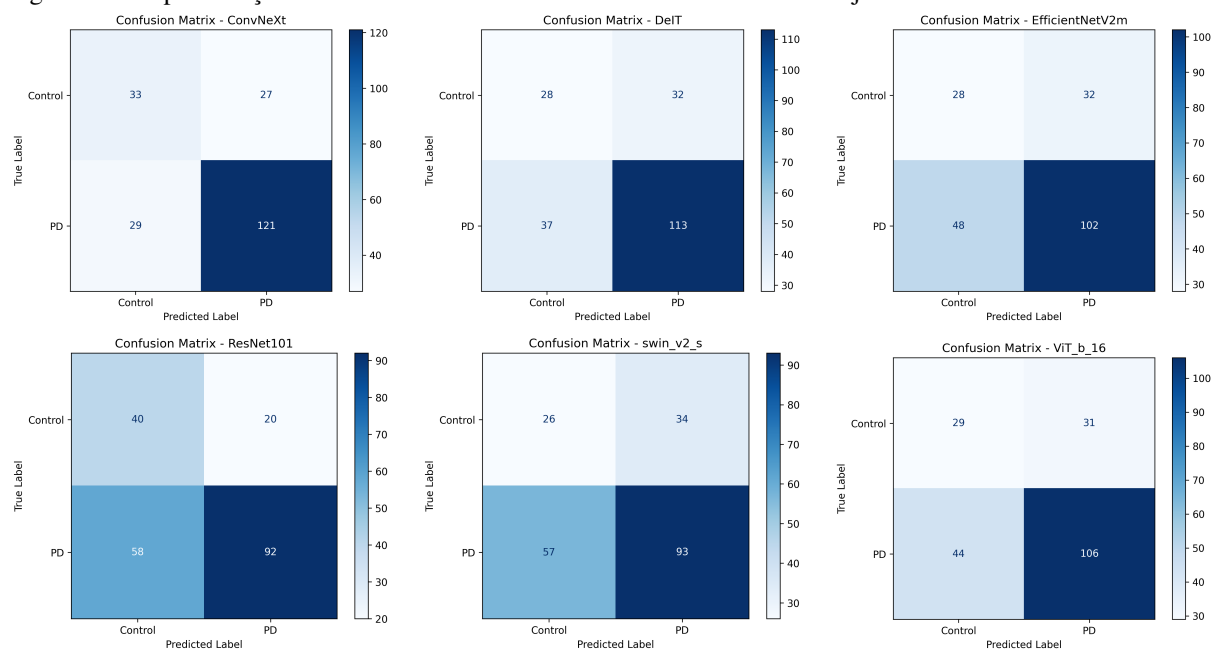
Fonte: Próprio autor.

As matrizes de confusão na Figura 16 fornecem uma visão detalhada do desempenho de cada modelo. Para o KNN (que obteve a maior acurácia), observa-se um número relativamente alto de Verdadeiros Positivos (TP) (8) e Verdadeiros Negativos (TN) (16). A quantidade de Falsos Negativos (FN) é baixa (1), o que explica sua alta sensibilidade. No entanto, o número de Falsos Positivos (FP) é de 5, indicando que, embora detecte bem a DP, pode erroneamente classificar alguns indivíduos saudáveis como tendo a doença. Para modelos com desempenho inferior, como *Logistic Regression* ou *MLP*, as matrizes mostram uma distribuição de erros mais equitativa entre Falsos Positivos e Falsos Negativos, ou uma tendência a classificar a maioria dos exemplos como a classe majoritária (se houver desbalanceamento, o que a Tabela 10 indica ser quase balanceado para áudio), o que resulta em baixa acurácia e AUC.

### 5.2.1.3 Imagens

Para as classificações a partir de imagens médicas (RM e TC), os resultados revelaram a complexidade inerente a esta modalidade, especialmente devido ao desbalanceamento de classes do conjunto de dados usado nesse trabalho e à sensibilidade a diferentes estratégias de treinamento. As Figuras abaixo fornecem as matrizes de confusão detalhadas para os modelos em diferentes abordagens de treinamento, aprendizagem por transferência, ou *transfer learning* e ajuste fino, complementadas pelas curvas de acurácia e perda, ilustradas na Figura 20 que permitem uma análise da dinâmica de aprendizado.

Figura 17 – Representação da matriz de confusão de todos os modelos usando ajuste fino.

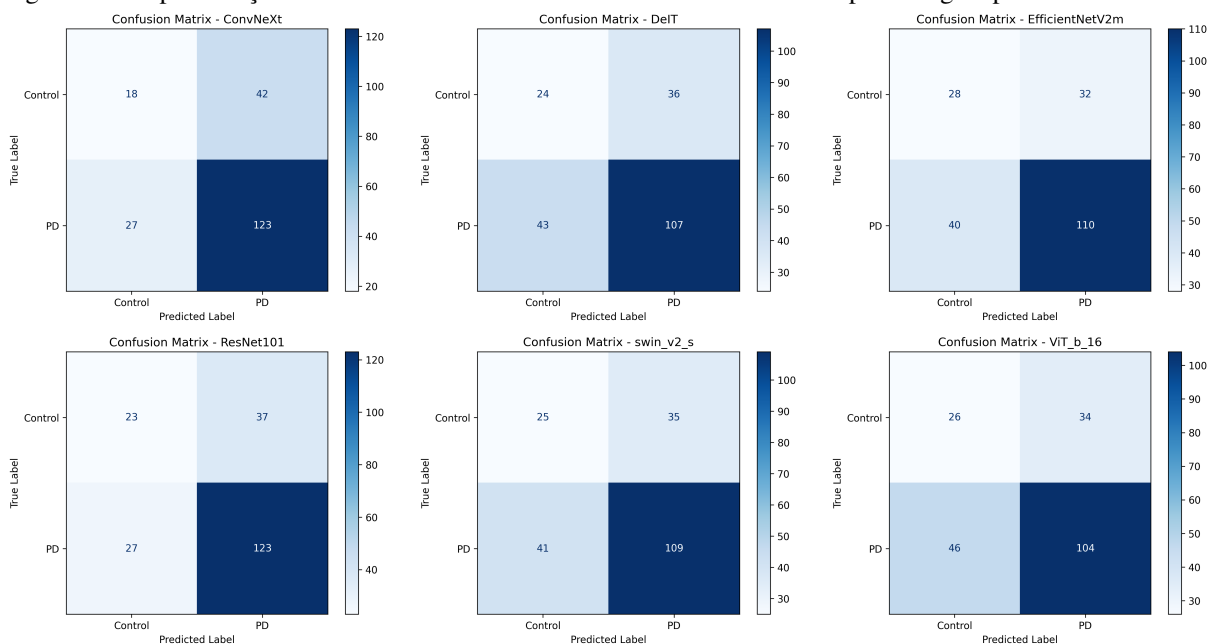


Fonte: Próprio autor.

A Figura 17 demonstra visualmente a distribuição dos Verdadeiros Positivos, Verdadeiros Negativos, Falsos Positivos e Falsos Negativos para cada arquitetura. É evidente que todos os modelos lutam mais para classificar corretamente a classe Controle em comparação com a classe DP, refletindo o desbalanceamento de classes no *dataset*. O *ConvNeXt*, em particular, exibe uma distribuição mais favorável, com um número significativamente maior de Verdadeiros Positivos para DP e um número comparativamente baixo de Falsos Negativos, tornando-o um forte candidato para aplicações clínicas na detecção da DP. Por outro lado, modelos como *Swin-V2-S* e *ResNet101* apresentam um número mais elevado de Falsos Negativos, o que poderia levar a diagnósticos perdidos em um cenário real. A análise desses erros específicos é crucial para refinar os modelos e, possivelmente, implementar estratégias de mitigação do desequilíbrio

de classes.

Figura 18 – Representação da matriz de confusão de todos os modelos usando aprendizagem por transferência.



Fonte: Próprio autor.

