Aprendizado Multimodal Tabular-Texto em Registros Clínicos Baseado em Modelos Fundacionais

Rafaela de Fátima Silva Alexandre, Pedro Henrique de Souza Barros, Heitor Soares Ramos Filho
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil

***Resumo*—Registros eletrônicos de saúde combinam dados estruturados, como exames, sinais vitais e histórico do paciente, com informações textuais, como anotações e relatórios clínicos. Entretanto, essas modalidades são frequentemente analisadas separadamente, limitando a representação integrada do contexto clínico. Este trabalho propõe investigar o alinhamento multimodal entre dados tabulares e textuais por meio de modelos fundacionais. A abordagem prevê a extração independente de embeddings, sua projeção em um espaço latente compartilhado e o uso de aprendizado contrastivo para aproximar representações correspondentes. Espera-se obter representações mais completas, robustas e transferíveis para tarefas clínicas posteriores.**

***Index Terms*—aprendizado multimodal, registros eletrônicos de saúde, dados tabulares, texto clínico, modelos fundacionais, aprendizado contrastivo**

I. Introdução

O aprendizado profundo tem produzido avanços em visão computacional, processamento de linguagem natural e sistemas multimodais. Modelos como o *Contrastive Language-Image Pre-Training* (*CLIP*) demonstraram que modalidades distintas podem ser alinhadas em um espaço representacional compartilhado, permitindo generalização e transferência entre tarefas [1].

Esse paradigma relaciona-se aos modelos fundacionais, treinados em larga escala sobre dados heterogêneos e posteriormente adaptados a diferentes aplicações [2]. Na saúde, sua utilização é particularmente relevante porque os registros eletrônicos de saúde (*Electronic Health Records* – EHRs) reúnem informações estruturadas e não estruturadas [3], [4].

Os dados estruturados registram elementos como exames, sinais vitais e histórico do paciente, enquanto os textos clínicos acrescentam contexto semântico, raciocínio médico e nuances diagnósticas. A análise isolada dessas fontes pode limitar a identificação das relações existentes entre diferentes aspectos do estado clínico [5].

Este trabalho propõe investigar o alinhamento multimodal tabular-texto em registros clínicos. O objetivo é estudar como modelos fundacionais podem ser utilizados para construir representações compartilhadas, mais completas e generalizáveis dos pacientes. O restante do artigo apresenta o referencial teórico, os materiais e métodos, os resultados obtidos e as conclusões do trabalho.

II. Referencial Teórico e Trabalhos Correlatos

A. Modelos Fundacionais e Aprendizado Multimodal

Modelos fundacionais são treinados, geralmente de forma autossupervisionada, em grandes volumes de dados e podem ser adaptados a diversas tarefas posteriores. Entre suas principais características estão o surgimento de novas capacidades com o aumento da escala e a reutilização de uma mesma arquitetura em diferentes aplicações [2].

No aprendizado multimodal, representações provenientes de diferentes modalidades são integradas em um espaço comum. O *CLIP* utiliza aprendizado contrastivo para aproximar imagens e textos correspondentes e afastar pares não relacionados [1]. Embora originalmente proposto para visão e linguagem, esse princípio também pode ser empregado no alinhamento entre dados tabulares e textos clínicos.

Bonnier [6] investiga modelos multimodais aplicados a tabelas que contêm campos textuais. O trabalho compara estratégias de representação dos atributos numéricos e diferentes formas de fusão entre informações tabulares e textuais, demonstrando que as escolhas de codificação e integração influenciam o desempenho em tarefas de classificação. Diferentemente dessa abordagem, que prioriza a predição supervisionada, o presente trabalho concentra-se no alinhamento contrastivo das modalidades em um espaço latente compartilhado.

Mais recentemente, Arazi et al. [7] apresentam o MulTa-Bench, um *benchmark* composto por tarefas que combinam dados tabulares com textos ou imagens. O estudo destaca que *embeddings* genéricos e congelados podem não preservar informações específicas necessárias para cada tarefa, enquanto representações adaptadas ao objetivo apresentam melhor desempenho. Esse resultado reforça a importância de avaliar tanto a qualidade do alinhamento multimodal quanto a adequação dos *encoders* ao domínio clínico.