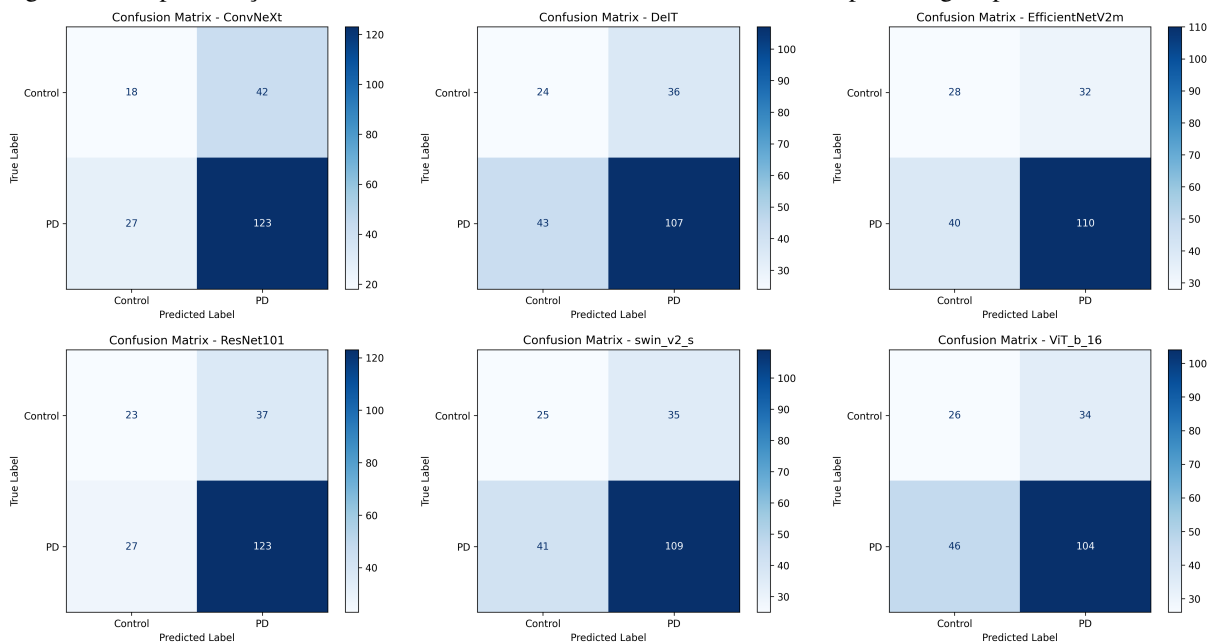
Analisando também as imagens acima usando aprendizagem por transferência, chegamos a uma análise geral, não há uma superioridade clara de uma abordagem sobre a outra para todos os modelos. Alguns modelos performaram melhor com *fine-tuning* para certas métricas, enquanto outros se beneficiaram mais do *transfer learning*. Isso sugere que a escolha da abordagem ideal é dependente da arquitetura específica do modelo e das características do *dataset*, especialmente considerando o desequilíbrio de classes.

Para complementar, foi realizada uma análise individual de cada abordagem de treinamento, por aprendizagem por transferência e ajuste fino, dos modelos para tornar possível a compreensão de aspectos cruciais sobre a dinâmica de aprendizado de cada modelo, como performance do modelo por épocas, verificar se existe presença de *overfitting* ou *underfitting*, a partir de uma inspeção visual dos gráficos. Então, a seguir apresentamos e discutimos as curvas de Perda (*Loss*) e Acurácia para os conjuntos de treinamento e validação dos modelos avaliados.

A Figura 19 apresentam as curvas de perda e acurácia para os conjuntos de treinamento e validação dos modelos testados quando treinados utilizando a abordagem de aprendizagem por transferência, onde apenas a camada classificadora foi treinada, mantendo os pesos da base pré-treinada congelados.

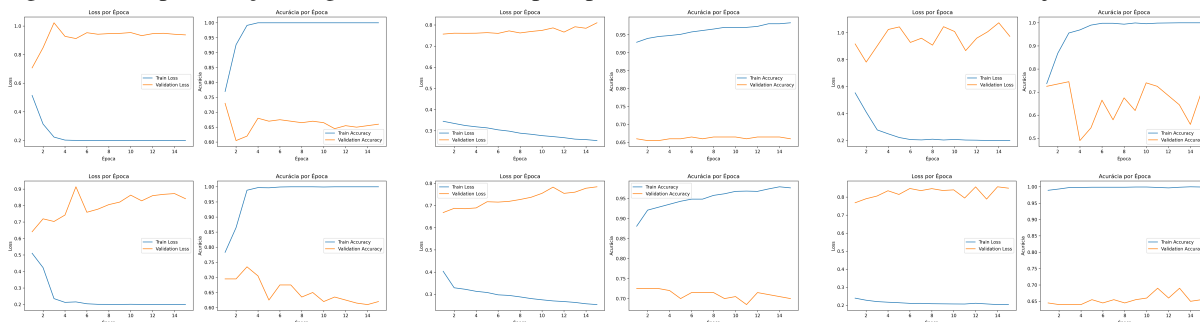
De forma análoga aos experimentos de *fine-tuning*, observa-se uma tendência genera-

Figura 19 – Representação da matriz de confusão de todos os modelos usando aprendizagem por transferência.



Fonte: Próprio autor.

Figura 20 – Representação do gráfico de acurácia por época do treinamento dos modelos usando ajuste fino.



Fonte: Próprio autor.

lizada de *overfitting* também nesta abordagem. A *Train Loss* decresce consistentemente ao longo das épocas, atingindo valores baixos para todos os modelos, enquanto a *Train Accuracy* ascende de maneira robusta, alcançando patamares elevados, tipicamente acima de 0.80, e em alguns casos, próximo a 0.90 ou mais. Este comportamento indica que os modelos estão aprendendo eficientemente a mapear as características de entrada para os rótulos de saída no conjunto de treinamento.

No entanto, a *Validation Loss* para a maioria dos modelos com *transfer learning* mantém-se em um patamar significativamente mais elevado que a *Train Loss* e, em diversos casos, como: *ViT\_B\_16*, *Swin\_V2\_S*, *EfficientNetV2m*, *DeIT*, *ConvNeXt\_B*, exibe uma tendência de aumento gradual após as primeiras épocas, ou permanece estagnada em um valor alto. Concomitantemente, a *Validation Accuracy* é consistentemente inferior à *Train Accuracy*, variando aproximadamente entre 0.60 e 0.75 para a maioria das arquiteturas. Embora em alguns

modelos, como a rede *ResNet101*, a *Validation Accuracy* possa mostrar maior estabilidade, a lacuna persistente entre as curvas de treinamento e validação é um diagnóstico claro de que, mesmo com a base pré-treinada, a camada classificadora está superajustando os dados de treinamento e falhando em generalizar eficazmente para o conjunto de validação. A complexidade intrínseca dos modelos pré-treinados, aliada à capacidade limitada da camada classificadora de capturar a variabilidade completa do *dataset* reduzido e desbalanceado, contribui para esse cenário, onde o modelo memoriza em vez de aprender representações robustas para a distinção entre classes Controle e DP.

Tabela 15 – Resultados obtidos a partir da base de dados de imagens

Modelo	F1-Score (Control)	F1-Score (PD)	F1-Score Médio (Geral)
EfficientNetV2-Medium	0.41	0.72	0.565
Swin-V2-Small	0.36	0.67	0.515
ConvNeX-Tiny	<b>0.54</b>	<b>0.81</b>	<b>0.675</b>
ViT-B-16	0.43	0.74	0.585
ResNet101	0.50	0.71	0.605
DeiT	0.44	0.76	0.600

Fonte: Próprio autor.

A Tabela 15 apresenta um resumo comparativo do *F1-Score* para cada classe Controle e DP e a média geral, *F1-Score* Médio, obtida pelos diferentes modelos de classificação avaliados. Podemos observar que, de forma geral, os modelos apresentaram um desempenho superior na classificação da classe DP em comparação com a classe Controle. Este comportamento pode ser atribuído a desbalanceamento de classes no conjunto de dados ou a características intrínsecas das amostras de Parkinson que são mais distintivas para o modelo aprender. O modelo ConvNeXt se destacou, apresentando os maiores *F1-Scores* em ambas as classes, indicando uma robustez superior na identificação de pacientes com Parkinson e indivíduos saudáveis. Modelos como *EfficientNetV2m* e *ViT-B-16*, embora competentes, demonstraram uma lacuna maior entre o desempenho das classes, sugerindo a necessidade de aumentar o fator de aumento de dados para a classe minoritária, controle, em futuras investigações ou até mesmo utilizar funções de perda que penalize *datasets* desbalanceados.

### 5.2.2 *Análise Comparativa das Abordagens de Treinamento: Ajuste Fino vs. Aprendizado por Transferência*

A comparação entre as abordagens de Ajuste Fino (apresentadas nas Figuras 20 anteriores) e Aprendizado por Transferência (Figuras 18 e 19) revela insights cruciais sobre a adaptação de modelos pré-treinados para a tarefa de classificação de Parkinson em um domínio específico.

A partir dos gráficos, notam-se pontos fortes na abordagem de aprendizado por transferência com camadas congeladas, pois é particularmente vantajosa quando o dataset é pequeno, pois reduz drasticamente o número de parâmetros a serem treinados. Isso mitiga, em tese, o risco de *overfitting*, uma vez que as características de alto nível aprendidas no pré-treinamento são mantidas. Para a classe DP, foi observado que alguns modelos, como *ResNet101* e *ConvNeXt\_B* em aprendizado por transferência, apresentaram um número elevado de Verdadeiros Positivos e um menor número de Falsos Negativos, indicando que a base de características pré-treinada é eficaz na identificação de padrões de Parkinson.