Apesar dos avanços observados, o alinhamento direto entre *embeddings* produzidos por modelos fundacionais tabulares e modelos de linguagem clínica ainda é menos explorado do que a integração entre visão e linguagem. É nessa direção que se posiciona o presente trabalho, buscando preservar relações semânticas entre pacientes em um espaço multimodal compartilhado.

B. Representação de Registros Clínicos

Os registros eletrônicos de saúde combinam informações estruturadas, como diagnósticos, sintomas, tratamentos e condições prévias, com textos livres que descrevem o estado clínico dos pacientes. Trabalhos anteriores demonstram que representações aprendidas a partir desses registros podem capturar padrões latentes úteis para tarefas como previsão de doenças, readmissão hospitalar e evolução clínica [8], [4]

Entretanto, abordagens tradicionais frequentemente processam cada modalidade separadamente, dificultando a modelagem das relações existentes entre as informações estruturadas e textuais [5]. A integração multimodal busca superar essa limitação ao representar o paciente a partir de diferentes fontes, combinando os aspectos quantitativos dos dados tabulares com o contexto semântico presente nos textos clínicos.

Rabaey et al. [9] apresentam o *SimSUM*, um conjunto de dados composto por registros clínicos simulados que associa informações estruturadas a textos clínicos. Cada registro representa um atendimento relacionado a doenças respiratórias e contém variáveis referentes a sintomas, diagnósticos e condições prévias. A correspondência entre as modalidades torna o *SimSUM* adequado para investigar o aprendizado multimodal tabular-texto. No presente trabalho, esse conjunto de dados é utilizado como base para a construção e avaliação das representações multimodais dos pacientes.

C. Modelos para Dados Tabulares e Textuais

Algoritmos baseados em árvores, como o *XGBoost*, apresentam desempenho consolidado em dados tabulares [10]. Mais recentemente, arquiteturas como *TabTransformer* e *FT-Transformer* passaram a empregar mecanismos de atenção para modelar interações entre atributos e produzir representações mais expressivas das amostras [11], [12].

No contexto dos modelos fundacionais tabulares, Qu et al. [13] propõem o TabICL, uma arquitetura que utiliza aprendizado em contexto para realizar classificação sem atualizações de parâmetros específicas para cada conjunto de dados. O modelo processa atributos e amostras tabulares para construir representações das linhas, permitindo lidar com conjuntos de dados de diferentes características e dimensões. Neste trabalho, o TabICL é empregado como extrator das representações tabulares dos pacientes.

Para a modalidade textual, modelos baseados no BERT foram adaptados ao vocabulário e às características dos documentos médicos. Alsentzer et al. [14] apresentam versões do BERT treinadas em textos clínicos gerais e em sumários de alta hospitalar. Os resultados indicam que o pré-treinamento específico do domínio pode melhorar o desempenho em tarefas de processamento de linguagem clínica. Com base nessa contribuição, o ClinicalBERT é utilizado neste trabalho para produzir *embeddings* dos textos associados aos pacientes.

Assim, o TabICL e o ClinicalBERT fornecem representações especializadas para as modalidades tabular e textual, respectivamente. O presente trabalho investiga o alinhamento dessas representações por meio do aprendizado contrastivo, buscando construir uma representação multimodal compartilhada que integre as diferentes informações clínicas de cada paciente.

D. Aprendizado Contrastivo e InfoNCE

O aprendizado contrastivo busca aprender representações em que exemplos relacionados fiquem próximos no espaço latente, enquanto exemplos não relacionados sejam afastados. No contexto multimodal, essa estratégia permite alinhar representações de diferentes modalidades, como texto e imagem no CLIP, ou dados tabulares e texto clínico neste trabalho.