No entanto, também traz consigo alguns pontos negativos, pois apesar da menor suscetibilidade teórica ao *overfitting* em comparação com o ajuste fino irrestrito, os resultados demonstram que o *overfitting* ainda é um desafio significativo, com a *Validation Accuracy* permanecendo em um patamar baixo e a *Validation Loss* divergindo. A desvantagem central reside na rigidez da base pré-treinada. Se as características aprendidas na tarefa original de pré-treinamento não forem perfeitamente alinhadas com as características discriminativas para a detecção de Parkinson no *dataset* específico, a camada classificadora tem um espaço limitado para compensar essa lacuna de representação. Isso pode resultar em um teto de desempenho para a generalização e uma incapacidade de aprender nuances específicas do domínio que poderiam ser capturadas por um ajuste fino mais flexível.

Em suma, enquanto o ajuste fino oferece maior flexibilidade e potencial para alcançar uma performance de treinamento quase perfeita, ele é mais propenso a um *overfitting* grave em *datasets* pequenos e desbalanceados. O aprendizado por transferência, por outro lado, pode proporcionar uma base de características mais estável, mas a rigidez dessas características pré-treinadas pode limitar a capacidade do modelo de se adaptar totalmente ao domínio específico e generalizar de forma mais eficaz para ambas as classes, especialmente a minoritária. Os resultados sugerem que uma estratégia intermediária, como o ajuste fino com taxas de aprendizado muito baixas para as camadas pré-treinadas ou a aplicação mais rigorosa de técnicas

de regularização, poderia ser um caminho promissor para mitigar o *overfitting* e aprimorar a generalização em futuros experimentos.

## 6 CONCLUSÃO

Neste trabalho, foi proposta uma abordagem multimodal inovadora para o diagnóstico da DP, estruturada sobre uma arquitetura agentiva colaborativa que integra dispositivos de IoT e modelos avançados de linguagem médica. A solução desenvolvida visa combinar e analisar dados provenientes de múltiplas modalidades, dos tipos, sensoriais, de áudio, imagem e, futuramente, texto clínico, de forma sinérgica, ampliando a capacidade de detecção precoce e monitoramento da doença. A proposta inicial contemplava a avaliação de três categorias de agentes inteligentes. Primeiro, algoritmos clássicos de ML aplicados a dados sensoriais e tabulares. Segundo, redes neurais profundas voltadas para a análise de imagens médicas; e, por último, modelos de linguagem pré-treinados especializados em dados clínicos textuais. Essa expectativa foi parcialmente atendida, com resultados promissores nas modalidades já exploradas, evidenciando o potencial da abordagem colaborativa e distribuída para aplicações em ambientes clínicos conectados.

Nossas análises demonstraram que, para os dados sensoriais, os métodos de ensemble baseados em gradiente, tais quais, *Gradient Boosting*, *HistGradient Boosting* e *XGBoost*, alcançam as melhores combinações de acurácia,  $F_1$ -score e *AUC*, com o *HistGradient Boosting* atingindo uma *AUC* notável de 0.909. Isso indica sua forte capacidade discriminatória e robustez para a classificação da DP. Além disso, mostramos que agentes voltados para maximizar a sensibilidade, como KNN e SVM, podem ser úteis em cenários clínicos de triagem, onde a minimização de falsos negativos é prioritária, mesmo que à custa de menor precisão.

Para a modalidade de áudio, os resultados foram mais desafiadores. Embora o KNN tenha se destacado com a melhor acurácia e *F1-score*, a performance geral dos modelos, especialmente evidenciada por baixos valores de *ROC AUC*, muitos próximos ou abaixo de 0.5, sugere que a extração de características acústicas ou a qualidade dos dados de áudio podem ser limitantes. Este achado ressalta a complexidade de capturar biomarcadores vocais da DP de forma robusta e indica uma área crítica para futuros aprimoramentos.

Na análise de imagens, a aplicação de redes neurais convolucionais e transformadores de visão, tanto com ajuste fino quanto com transferência por aprendizado, revelou que o desbalanceamento de classes é um desafio persistente, impactando a classificação da classe Controle. O modelo *ConvNeXt-Tiny* demonstrou o melhor equilíbrio, com os maiores *F1-Scores* para ambas as classes, sublinhando seu potencial para aplicações clínicas. No entanto, a análise das curvas de perda e acurácia evidenciou um desafio comum de *overfitting* em ambas as abordagens

de treinamento, mesmo com a transferência por aprendizado, indicando que os modelos tendem a memorizar os dados de treinamento em vez de generalizar eficazmente para dados não vistos.

Os resultados reforçam a viabilidade de explorar múltiplas fontes de informação de forma complementar, indicando que a fusão adequada de características sensoriais, acústicas e visuais potencializa a capacidade de discriminação entre pacientes com DP e controles saudáveis. A robustez dos modelos de gradiente para dados sensoriais e o desempenho promissor de arquiteturas como *ConvNeXt* para imagens suportam seu uso em sistemas de apoio à decisão clínica. A futura integração dos resultados provenientes da análise textual, conduzida por agentes especializados baseados em modelos de LLMs, é projetada para complementar e potencializar significativamente o poder preditivo do sistema multimodal. Ao incorporar informações semânticas extraídas de registros clínicos, anotações médicas e outros dados textuais, espera-se que o sistema alcance uma visão diagnóstica mais abrangente e contextualizada, reforçando a precisão e a confiabilidade da arquitetura colaborativa proposta.

### **6.0.1 Discussão, Oportunidades e Questões em aberto**

A partir do desenvolvimento e experimentação da plataforma multimodal agentiva para diagnóstico da DP, emergem importantes lições aprendidas, bem como oportunidades e desafios ainda em aberto, que poderão guiar estudos futuros.

#### **Lições Aprendidas:**

- A integração de modalidades sensoriais heterogêneas (áudio, imagem, sensores e texto) revela grande potencial, mas exige pipelines específicos de pré-processamento e extração de características adaptados à natureza de cada dado.
- Agentes baseados em RAG multimodal são capazes de realizar inferências contextualizadas sem necessidade de re-treinamento total, o que é crucial para cenários com dados escassos e em constante atualização.
- A execução distribuída entre borda e nuvem demonstrou-se eficiente em termos de latência, quando associada à compressão de modelos e separação de etapas computacionais.

#### **Desafios Abertos:**

- Fusão multimodal: lidar com o balanceamento entre modalidades, propagação de ruído e coordenação entre agentes em decisões complexas ainda demanda pesquisa.
- A ausência de arquiteturas abertas, padronizadas e reutilizáveis limita a reprodutibilidade experimental e dificulta a comparação sistemática de desempenho entre diferentes aborda-

gens propostas na literatura. Como continuidade deste trabalho, propõe-se a definição de uma arquitetura de referência para plataformas neurointeligentes, fundamentadas em mecanismos de RAG e agentes colaborativos, visando estabelecer bases sólidas para futuras pesquisas e aplicações clínicas.

- Padronização e interoperabilidade: ainda há lacunas na adaptação a formatos médicos como DICOM, HL7-FHIR e integração segura com repositórios como a RNDS (Rede Nacional de Dados em Saúde).

## 6.1 Trabalhos Futuros

Nas próximas etapas, será finalizada a análise da modalidade textual, incorporando extrações semânticas derivadas de agentes LLM para enriquecer o entendimento de prontuários clínicos e anotações médicas. Em paralelo, serão abordadas em estratégias de fusão multimodal, desde a combinação de vetores de característica, *feature-level fusion*, até métodos de decisão conjunta ou em inglês, *decision-level fusion*, e arquiteturas ponta a ponta, visando integrar de forma precisa todas as fontes de informação.

Também serão investigadas e aplicadas técnicas robustas para mitigar o *overfitting* em modelos de imagem, incluindo regularização avançada, aumentos de dados específicos para a classe controle e funções de perda adaptadas a *datasets* desbalanceados. Na linha do áudio, vai ser otimizado o pré-processamento e a extração de atributos para gerar representações mais discriminativas e elevar o desempenho desses modelos.

Em seguida, desenvolveremos um protótipo de sistema de apoio clínico totalmente integrado, validado em colaboração com profissionais de saúde para assegurar aplicabilidade real. Paralelamente, será aplicados métodos de *Explainable AI* (XAI) para revelar quais características multimodais mais influenciam o diagnóstico, fortalecendo a confiança e aceitação clínica. Por fim, realizaremos estudos longitudinais em pacientes para monitorar a progressão da DP e avaliar a capacidade preditiva dos modelos diante de diferentes tratamentos.