Uma das funções de perda mais utilizadas nesse cenário é a InfoNCE. Dado um exemplo de consulta, a InfoNCE aumenta a similaridade com seu par positivo e reduz a similaridade relativa com os demais exemplos do lote, tratados como negativos. Para uma consulta tabular z_i^{tab} e textos candidatos z_j^{txt}, a perda pode ser escrita como:

$$\mathcal{L}_{tab \to txt} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\mathrm{sim}(z_i^{tab}, z_i^{txt})/\tau)}{\sum_{j=1}^{N} \exp(\mathrm{sim}(z_i^{tab}, z_j^{txt})/\tau)}. \quad (1)$$

Nessa formulação, $\mathrm{sim}(\cdot)$ representa a similaridade entre embeddings, τ é o parâmetro de temperatura e N é o número de exemplos no lote. O par correspondente ao mesmo paciente é considerado positivo, enquanto os demais exemplos do lote são considerados negativos.

Neste trabalho, a InfoNCE é utilizada inicialmente para alinhar embeddings tabulares do TabICL e embeddings textuais do ClinicalBERT. Posteriormente, sua limitação no tratamento uniforme dos negativos motiva a extensão ontológica proposta.

E. Soft-InfoNCE e Extensão Hierárquica

Apesar da utilidade da InfoNCE para alinhamento multimodal, sua formulação tradicional assume que todos os exemplos não correspondentes no lote são negativos igualmente válidos. Essa suposição pode ser inadequada quando há exemplos parcialmente relacionados. Li et al. [15] discutem essa limitação no contexto de busca de código e propõem a Soft-InfoNCE, que introduz pesos para diferenciar pares negativos de acordo com seu grau de relevância.

De forma geral, a Soft-InfoNCE substitui o alvo rígido da InfoNCE por uma distribuição ponderada sobre os exemplos do lote. Assim, em vez de considerar apenas o par exato como positivo e todos os demais como negativos fortes, a função permite atribuir pesos intermediários a exemplos semanticamente relacionados.

Neste trabalho, essa ideia é adaptada ao contexto clínico por meio da hierarquia ICD-10. A adaptação proposta recebe o nome de Soft Hierarchical InfoNCE, pois os pesos entre pacientes são definidos a partir da proximidade ontológica entre suas doenças. Dessa forma, pacientes distintos, mas com condições clinicamente próximas, deixam de ser tratados como negativos absolutos.

A formulação geral da Soft Hierarchical InfoNCE na direção Tabular $\rightarrow$ Texto é dada por:

$$\mathcal{L}^{soft}_{tab\to txt} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\tilde{W}_{ij}\log\frac{\exp(\mathrm{sim}(z_i^{tab}, z_j^{txt})/\tau)}{\sum_{k=1}^{N}\exp(\mathrm{sim}(z_i^{tab}, z_k^{txt})/\tau)}. \quad (2)$$

Nessa equação, $\tilde{W}_{ij}$ representa o peso normalizado entre os pacientes i e j, calculado a partir da proximidade ontológica entre suas doenças, e τ é o parâmetro de temperatura. Quando $\tilde{W}_{ij}$ concentra todo o peso no par correspondente, a formulação se aproxima da InfoNCE tradicional. Quando pesos intermediários são atribuídos a pacientes clinicamente relacionados, a perda passa a favorecer uma organização mais gradual do espaço latente.

A perda final utilizada neste trabalho é bidirecional, considerando tanto a direção Tabular $\rightarrow$ Texto quanto a direção Texto $\rightarrow$ Tabular:

$$\mathcal{L}^{soft} = \frac{1}{2}\left(\mathcal{L}^{soft}_{tab\to txt} + \mathcal{L}^{soft}_{txt\to tab}\right). \quad (3)$$

Diante desse contexto, este trabalho propõe um *pipeline* multimodal tabular-texto para o conjunto SimSUM, no qual *embeddings* tabulares extraídos com o TabICL e *embeddings* textuais extraídos com o ClinicalBERT são projetados para um espaço latente compartilhado. Além do alinhamento inicial com InfoNCE, investiga-se uma extensão baseada em pesos ontológicos derivados da hierarquia ICD-10, denominada neste trabalho de Soft Hierarchical InfoNCE, com o objetivo de preservar relações clínicas entre doenças semanticamente próximas. Assim, a contribuição central do trabalho consiste em avaliar o alinhamento entre modalidades clínicas e em analisar se a incorporação de conhecimento ontológico torna o espaço latente mais coerente com relações médicas.

III. Materiais e Métodos

A metodologia foi organizada em duas etapas experimentais. Na primeira, foram extraídos *embeddings* tabulares com o TabICL e *embeddings* textuais com o ClinicalBERT, posteriormente alinhados em um espaço latente compartilhado por meio da InfoNCE. Na segunda, foi incorporado conhecimento ontológico da ICD-10 ao alinhamento, substituindo a separação rígida entre positivos e negativos por pesos contínuos baseados na proximidade clínica entre doenças.

A. Conjunto de Dados

O trabalho considera registros eletrônicos de saúde nos quais cada paciente é representado por duas modalidades complementares. A modalidade tabular reúne informações estruturadas, como exames, sinais vitais e histórico do paciente. A modalidade textual contém informações não estruturadas, como anotações e relatórios clínicos.

O conjunto de dados utilizado foi o *SimSUM*, um conjunto clínico multimodal composto por 10.000 registros simulados de pacientes com condições respiratórias. Cada registro contém uma modalidade tabular, formada por variáveis clínicas estruturadas, e uma modalidade textual, composta por descrições clínicas associadas ao mesmo caso.

A base possui 18 colunas, incluindo variáveis demográficas/contextuais, condições clínicas, sintomas, tratamento e texto clínico. Entre as variáveis estruturadas estão doenças respiratórias, como asma, Doença Pulmonar Obstrutiva Crônica (DPOC), pneumonia, resfriado comum e rinite alérgica, além de sintomas como dispneia, tosse, dor, febre e sintomas nasais. Também há atributos como estação do ano, uso de antibiótico e número de dias de afastamento domiciliar.

No conjunto utilizado, cada linha foi tratada como um registro de paciente. Como a base não possui identificador longitudinal de paciente, não foi realizada agregação temporal ou acompanhamento de múltiplas visitas. Após a separação entre exemplos de contexto e consulta utilizada pelo TabICL, foram mantidos 256 registros como contexto e 9.744 registros para geração dos embeddings e avaliação experimental.

O conjunto apresenta desbalanceamento entre as condições clínicas. O resfriado comum é a condição mais frequente, com 2.335 registros positivos, enquanto pneumonia é a mais rara, com 103 registros positivos. As demais condições consideradas incluem asma, com 949 registros positivos, DPOC, com 201, e rinite alérgica, com 145. Esse desbalanceamento foi considerado na interpretação dos resultados, especialmente na tarefa *downstream* de classificação de pneumonia.

Exemplos representativos dos registros são apresentados na Tabela III, no Apêndice.

B. Preparação dos Dados e Geração dos Embeddings

O conjunto de dados *SimSUM* foi carregado a partir de um arquivo *CSV* com separador ponto e vírgula. Cada registro contém atributos clínicos estruturados e duas possíveis representações textuais, denominadas `text` e `advanced_text`. A variável `pneu`, indicativa de pneumonia, foi adotada como alvo da tarefa *downstream* tabular e convertida de `yes`/`no` para valores binários 1/0. Registros com valores inválidos nessa variável foram removidos.

Em seguida, as modalidades foram separadas. As colunas `text` e `advanced_text` foram reservadas para o processamento textual, enquanto as demais variáveis compuseram a modalidade tabular. Os atributos binários, como presença de sintomas e condições clínicas, foram convertidos de `yes`/`no` para 1/0. A variável ordinal `fever` foi representada pelos valores 0, 1 e 2, correspondentes a ausência, febre baixa e febre alta, respectivamente. As variáveis categóricas restantes foram transformadas por codificação *one-hot*, e todos os atributos tabulares foram convertidos para ponto flutuante.