## REFERÊNCIAS

ABOOTORABI, M. M.; ZOBEIRI, A.; DEHGHANI, M.; MOHAMMADKHANI, M.; MOHAMMADI, B.; GHAHROODI, O.; BAGHSHAH, M. S.; ASGARI, E. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. **arXiv preprint**, 2025. Disponível em: <https://arxiv.org/pdf/2502.08826>. Acesso em: 19 abr. 2025.

AKILA, B.; NAYAH, J. J. V. Parkinson classification neural network with mass algorithm for processing speech signals. **Neural Computing and Applications**, New York, NY, v. 36, n. 17, p. 10165–10181, 2024.

ALAYRAC, J.-B.; DONAHUE, J.; LUC, P.; MIECH, A.; BARR, I.; HASSON, Y.; LENC, K.; MENSCH, A.; MILLICAN, K.; REYNOLDS, M. *et al.* Flamingo: a visual language model for few-shot learning. *In: ANNUAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NEURIPS)*, 36., 2022, New Orleans, Louisiana, USA. *Proceedings [...]*. San Diego, CA, USA: Neural Information Processing Systems Foundation, Inc., 2022. p. 23716–23736.

ALI-ELDIN, A.; WANG, B.; SHENOY, P. The hidden cost of the edge: a performance comparison of edge and cloud latencies. *In: INTERNATIONAL CONFERENCE FOR HIGH PERFORMANCE COMPUTING, NETWORKING, STORAGE AND ANALYSIS*, 2021, St. Louis. *Proceedings [...]*. New York, NY: Association for Computing Machinery, 2021. p. 1–12.

ASSIS, A. C.; CARVALHO, P. G. B. de; MALAQUIAS, D. T. M.; ALMEIDA, I. C.; MACIEL, C. V.; FRANCO, C. E. F.; MELO, K. R. S. T.; SOUZA, A. G. R.; ABREU, G. P.; MIRANDA, F. A. C. de P. *et al.* Uso de algoritmos para diagnóstico precoce de alzheimer, parkinson e esclerose múltipla. **LUMEN ET VIRTUS**, v. 15, n. 43, p. 9060–9069, 2024.

AYARI, A.; HAMDI, H.; ALSULBI, K. A. E-health application in iomt environment deployed in an edge and cloud computing platforms. **Procedia Computer Science**, Amsterdam, v. 246, p. 1019–1028, 2024.

BASIT, A.; HUSSAIN, K.; HANIF, M. A.; SHAFIQUE, M. Medaide: leveraging large language models for on-premise medical assistance on edge devices. **arXiv preprint**, 2024. Disponível em: <https://arxiv.org/pdf/2403.00830>. Acesso em: 15 abr. 2025.

BENGIO, Y.; DUCHARME, R.; VINCENT, P.; JAUVIN, C. A neural probabilistic language model. **Journal of machine learning research**, Cambridge, MA, v. 3, n. Feb, p. 1137–1155, 2003.

BERNARDI, É. A. T. **Padrão de intercomunicação em saúde e prontuário eletrônico do paciente**. [S. l.]: Editora Senac São Paulo, 2024.

BRASIL. Ministério da Saúde. **11/4 – Dia Mundial de Conscientização da Doença de Parkinson**. 2023. Disponível em: <https://bvsmis.saude.gov.br/?p=4893>. Acesso em: 11 abr. 2025.

CHAUHAN, M. K.; GHOSAL, P. A hybrid cnn-bilstm neural network architecture for early prediction of parkinson’s disease. *In: IEEE INTERNATIONAL SYMPOSIUM ON SMART ELECTRONIC SYSTEMS*, 10., 2024, New Delhi. *Proceedings [...]*. New Delhi: IEEE, 2024. p. 303–308.

CHEN, H.; FU, J.; LIU, X.; ZHENG, Z.; LUO, X.; ZHOU, K.; XU, Z.; GENG, D. A parkinson's disease-related nuclei segmentation network based on cnn-transformer interleaved encoder with feature fusion. **Computerized Medical Imaging and Graphics**, Amsterdam, v. 118, p. 102465, 2024. Disponível em: <https://doi.org/10.1016/j.compmedimag.2024.102465>. Acesso em: 23 fev. 2025.

CHEN, W.; HU, H.; CHEN, X.; VERGA, P.; COHEN, W. W. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. **arXiv preprint**, 2022. Disponível em: <https://arxiv.org/pdf/2210.02928>. Acesso em: 10 mai. 2025.

CHO, Y.-M.; SHU, R.; DAS, N.; ALKHOULI, T.; LAI, Y.-A.; CAI, J.; SUNKARA, M.; ZHANG, Y. Roundtable: Investigating group decision-making mechanism in multi-agent collaboration. **arXiv preprint**, 2024. Disponível em: <https://arxiv.org/pdf/2411.07161>. Acesso em: 5 abr. 2025.

CHOMSKY, N. **Syntactic structures**. [S. l.]: Mouton de Gruyter, 2002.

DETTMERS, T.; PAGNONI, A.; HOLTZMAN, A.; ZETTLEMOYER, L. Qlora: Efficient finetuning of quantized llms. In: ANNUAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NEURIPS), 37., 2023, New Orleans, Louisiana, USA. *Proceedings [...]*. San Diego, CA, USA: Neural Information Processing Systems Foundation, Inc., 2023. p. 10088–10115.

DEVI, K. V. R.; SMITHA, B. S.; LAKHANPAL, S.; KALRA, R.; SETHI, V. A.; THAJIL, S. K. A review: Swarm robotics: Cooperative control in multi-agent systems. EDP Sciences, Cape Town, South Africa, v. 505, p. 03013, 2024. Disponível em: <https://doi.org/10.1051/e3sconf/202450503013>. Acesso em: 23 mai. 2025.

DEVLIN, J. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint**, 2018. Disponível em: <https://arxiv.org/pdf/1810.04805>. Acesso em: 20 ago. 2025.

DIAZ-DE-ARCAYA, J.; LÓPEZ-DE-ARMENTIA, J.; MIÑÓN, R.; OJANGUREN, I. L.; TORRE-BASTIDA, A. I. Large language model operations (llmops): Definition, challenges, and lifecycle management. In: INTERNATIONAL CONFERENCE ON SMART AND SUSTAINABLE TECHNOLOGIES (SPLITECH), 9., 2024, Bol and Split, Croatia. *Proceedings [...]*. Bol and Split: IEEE, 2024. p. 499–502.

DONG-CHEN, X.; YONG, C.; YANG, X.; CHEN-YU, S.; LI-HUA, P. Signaling pathways in parkinson's disease: molecular mechanisms and therapeutic interventions. **Signal transduction and targeted therapy**, London, v. 8, n. 1, p. 73, 2023.

DORRI, A.; KANHERE, S. S.; JURDAK, R. Multi-agent systems: A survey. **IEEE Access**, Piscataway, NJ, v. 6, p. 28573–28593, 2018.

DUQUE, K. R.; VIZCARRA, J. A.; HILL, E. J.; ESPAY, A. J. Disease-modifying vs symptomatic treatments: Splitting over lumping. **Handbook of clinical neurology**, Amsterdam, v. 193, p. 187–209, 2023.

EDGE, D.; TRINH, H.; CHENG, N.; BRADLEY, J.; CHAO, A.; MODY, A.; TRUITT, S.; METROPOLITANSKY, D.; NESS, R. O.; LARSON, J. From local to global: A graph rag approach to query-focused summarization. **arXiv preprint**, 2024. Disponível em: <https://arxiv.org/pdf/2404.16130>. Acesso em: 20 ago. 2025.

EHTESHAM, A.; SINGH, A.; KUMAR, S. Enhancing clinical decision support and ehr insights through llms and the model context protocol: An open-source mcp-fhir framework. **arXiv preprint**, 2025. Disponível em: <https://arxiv.org/pdf/2506.13800>. Acesso em: 23 jun. 2025.

ERETH, J. Dataops-towards a definition. **LWDA**, v. 2191, p. 104–112, 2018.

ES, S.; JAMES, J.; ESPINOSA-ANKE, L.; SCHOCKAERT, S. Ragas: Automated evaluation of retrieval augmented generation. **arXiv preprint**, 2023. Disponível em: <https://arxiv.org/pdf/2309.15217>. Acesso em: 12 mai. 2025.

FILIPPI, M.; BALESTRINO, R.; AGOSTA, F. Biomarkers of non-motor symptoms in parkinson's disease and parkinsonisms. **Frontiers in Neurology**, v. 14, p. 1257064, 2023. Disponível em: <https://www.frontiersin.org/journals/neurology/articles/10.3389/fneur.2023.1257064>. Acesso em: 15 mai. 2025.

FOERSTER, J.; FARQUHAR, G.; AFOURAS, T.; NARDELLI, N.; WHITESON, S. Counterfactual multi-agent policy gradients. *In: AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE*, 32., 2018, New Orleans, Louisiana, USA. *Proceedings [...]*. New Orleans, Louisiana, USA: [s. n.], 2018. p. 2974 – 2982.