1) Geração dos Embeddings Tabulares: Para a extração dos *embeddings* tabulares, foi utilizado o TabICL com o *checkpoint* `tabicl-classifier-v2-20260212.ckpt`. Os registros foram divididos em um conjunto de contexto, composto pelos primeiros 256 pacientes, e um conjunto de consulta, formado pelos 9.744 registros restantes. Os rótulos de pneumonia foram fornecidos apenas para os pacientes de contexto, enquanto os pacientes de consulta foram apresentados sem seus respectivos rótulos.

A geração das representações ocorreu em três etapas internas. Inicialmente, o módulo de representação por colunas

produziu *embeddings* sensíveis à distribuição dos atributos. Em seguida, o módulo de interação por linhas combinou as informações das diferentes colunas para construir uma representação individual de cada paciente. Por fim, o módulo de aprendizado em contexto incorporou os rótulos dos pacientes de contexto e utilizou esses exemplos para contextualizar as representações dos pacientes de consulta.

Os *embeddings* foram extraídos após a etapa de aprendizado em contexto e antes do decodificador responsável pela classificação. Dessa forma, foram obtidas representações tabulares contextualizadas pelo conjunto de exemplos, em vez das probabilidades finais produzidas pelo classificador. O processamento foi realizado em modo de avaliação e sem cálculo de gradientes.

2) Geração dos Embeddings Textuais: Os textos clínicos correspondentes aos pacientes de consulta foram processados com o modelo pré-treinado Bio_ClinicalBERT. A coluna `advanced_text` foi utilizada como fonte textual principal e, quando seu conteúdo estava vazio, empregou-se o texto disponível na coluna `text`.

A *tokenização* foi realizada com preenchimento e truncamento para um comprimento máximo de 256 *tokens*. Os textos foram processados em lotes de 32 amostras. O ClinicalBERT foi mantido congelado e executado em modo de avaliação, sem atualização de seus parâmetros. Como representação de cada registro, foi selecionado o estado oculto associado ao *token* `[CLS]` da última camada, resultando em *embeddings* textuais de 768 dimensões.

3) Pareamento e Armazenamento: O mesmo conjunto de índices de consulta foi utilizado nas modalidades tabular e textual. Assim, cada *embedding* produzido pelo TabICL permaneceu diretamente associado ao *embedding* textual e ao rótulo do mesmo paciente.

Ao final do processamento, os *embeddings* tabulares pós-ICL, os *embeddings* textuais, os rótulos de pneumonia, os índices dos pacientes de consulta e a identificação da coluna textual utilizada foram armazenados conjuntamente no arquivo `clinical_multimodal_embeddings.pt`. Esse arquivo constituiu a entrada para as etapas posteriores de projeção e alinhamento multimodal.

C. Protocolo Experimental

Os experimentos foram organizados em duas etapas. A primeira etapa avaliou o alinhamento multimodal entre os *embeddings* tabulares do TabICL e os *embeddings* textuais do *ClinicalBERT* utilizando a *InfoNCE* tradicional. A segunda etapa investigou uma extensão ontológica da função de perda, incorporando pesos derivados da hierarquia ICD-10 por meio da *Soft Hierarchical InfoNCE*.

Na primeira etapa, foram conduzidos cinco experimentos principais. Inicialmente, avaliou-se a utilidade das representações em uma tarefa *downstream* de classificação de pneumonia, comparando *ClinicalBERT* congelado, *ClinicalBERT* com ajuste fino, *TabICLClassifier* direto e representações alinhadas. Em seguida, analisou-se o espaço latente por meio de *PCA*, verificando a organização das doenças no espaço multimodal. Também foi calculada a similaridade doença–sintoma no espaço latente, com o objetivo de avaliar se as representações preservavam relações clinicamente plausíveis. Por fim, foram avaliadas métricas de *retrieval* semântico, como *Recall@K* e *Precision@K*, nas direções Tabular $\rightarrow$ Texto e Texto $\rightarrow$ Tabular.

Na segunda etapa, avaliou-se o efeito da incorporação de conhecimento ontológico no alinhamento multimodal. Para isso, foram construídos pesos entre doenças a partir da hierarquia *ICD-10* e treinado um alinhador com *Soft Hierarchical InfoNCE*. O desempenho foi medido com *Ontology Weighted Recall@K*, que atribui crédito parcial para recuperações clinicamente próximas, além das correspondências exatas. Essa etapa foi analisada inicialmente com *ClinicalBERT* congelado.

Assim, os experimentos buscaram avaliar não apenas a proximidade entre modalidades, mas também a organização clínica do espaço latente, a recuperação de pacientes semanticamente relacionados e o impacto da inclusão de conhecimento ontológico no processo de alinhamento.

D. Primeira Etapa: Alinhamento Multimodal com InfoNCE

A primeira etapa teve como objetivo verificar se os *embeddings* tabulares e textuais poderiam ser alinhados em um espaço latente compartilhado. Foram utilizados os *embeddings* tabulares produzidos pelo TabICL e os *embeddings* textuais extraídos pelo *ClinicalBERT*. Os registros foram mantidos na mesma ordem, garantindo a correspondência entre a representação tabular e o texto clínico de cada paciente.

1) Projeção para o Espaço Compartilhado: Como os *embeddings* tabulares e textuais possuíam dimensionalidades diferentes, foram utilizadas duas cabeças de projeção independentes. Cada cabeça foi composta por uma transformação linear, ativação *GELU*, uma segunda camada linear, *dropout*, conexão residual e normalização em camadas.

As saídas foram normalizadas pela norma L_2, permitindo que a similaridade entre pacientes fosse calculada por meio do produto interno, equivalente à similaridade cosseno para vetores normalizados. Dessa forma, as duas modalidades foram projetadas para um espaço latente comum.

2) Alinhamento com InfoNCE: O alinhamento inicial foi realizado com a função InfoNCE apresentada na Equação 1. Para cada lote, o *embedding* tabular de um paciente e o *embedding* textual correspondente foram tratados como um par positivo. Os textos pertencentes aos demais pacientes do lote foram considerados exemplos negativos.

Na implementação, a similaridade entre as modalidades foi calculada pelo produto interno entre os vetores projetados e normalizados, equivalente à similaridade cosseno. A perda foi aplicada de forma simétrica, considerando as direções Tabular $\rightarrow$ Texto e Texto $\rightarrow$ Tabular. Assim, o modelo foi treinado para recuperar o texto correspondente a partir da tabela e, simultaneamente, recuperar a tabela correspondente a partir do texto.

3) Configuração do Treinamento: O alinhador foi treinado durante 200 épocas, utilizando lotes de 64 pacientes. Foi empregado o otimizador *AdamW*, com taxa de aprendizado

de 10^{-4} e regularização por decaimento de peso igual a 0,01. A taxa de aprendizado foi ajustada ao longo das épocas por um escalonador cosseno.

O parâmetro de temperatura da perda contrastiva foi aprendido durante o treinamento e limitado para evitar valores extremos. Após o treinamento, os pesos do alinhador e os *embeddings* projetados de todos os pacientes foram armazenados para as análises posteriores.

4) Representação Multimodal: Além das representações tabular e textual alinhadas, foi construída uma representação multimodal por meio da média das duas modalidades:

$$\mathbf{z}_i^{mm} = \frac{\mathbf{z}_i^{tab} + \mathbf{z}_i^{text}}{2}. \tag{4}$$

O vetor resultante foi novamente normalizado pela norma L_2. Essa representação foi utilizada para analisar conjuntamente as informações estruturadas e textuais de cada paciente.

5) Avaliação do Espaço Latente: A organização do espaço latente foi analisada por meio da Análise de Componentes Principais (PCA). Os *embeddings* foram padronizados e reduzidos para duas dimensões, permitindo visualizar a distribuição das modalidades e das condições clínicas.

As visualizações consideraram cinco doenças respiratórias: pneumonia, asma, doença pulmonar obstrutiva crônica, resfriado comum e rinite alérgica. Também foram analisados sintomas como dispneia, tosse, dor, febre e sintomas nasais.