FRANKLIN, S.; GRAESSER, A. Is it an agent, or just a program?: A taxonomy for autonomous agents. *In: INTERNATIONAL WORKSHOP ON AGENT THEORIES, ARCHITECTURES, AND LANGUAGES*, 3., 1996, Budapest, Hungary. *Proceedings [...]*. Budapest, Hungary: Springer-Verlag, 1996. p. 21–35.

GAO, L.; GAO, K.; SUN, W.; ZHANG, Y.; ZHANG, R.; ZHANG, L. Effect simulation and local adaptation of multi-agent collaborative governance in marine eco-economic systems: Evidence from china. **Ecological Economics**, Amsterdam, v. 228, p. 108451, 2025. Disponível em: <https://doi.org/10.1016/j.ecolecon.2024.108451>. Acesso em: 15 fev. 2025.

GAWADE, S.; AKHOURI, S.; KULKARNI, C.; SAMANT, J.; SAHU, P.; PAHAL, J.; MEHER, S. *et al.* Multi agent based medical assistant for edge devices. **arXiv preprint**, 2025. Disponível em: <https://arxiv.org/pdf/2503.05397>. Acesso em: 28 abr. 2025.

GBD Parkinson's Disease Collaborators. Global, regional, and national burden of parkinson's disease and its attributable risk factors, 1990–2021: a systematic analysis for the global burden of disease study. **The Lancet Neurology**, v. 23, n. 4, p. 123–145, 2024. Disponível em: [https://doi.org/10.1016/S1474-4422\(23\)00456-7](https://doi.org/10.1016/S1474-4422(23)00456-7). Acesso em: 11 abr. 2025.

GELDEREN, L. van; TEJEDOR-GARCÍA, C. Innovative speech-based deep learning approaches for parkinson's disease classification: A systematic review. **arXiv preprint**, 2024. Disponível em: <https://arxiv.org/pdf/2407.17844>. Acesso em: 20 fev. 2025.

GHAYVAT, H.; PANDYA, S. *et al.* Revolutionizing healthcare: Iomt-enabled digital enhancement via multimodal adl data fusion. **Information Fusion**, Amsterdam, v. 111, p. 102518, 2024. Disponível em: <https://doi.org/10.1016/j.inffus.2024.102518>. Acesso em: 15 mai. 2025.

GHEZLOO, F.; SEYFIOGLU, M. S.; SORAKI, R.; IKEZOGWO, W. O.; LI, B.; VIVEKANANDAN, T.; ELMORE, J. G.; KRISHNA, R.; SHAPIRO, L. Pathfinder: A multi-modal multi-agent system for medical diagnostic decision-making applied to histopathology. **arXiv preprint**, 2025. Disponível em: <https://arxiv.org/pdf/2502.08916>. Acesso em: 25 fev. 2025.

GOEL, A.; NEDUNCHELIYAN, S. Health monitoring and diagnostic platform based on ai and iomt sensors: An overview of methodologies and challenges. *In: INTERNATIONAL CONFERENCE ON TRENDS IN ENGINEERING SYSTEMS AND TECHNOLOGIES (ICTEST)*, 1., 2024, Kochi. *Proceedings [...]*. Kochi: IEEE, 2024. p. 216–220.

GOLDBARG, M. A. S. S. (**Compressão consciente de modelos de redes neurais profundas: baseada em poda seguida de quantização**). Dissertação (Mestrado em Engenharia Elétrica) – Universidade Federal do Rio Grande do Norte, Natal, 2024.

HE, T.; CHEN, J.; HOSSAIN, M. S.; LYU, Z. Enhanced detection of early parkinson’s disease through multi-sensor fusion on smartphone-based iomt platforms. **Information Fusion**, Amsterdam, v. 117, p. 102889, 2025. Disponível em: <https://doi.org/10.1016/j.inffus.2024.102889>. Acesso em: 15 mar. 2025.

HENDRIKS, M.; VINKE, R. S.; GEORGIEV, D. Gender discrepancies and differences in motor and non-motor symptoms, cognition, and psychological outcomes in the treatment of parkinson’s disease with subthalamic deep brain stimulation. **Frontiers in Neurology**, Lausanne, v. 14, p. 1257781, 2024. Disponível em: <https://doi.org/10.3389/fneur.2023.1257781>. Acesso em: 24 mai. 2025.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural computation**, v. 9, n. 8, p. 1735–1780, 1997.

HUANG, C.; WANG, J.; WANG, S.; ZHANG, Y. Internet of medical things: A systematic review. **Neurocomputing**, Amsterdam, v. 557, p. 126719, 2023.

HUANG, Y.; ZHANG, J.; WANG, Z.; BIE, B.; HU, X.; HE, X. *et al.* Repl: Recalibrating perplexity by uncertainty in semantic propagation and language generation for explainable qa hallucination detection. **arXiv preprint**, 2025. Disponível em: <https://arxiv.org/pdf/2505.15386>. Acesso em: 23 mai. 2025.

HUTCHINS, W. J. The georgetown-ibm experiment demonstrated in january 1954. *In: CONFERENCE OF THE ASSOCIATION FOR MACHINE TRANSLATION IN THE AMERICAS*, 6., 2004, Washington, USA. *Proceedings [...]*. Washington, DC: Springer, 2004. p. 102–114.

ISLAM, U.; ALATAWI, M. N.; ALQAZZAZ, A.; ALAMRO, S.; SHAH, B.; MOREIRA, F. A hybrid fog-edge computing architecture for real-time health monitoring in iomt systems with optimized latency and threat resilience. **Scientific Reports**, London, v. 15, n. 1, p. 25655, 2025. Disponível em: <https://doi.org/10.1038/s41598-025-09696-3>. Acesso em: 4 mai. 2025.

IUCKSCH, D. D.; SIEGA, J.; LEVECK, G. C.; ARAUJO, L. B. d.; MÉLO, T. R.; ISRAEL, V. L. *et al.* Improvement of balance, motor aspects, and activities of daily living in parkinson’s disease after a sequential multimodal aquatic-and land-based intervention program. **Rehabilitation Research and Practice**, v. 2023, p. 2762863, 2023. Disponível em: <https://onlinelibrary.wiley.com/doi/pdf/10.1155/2023/2762863>. Acesso em: 23 abr. 2025.

JEONG, C. Beyond text: Implementing multimodal large language model-powered multi-agent systems using a no-code platform. **arXiv preprint**, 2025. Disponível em: <https://arxiv.org/pdf/2501.00750>. Acesso em: 25 abr. 2025.

JÚNIOR, E. P.; DELMIRO, I. L.; MAGAIA, N.; MAIA, F. M.; HASSAN, M. M.; ALBUQUERQUE, V. H. C.; FORTINO, G. Intelligent sensory pen for aiding in the diagnosis of parkinson's disease from dynamic handwriting analysis. **Sensors**, Basel, v. 20, n. 20, p. 5840, 2020. Disponível em: <https://doi.org/10.3390/s20205840>. Acesso em: 19 fev. 2025.

KARPUKHIN, V.; OĞUZ, B.; MIN, S.; LEWIS, P.; WU, L.; EDUNOV, S.; CHEN, D.; YIH, W.-t. Dense passage retrieval for open-domain question answering. **arXiv preprint**, 2020. Disponível em: <https://arxiv.org/pdf/2004.04906>. Acesso em: 15 mai. 2025.

KHANOM, F.; UDDIN, M. S.; MOSTAFIZ, R. Pd\_ebm: An integrated boosting approach based on selective features for unveiling parkinson's disease diagnosis with global and local explanations. **Engineering Reports**, Hoboken, NJ, v. 7, n. 1, p. e13091, 2025. Disponível em: <https://doi.org/10.1002/eng2.13091>. Acesso em: 20 fev. 2025.

KHARCHOUF, I.; MOHAMMED, O. A. Controller hardware-in-the-loop testbed of a distributed consensus multi-agent system control under deception and disruption cyber-attacks. **Energies**, Basel, v. 17, n. 7, p. 1669, 2024.

KHATUN, M. A.; MEMON, S. F.; EISING, C.; DHIRANI, L. L. Machine learning for healthcare-iot security: A review and risk mitigation. **IEEE Access**, Piscataway, NJ, v. 11, p. 145869–145896, 2023. Disponível em: <https://doi.org/10.1109/ACCESS.2023.3346320>. Acesso em: 21 mai. 2025.

KHUSHBU, A.-N.; YANG, Z.; LIU, Y.; ZHANG, X. Parkinson's disease recognition using hybrid cnn-transformer model. *In: INTERNATIONAL CONFERENCE ON INDUSTRIAL AUTOMATION, ROBOTICS AND CONTROL ENGINEERING (IARCE)*, 4., 2024, Chengdu, China. *Proceedings [...]*. Chengdu: IEEE, 2024. p. 307–311.

KIM, Y.; PARK, C.; JEONG, H.; CHAN, Y. S.; XU, X.; MCDUFF, D.; LEE, H.; GHASSEMI, M.; BREAZEAL, C.; PARK, H. W. Mdagents: An adaptive collaboration of llms for medical decision-making. *In: ANNUAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NEURIPS)*, 38., 2024, Vancouver, Canada. *Proceedings [...]*. Vancouver, Canada: Curran Associates, Inc., 2024.

KINGER, S. B.; JUNEAU, T.; KAPLAN, R. I.; PLUIM, C. F.; FOX-FULLER, J. T.; WANG, T.; MUKADAM, N.; NEARGARDER, S.; SALAZAR, R. D.; CRONIN-GOLOMB, A. Changes in apathy, depression, and anxiety in parkinson's disease from before to during the covid-19 era. **Brain Sciences**, Basel, v. 13, n. 2, p. 199, 2023. Disponível em: <https://doi.org/10.3390/brainsci13020199>. Acesso em: 28 mar. 2025.

KRAEMER, L.; BANERJEE, B. Multi-agent reinforcement learning as a rehearsal for decentralized planning. **Neurocomputing**, Amsterdam, v. 190, p. 82–94, 2016.

KUMAR, A.; SAMRIYA, J. K.; BHANSALI, A.; MALIK, M.; ARYA, V.; ALHALABI, W.; ALHARBI, E.; GUPTA, B. B. Energy efficient task offloading from iomt in delay sensitive edge networks. **Alexandria Engineering Journal**, Amsterdam, v. 128, p. 476–483, 2025.

LAURETANI, F.; TESTA, C.; SALVI, M.; ZUCCHINI, I.; GIALLAURIA, F.; MAGGIO, M. Clinical evaluation of sleep disorders in parkinson's disease. **Brain Sciences**, Basel, v. 13, n. 4, p. 609, 2023. Disponível em: <https://doi.org/10.3390/brainsci13040609>. Acesso em: 20 mai. 2025.

LEWIS, P.; PEREZ, E.; PIKTUS, A.; PETRONI, F.; KARPUKHIN, V.; GOYAL, N.; KÜTTLER, H.; LEWIS, M.; YIH, W.-t.; ROCKTÄSCHEL, T. *et al.* Retrieval-augmented generation for knowledge-intensive nlp tasks. *In: ANNUAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NEURIPS)*, 34., 2020, virtual. *Proceedings [...]*. San Diego, CA, USA: Neural Information Processing Systems Foundation, Inc., 2020. p. 9459–9474.

LI, D.; TAN, Z.; QIAN, P.; LI, Y.; CHAUDHARY, K. S.; HU, L.; SHEN, J. Smoa: Improving multi-agent large language models with sparse mixture-of-agents. **arXiv preprint**, 2024. Disponível em: <https://arxiv.org/pdf/2411.03284>. Acesso em: 24 abr. 2025.

LI, G.; GAO, W.; GAO, W. Quantization techniques. *In: Point cloud compression: technologies and standardization*. Singapore: Springer Nature Singapore, 2024. p. 97–112. ISBN 978-981-97-1957-0. Disponível em: <https://link.springer.com/>. Acesso em: 29 abr. 2025.

LI, J.; LI, D.; SAVARESE, S.; HOI, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING*, 40., 2023, Honolulu, Hawaii, USA. *Proceedings [...]*. Honolulu, Hawaii, USA: PMLR, 2023. p. 19730–19742.

LI, X.; CHEN, C.; PAN, T.; ZHOU, X.; SUN, X.; ZHANG, Z.; WU, D.; CHEN, X. Trends and hotspots in non-motor symptoms of parkinson’s disease: a 10-year bibliometric analysis. **Frontiers in Aging Neuroscience**, Lausanne, v. 16, 2024. Disponível em: 10.3389/fnagi.2024.1335550. Acesso em: 15 mai. 2025.

LILLICRAP, T. P.; HUNT, J. J.; PRITZEL, A.; HEESS, N.; EREZ, T.; TASSA, Y.; SILVER, D.; WIERSTRA, D. Continuous control with deep reinforcement learning. **arXiv preprint**, 2015. Disponível em: <https://arxiv.org/pdf/1509.02971>. Acesso em: 13 jan. 2025.

LIMA, A. B. R.; FILHO, J. D. de L.; CAMAPUM, L. F. C.; SANTINI, J. G.; PLETSCHE, B. P. H.; ARAÚJO, B. M. B. de; COSTA, K. G.; SANTHIAGO, G. S.; MANN, N. C. B.; COSTA, C. K. P. *et al.* Panorama das internações por doenças neurológicas degenerativas no brasil: Parkinson, alzheimer e esclerose múltipla (2013-2023). **Brazilian Journal of Implantology and Health Sciences**, Macapá, AP, v. 6, n. 9, p. 1358–1368, 2024.

LIN, C.-Y. ROUGE: a package for automatic evaluation of summaries. *In: TEXT SUMMARIZATION BRANCHES OUT*, 2004, Barcelona, Spain. *Text Summarization Branches Out*. East Stroudsburg, USA: Association for Computational Linguistics, 2004. p. 74–81. Disponível em: <https://aclanthology.org/W04-1013/>. Acesso em: 23 mai. 2025.

LIU, G.; DAI, F.; XU, X.; FU, X.; DOU, W.; KUMAR, N.; BILAL, M. An adaptive dnn inference acceleration framework with end–edge–cloud collaborative computing. **Future Generation Computer Systems**, Amsterdam, v. 140, p. 422–435, 2023.

LIU, P.; LIU, X.; YAO, R.; LIU, J.; MENG, S.; WANG, D.; MA, J. Hm-rag: Hierarchical multi-agent multimodal retrieval augmented generation. **arXiv preprint**, 2025. Disponível em: <https://arxiv.org/pdf/2504.12330>. Acesso em: 23 jun. 2025.

LIU, Y.; LIU, W.; GU, X.; RUI, Y.; HE, X.; ZHANG, Y. Lmagent: A large-scale multimodal agents society for multi-user simulation. **arXiv preprint**, 2024. Disponível em: <https://arxiv.org/pdf/2412.09237>. Acesso em: 25 abr. 2025.

LUCK, M.; D'INVERNO, M. A formal framework for agency and autonomy. *In: INTERNATIONAL CONFERENCE ON MULTI-AGENT SYSTEMS (ICMAS), 1., 1995, San Francisco, California, USA. Proceedings [...].* San Francisco, California, USA: AAAI Press, 1995. p. 254–260.

MACHORRO-CANO, I.; OLMEDO-AGUIRRE, J. O.; ALOR-HERNÁNDEZ, G.; RODRÍGUEZ-MAZAHUA, L.; SÁNCHEZ-MORALES, L. N.; PÉREZ-CASTRO, N. Cloud-based platforms for health monitoring: a review. **Informatics**, Basel, v. 11, n. 1, p. 2, 2023. Disponível em: <https://doi.org/10.3390/informatics11010002>. Acesso em: 19 mar. 2025.

MAI, A. S.; DENG, X.; TAN, E.-K. Epidemiology of early-onset parkinson disease (eopd) worldwide: East versus west. **Parkinsonism & Related Disorders**, Amsterdam, v. 129, p. 107126, 2024. Disponível em: <https://doi.org/10.1016/j.parkreldis.2024.107126>. Acesso em: 17 fev. 2025.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. **arXiv preprint**, 2013. Disponível em: <https://arxiv.org/pdf/1301.3781>. Acesso em: 13 jan. 2025.

OLIEHOEK, F. A.; SPAAN, M. T.; VLASSIS, N. Optimal and approximate q-value functions for decentralized pomdps. **Journal of Artificial Intelligence Research**, v. 32, p. 289–353, 2008.

PAHUNE, S.; AKHTAR, Z. Transitioning from mlops to llmops: Navigating the unique challenges of large language models. **Information**, v. 16, n. 2, p. 87, 2025.

PANDEY, H.; AMOD, A. *et al.* Advancing healthcare automation: Multi-agent system for medical necessity justification. **arXiv preprint**, 2024. Disponível em: <https://arxiv.org/pdf/2404.17977>. Acesso em: 24 abr. 2025.

PAPINENI, K.; ROUKOS, S.; WARD, T.; ZHU, W.-J. Bleu: a method for automatic evaluation of machine translation. *In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 40., 2002, Philadelphia, Pennsylvania, USA. Proceedings [...].* Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002. p. 311–318.