A qualidade do alinhamento foi adicionalmente examinada pela comparação entre a similaridade cosseno dos pares tabular–texto correspondentes e a similaridade de pares formados aleatoriamente.

6) Avaliação da Similaridade Clínica: Para investigar a estrutura semântica aprendida, foram calculados centroides para cada doença e sintoma. O centroide de uma condição foi obtido pela média dos *embeddings* multimodais dos pacientes que apresentavam o respectivo rótulo.

A similaridade cosseno entre os centroides foi utilizada para analisar a proximidade entre doenças e as relações entre doenças e sintomas. Também foi calculada a consistência dos vizinhos mais próximos, definida como a proporção dos K vizinhos que compartilhavam um determinado rótulo clínico com o paciente consultado.

7) Avaliação por Retrieval Semântico: A capacidade de recuperação entre modalidades foi avaliada nas direções Tabular $\rightarrow$ Texto e Texto $\rightarrow$ Tabular. Dois pacientes foram considerados semanticamente relacionados quando compartilhavam pelo menos uma doença ou sintoma clínico.

Foram calculadas as métricas *Semantic Recall@K* e *Semantic Precision@K* para diferentes valores de K. O *Semantic Recall@K* indicou se pelo menos um paciente clinicamente relacionado apareceu entre os primeiros resultados, enquanto o *Semantic Precision@K* mediu a proporção de resultados recuperados que compartilhavam informações clínicas com a consulta.

8) Comparação com Representações Unimodais: Por fim, o *retrieval* obtido após o alinhamento foi comparado com duas referências unimodais: recuperação Tabular $\rightarrow$ Tabular utilizando os *embeddings* originais do TabICL e recuperação Texto $\rightarrow$ Texto utilizando os *embeddings* originais do ClinicalBERT.

Essa comparação permitiu avaliar se o alinhamento multimodal preservou ou melhorou a organização clínica encontrada nas representações produzidas separadamente por cada modelo.

E. Motivação para a Extensão Ontológica

Na primeira etapa, o alinhamento entre *embeddings* tabulares e textuais foi realizado com a função *InfoNCE* tradicional. Essa função aproxima o par correspondente, isto é, a representação tabular e textual do mesmo paciente, e afasta os demais exemplos presentes no lote de treinamento.

Embora essa estratégia seja adequada para alinhar pares multimodais, ela assume que todos os exemplos não correspondentes são negativos igualmente válidos. Essa limitação também é discutida por Li et al. [15], que argumentam que a *InfoNCE* tradicional pode ser prejudicada pela presença de falsos negativos e pela incapacidade de diferenciar graus de relevância entre pares negativos.

No contexto clínico, essa suposição é particularmente restritiva, pois pacientes distintos podem apresentar condições clinicamente relacionadas ou sintomas semelhantes. Assim, exemplos associados a doenças próximas, como asma e DPOC, podem ser tratados como negativos fortes, mesmo compartilhando características clínicas relevantes.

Essa limitação motivou a segunda etapa do trabalho, na qual foi incorporada informação ontológica da Classificação Internacional de Doenças (CID-10) (em inglês - *International Statistical Classification of Diseases and Related Health Problems* (*ICD-10*)) à função de perda. Inspirada pela ideia de ponderar pares negativos na Soft-InfoNCE [15], a separação rígida entre pares positivos e negativos foi substituída por uma ponderação contínua baseada na proximidade clínica entre doenças.

F. Segunda Etapa: Alinhamento com Ontologia Clínica

Na segunda etapa, foi incorporado conhecimento clínico ao alinhamento entre as modalidades por meio da hierarquia da *ICD-10*. Foram utilizados os *embeddings* tabulares pós-ICL (em inglês, *In-Context Learning*) do *TabICL*, com 512 dimensões, e os *embeddings* textuais previamente extraídos pelo *ClinicalBERT*, com 768 dimensões.

O *ClinicalBERT* permaneceu congelado durante essa etapa. Portanto, o treinamento foi realizado somente sobre as cabeças de projeção responsáveis por mapear as representações tabulares e textuais para o espaço latente compartilhado.