PEIGNÉ, P.; KNIEJSKI, M.; SONDEJ, F.; DAVID, M.; HOELSCHER-OBERMAIER, J.; WITT, C. S. de; KRAN, E. Multi-agent security tax: Trading off security and collaboration capabilities in multi-agent systems. *In: AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE, 39., 2025, Philadelphia, Pennsylvania, USA. Proceedings [...].* Philadelphia, Pennsylvania, USA: AAAI Press, 2025. v. 39, n. 26, p. 27573–27581.

PENG, Z.; WANG, W.; DONG, L.; HAO, Y.; HUANG, S.; MA, S.; WEI, F. Kosmos-2: Grounding multimodal large language models to the world. **arXiv preprint**, 2023. Disponível em: <https://arxiv.org/pdf/2306.14824>. Acesso em: 14 mai. 2025.

PILOTO, G. C.; SILVA, J. C. da; GERMANN, K. C.; CZEPULA, A. Efetividade e segurança do uso da cannabis sativa no tratamento dos sintomas não-motores na doença de parkinson: Uma revisão de escopo. **Research, Society and Development**, Vargem Grande Paulista, SP, v. 13, n. 12, p. e115131247643–e115131247643, 2024.

RABIE, H.; AKHLOUFI, M. A. A review of machine learning and deep learning for parkinson's disease detection. **Discover Artificial Intelligence**, New York, NY, v. 5, n. 1, p. 24, 2025. Disponível em: <https://doi.org/10.1007/s44163-025-00241-9>. Acesso em: 10 mar. 2025.

RADFORD, A.; KIM, J. W.; HALLACY, C.; RAMESH, A.; GOH, G.; AGARWAL, S.; SASTRY, G.; ASKELL, A.; MISHKIN, P.; CLARK, J.; KRUEGER, G.; SUTSKEVER, I. Learning transferable visual models from natural language supervision. *In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING*, 38., 2021, Virtual. *Proceedings [...]*. Virtual: PMLR, 2021. p. 8748–8763.

RAFAILOV, R.; SHARMA, A.; MITCHELL, E.; MANNING, C. D.; ERMON, S.; FINN, C. Direct preference optimization: Your language model is secretly a reward model. *In: ANNUAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NEURIPS)*, 37., 2023, New Orleans, Louisiana, USA. *Proceedings [...]*. San Diego, CA, USA: Neural Information Processing Systems Foundation, Inc., 2023. p. 53728–53741.

RAHUTOMO, F.; KITASUKA, T.; ARITSUGI, M. Semantic cosine similarity. *In: INTERNATIONAL STUDENT CONFERENCE ON ADVANCED SCIENCE AND TECHNOLOGY (ICAST)*, 7., 2012, Seoul, South Korea. *Proceedings [...]*. Seoul: University of Seoul, 2012. v. 4, n. 1, p. 1.

RAJAGOPAL, D.; SUBRAMANIAN, P. K. T. Ai augmented edge and fog computing for internet of health things (ioht). **PeerJ Computer Science**, London, v. 11, p. e2431, 2025. Disponível em: <https://doi.org/10.7717/peerj-cs.2431>. Acesso em: 21 fev. 2025.

RANCEA, A.; ANGHEL, I.; CIOARA, T. Edge computing in healthcare: Innovations, opportunities, and challenges. **Future internet**, Basel, v. 16, n. 9, p. 329, 2024. Disponível em: <https://doi.org/10.3390/fi16090329>. Acesso em: 19 mar. 2025.

RASHID, T.; SAMVELYAN, M.; WITT, C. S. D.; FARQUHAR, G.; FOERSTER, J.; WHITESON, S. Monotonic value function factorisation for deep multi-agent reinforcement learning. **Journal of Machine Learning Research**, Cambridge, MA, v. 21, n. 178, p. 1–51, 2020.

RAZZOUKI, A. F.; JEANCOLAS, L.; MANGONE, G.; SAMBIN, S.; CHALANÇON, A.; GOMES, M.; LEHÉRICY, S.; CORVOL, J.-C.; VIDAILHET, M.; ARNULF, I.; EL-YACOUBI, M. A.; PETROVSKA-DELACRÉTAZ, D. Early-stage parkinson's disease detection based on optical flow and video vision transformer. *In: INTERNATIONAL CONFERENCE ON HUMAN SYSTEM INTERACTION (HSI)*, 16., 2024, Paris, France. *Proceedings [...]*. Paris, France: IEEE, 2024. p. 1–6.

REDDY, A.; REDDY, R. P.; ROGHANI, A. K.; GARCIA, R. I.; KHEMKA, S.; PATTOOR, V.; JACOB, M.; REDDY, P. H.; SEHAR, U. Artificial intelligence in parkinson's disease: Early detection and diagnostic advancements. **Ageing research reviews**, Amsterdam, v. 99, p. 102410, 2024. Disponível em: <https://doi.org/10.1016/j.arr.2024.102410>. Acesso em: 18 mai. 2025.

RUSSELL, S. J.; NORVIG, P. **Artificial intelligence: a modern approach**. 3. ed. [S. l.]: Pearson, 2016.

SARKAR, M.; LEE, T.-H.; SAHOO, P. K. Smart healthcare: exploring the internet of medical things with ambient intelligence. **Electronics**, Basel, v. 13, n. 12, p. 2309, 2024.

SATHUPADI, K.; ACHAR, S.; BHASKARAN, S. V.; FARUQUI, N.; ABDULLAH-AL-WADUD, M.; UDDIN, J. Edge-cloud synergy for ai-enhanced sensor network data: A real-time predictive maintenance framework. **Sensors**, Basel, v. 24, n. 24, p. 7918, 2024. Disponível em: <https://doi.org/10.3390/s24247918>. Acesso em: 10 fev. 2025.

SHAN, R.; SHAN, T. Enterprise llmops: Advancing large language models operations practice. *In: IEEE CLOUD SUMMIT, 2024, Washington, DC, USA. Proceedings [...].* Washington, DC, USA: IEEE, 2024. p. 143–148.

SHAO, Z.; WANG, P.; ZHU, Q.; XU, R.; SONG, J.; BI, X.; ZHANG, H.; ZHANG, M.; LI, Y.; WU, Y. *et al.* Deepseekmath: Pushing the limits of mathematical reasoning in open language models. **arXiv preprint**, 2024. Disponível em: <https://arxiv.org/pdf/2402.03300>. Acesso em: 24 fev. 2025.

SHARMA, S. J.; GUPTA, R. Deep learning models for predicting cognitive impairment in parkinson's disease detection. *In: IEEE WORLD CONFERENCE ON APPLIED INTELLIGENCE AND COMPUTING (AIC), 3., 2024, Gwalior, India. Proceedings [...].* Gwalior, MP: IEEE, 2024. p. 1172–1178.

SHEN, Z.; HE, Y.; WANG, Z.; ZHANG, Y.; SUN, G.; YE, W.; LI, A. Edgelora: An efficient multi-tenant llm serving system on edge devices. **arXiv preprint**, 2025. Disponível em: <https://arxiv.org/pdf/2507.01438>. Acesso em: 23 mai. 2025.

SINGH, A.; EHTESHAM, A.; KUMAR, S.; KHOEI, T. T. Agentic retrieval-augmented generation: A survey on agentic rag. **arXiv preprint**, 2025. Disponível em: <https://arxiv.org/pdf/2501.09136>. Acesso em: 27 abr. 2025.

SKARAMAGKAS, V.; PENTARI, A.; KEFALOPOULOU, Z.; TSIKNAKIS, M. Multi-modal deep learning diagnosis of parkinson's disease-a systematic review. **IEEE Transactions on Neural Systems and Rehabilitation Engineering**, Piscataway, NJ, v. 31, p. 2399–2423, 2023.

SONG, H.; FRANKLIN, S.; NEGATU, A. Sumpy: A fuzzy software agent. *In: ISCA CONFERENCE ON INTELLIGENT SYSTEMS, 7., 1996, Reno, Nevada, USA. Proceedings [...].* Reno, Nevada, USA: International Society for Computers and Their Applications (ISCA), 1996. p. 124–129.

SU, C.; WEN, J.; KANG, J.; WANG, Y.; SU, Y.; PAN, H.; ZHONG, Z.; HOSSAIN, M. S. Hybrid rag-empowered multimodal llm for secure data management in internet of medical things: A diffusion-based contract approach. **IEEE Internet of Things Journal**, Piscataway, NJ, v. 12, n. 10, p. 13428–13440, 2025.

SUN, Y.-m.; WANG, Z.-y.; LIANG, Y.-y.; HAO, C.-w.; SHI, C.-h. Digital biomarkers for precision diagnosis and monitoring in parkinson's disease. **NPJ digital medicine**, London, v. 7, n. 1, p. 218, 2024.

SUNEHAG, P.; LEVER, G.; GRUSLYS, A.; CZARNECKI, W. M.; ZAMBALDI, V.; JADERBERG, M.; LANCTOT, M.; SONNERAT, N.; LEIBO, J. Z.; TUYLS, K. *et al.* Value-decomposition networks for cooperative multi-agent learning. **arXiv preprint**, 2017. Disponível em: <https://arxiv.org/pdf/1706.05296>. Acesso em: 13 jul. 2025.

THE MICHAEL J. FOX FOUNDATION FOR PARKINSON'S RESEARCH. **The Michael J. Fox Foundation**. 2025. Disponível em: <https://www.michaeljfox.org/>. Acesso em: 15 abr. 2025.

TRAN, K.-T.; DAO, D.; NGUYEN, M.-D.; PHAM, Q.-V.; O'SULLIVAN, B.; NGUYEN, H. D. Multi-agent collaboration mechanisms: A survey of llms. **arXiv preprint**, 2025. Disponível em: <https://arxiv.org/pdf/2501.06322>. Acesso em: 27 abr. 2025.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. *In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NIPS)*, 31., 2017, Long Beach, CA, USA. *Advances in Neural Information Processing Systems 30*. Long Beach, CA, USA: Neural Information Processing Systems Foundation, Inc., 2017. p. 5998–6008.

WANG, S.; LIU, J.; XU, H.; YAN, J.; GAO, X. Efficient federated fine-tuning of large language models with layer dropout. **arXiv preprint**, 2025. Disponível em: <https://arxiv.org/pdf/2503.10217>. Acesso em: 28 abr. 2025.

WANG, S.; SHUAI, H.; LIU, Q.; WANG, F. Cooperative sentiment agents for multimodal sentiment analysis. **arXiv preprint**, 2024. Disponível em: <https://arxiv.org/pdf/2404.12642>. Acesso em: 25 abr. 2025.

WANG, S.; ZHANG, G.; YU, M.; WAN, G.; MENG, F.; GUO, C.; WANG, K.; WANG, Y. G-safeguard: A topology-guided security lens and treatment on llm-based multi-agent systems. **arXiv preprint**, 2025. Disponível em: <https://arxiv.org/pdf/2502.11127>. Acesso em: 25 fev. 2025.

WANG, X.; HUANG, J.; CHATZAKOU, M.; MEDIJAINEN, K.; TOOMELA, A.; NÖMM, S.; RUZHANSKY, M. Lstm-cnn: An efficient diagnostic network for parkinson’s disease utilizing dynamic handwriting analysis. **Computer Methods and Programs in Biomedicine**, Amsterdam, v. 247, p. 108066, 2024.

WANG, Y.; CHEN, Y.; LI, Z.; TANG, Z.; GUO, R.; WANG, X.; WANG, Q.; ZHOU, A. C.; CHU, X. Towards efficient and reliable llm serving: A real-world workload study. **arXiv preprint**, 2024. Disponível em: <https://arxiv.org/pdf/2401.17644>. Acesso em: 24 jan. 2025.

WU, H.-H.; SEETHARAMAN, P.; KUMAR, K.; BELLO, J. P. Wav2clip: Learning robust audio representations from clip. *In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP)*, 47., 2022, Singapore. *Proceedings [...]*. Singapore: IEEE, 2022. p. 4563–4567.

WU, J.; ZHU, J.; QI, Y.; CHEN, J.; XU, M.; MENOLASCINA, F.; GRAU, V. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. **arXiv preprint**, 2024. Disponível em: <https://arxiv.org/pdf/2408.04187>. Acesso em: 20 ago. 2025.

XAVIER, D.; FELIZARDO, V.; FERREIRA, B.; ZACARIAS, H.; POURVAHAB, M.; SOUZA-PEREIRA, L.; GARCIA, N. M. Voice analysis in parkinson’s disease-a systematic literature review. **Artificial Intelligence in Medicine**, Amsterdam, p. 103109, 2025. Disponível em: <https://doi.org/10.1016/j.artmed.2025.103109>. Acesso em: 20 mar. 2025.

XIA, P.; WANG, J.; PENG, Y.; ZENG, K.; WU, X.; TANG, X.; ZHU, H.; LI, Y.; LIU, S.; LU, Y. *et al.* Mmedagent-rl: Optimizing multi-agent collaboration for multimodal medical reasoning. **arXiv preprint**, 2025. Disponível em: <https://arxiv.org/pdf/2506.00555>. Acesso em: 23 jun. 2025.

YANG, R.; NING, Y.; KEPPO, E.; LIU, M.; HONG, C.; BITTERMAN, D. S.; ONG, J. C. L.; TING, D. S. W.; LIU, N. Retrieval-augmented generation for generative artificial intelligence in health care. **npj Health Systems**, London, v. 2, n. 1, p. 2, 2025.

YASUNAGA, M.; AGHAJANYAN, A.; SHI, W.; JAMES, R.; LESKOVEC, J.; LIANG, P.; LEWIS, M.; ZETTLEMOYER, L.; YIH, W.-t. Retrieval-augmented multimodal language modeling. **arXiv preprint**, 2022. Disponível em: <https://arxiv.org/pdf/2211.12561>. Acesso em: 19 mai. 2025.

ZHANG, D.; LI, Z.; WANG, P.; ZHANG, X.; ZHOU, Y.; QIU, X. Speechagents: Human-communication simulation with multi-modal multi-agent systems. **arXiv preprint**, 2024. Disponível em: <https://arxiv.org/pdf/2401.03945>. Acesso em: 22 abr. 2025.

ZHANG, H.; RUAN, G.; HARTICH, R.; PATEL, N.; ZHONG, M.; KACZOREK, A.; MILLER, L.; GIESTING, R.; LI, W.; ZHANG, B.; XIAO, H.; PORTER, R.; WINGER, B. E.; YANG, J. Magnol. ai—an internet of medical things (iomt) platform for digital health research. *In: IEEE INTERNATIONAL CONFERENCE ON BIG DATA (BIGDATA)*, 2024, Washington, DC, USA. *Proceedings [...]*. Washington, DC, USA: IEEE, 2024. p. 5103–5112.

ZHANG, T.; KISHORE, V.; WU, F.; WEINBERGER, K. Q.; ARTZI, Y. Bertscore: Evaluating text generation with bert. **arXiv preprint**, 2019. Disponível em: <https://arxiv.org/pdf/1904.09675>. Acesso em: 12 mai. 2025.

ZHANG, W.; XIAO, D.; MAO, Q.; XIA, H. Role of neuroinflammation in neurodegeneration development. **Signal transduction and targeted therapy**, London, v. 8, n. 1, p. 267, 2023.

ZHANG, Y.; LI, J.; WANG, T. Edge–cloud collaborative inference for multimodal medical image analysis. **IEEE Journal of Biomedical and Health Informatics**, Piscataway, NJ, v. 28, p. 1024–1035, 2024.

ZHAO, A.; LIU, Y.; YU, X.; XING, X.; ZHOU, H. Artificial intelligence-enabled detection and assessment of parkinson’s disease using multimodal data: A survey. **Information Fusion**, Amsterdam, p. 103175, 2025.

ZHAO, R.; CHEN, H.; WANG, W.; JIAO, F.; DO, X. L.; QIN, C.; DING, B.; GUO, X.; LI, M.; LI, X. *et al.* Retrieving multimodal information for augmented generation: A survey. **arXiv preprint**, 2023. Disponível em: <https://arxiv.org/pdf/2303.10868>. Acesso em: 20 mai. 2025.

ZHENG, L.; CHIANG, W.-L.; SHENG, Y.; ZHUANG, S.; WU, Z.; ZHUANG, Y.; LIN, Z.; LI, Z.; LI, D.; XING, E. *et al.* Judging llm-as-a-judge with mt-bench and chatbot arena. *In: ANNUAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NEURIPS)*, 37., 2023, New Orleans, Louisiana, USA. *Proceedings [...]*. San Diego, CA, USA: Neural Information Processing Systems Foundation, Inc., 2023. p. 46595–46623.

ZHOU, C.; HAN, S.; ZHANG, S.; ZHOU, Y.; ZHANG, W.; JIN, C. Efficient fine-tuning of quantized models via adaptive rank and bitwidth. **arXiv preprint**, 2025. Disponível em: <https://arxiv.org/pdf/2505.03802>. Acesso em: 21 abr. 2025.

ZHU, Y.; REN, C.; WANG, Z.; ZHENG, X.; XIE, S.; FENG, J.; ZHU, X.; LI, Z.; MA, L.; PAN, C. Emerge: Enhancing multimodal electronic health records predictive modeling with retrieval-augmented generation. *In: ACM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT (CIKM)*, 33., 2024, Boise, ID, USA. *Proceedings [...]*. Boise, ID, USA: Association for Computing Machinery, 2024. p. 3549–3559.