1) Construção da Hierarquia ICD-10: As cinco doenças respiratórias presentes no *SimSUM* foram associadas aos respectivos códigos da ICD-10: resfriado comum (`J00`), pneumonia (`J18.9`), rinite alérgica (`J30`), DPOC (`J44`) e asma (`J45`).

Cada código foi representado por um caminho hierárquico, iniciado na raiz da classificação e composto pelos blocos intermediários até a doença. A distância entre duas condições

foi definida pelo número de arestas entre seus códigos por meio do ancestral comum mais baixo, denominado *Lowest Common Ancestor* (LCA):

$$d_{\mathrm{ICD}}(a,b) = \mathrm{depth}(a) + \mathrm{depth}(b) - 2\,\mathrm{depth}(\mathrm{LCA}(a,b)). \tag{5}$$

Essa formulação atribui distâncias menores a doenças localizadas em regiões próximas da hierarquia.

2) Definição dos Pesos Ontológicos: As distâncias entre doenças foram transformadas em pesos contínuos por meio de uma função de decaimento exponencial:

$$w(a,b) = \exp\left(-\lambda_{\mathrm{ICD}} d_{\mathrm{ICD}}(a,b)\right), \tag{6}$$

em que λ_{ICD} controla a velocidade com que a similaridade diminui conforme aumenta a distância hierárquica.

Doenças iguais recebem peso máximo, enquanto doenças mais distantes recebem pesos progressivamente menores. Dessa forma, a proximidade clínica deixa de ser representada por uma decisão binária e passa a ser expressa de maneira contínua.

3) Matriz Ontológica entre Pacientes: A matriz de pesos entre doenças foi expandida para uma matriz paciente–paciente, representada por $W \in \mathbb{R}^{N\times N}$. Cada elemento W_{ij} expressa a proximidade ontológica entre os diagnósticos dos pacientes i e j.

Como um paciente pode apresentar mais de uma doença, o peso entre dois pacientes foi definido pela maior similaridade encontrada entre seus respectivos diagnósticos:

$$W_{ij} = \max_{a\in D_i,\, b\in D_j} w(a,b), \tag{7}$$

em que D_i e D_j representam os conjuntos de doenças dos pacientes. A diagonal da matriz foi definida como 1, garantindo peso máximo para a correspondência de um paciente consigo mesmo.

A matriz completa foi calculada previamente ao treinamento. Durante o processamento de cada lote, foram selecionadas apenas as linhas e colunas correspondentes aos pacientes presentes naquele lote.

4) Arquitetura de Projeção: As representações tabulares e textuais foram processadas por duas cabeças de projeção independentes. Cada cabeça foi composta por uma projeção linear, ativação *GELU*, uma segunda camada linear, *dropout*, conexão residual e normalização em camadas.

Os *embeddings* projetados foram normalizados pela norma L_2 e representados em um espaço latente compartilhado de 256 dimensões:

$$\mathbf{z}_i^{tab} = f_{tab}(\mathbf{x}_i^{tab}), \qquad \mathbf{z}_i^{text} = f_{text}(\mathbf{x}_i^{text}). \tag{8}$$

Como os *encoders TabICL* e *ClinicalBERT* permaneceram congelados, somente os parâmetros das duas cabeças de projeção foram atualizados durante o alinhamento.

5) Soft Hierarchical InfoNCE: O alinhamento ontológico foi realizado com a Soft Hierarchical InfoNCE apresentada nas Equações 2 e 3. Diferentemente da InfoNCE tradicional, essa formulação utiliza pesos ontológicos normalizados como alvos suaves, permitindo que pacientes clinicamente próximos contribuam parcialmente para o alinhamento.

Na implementação, a matriz de pesos paciente–paciente W foi normalizada por linha para produzir os pesos $\tilde{W}_{ij}$ utilizados na função de perda. A similaridade entre os *embeddings* tabulares e textuais foi calculada no espaço latente compartilhado, após a projeção e normalização dos vetores.

A perda foi calculada nas duas direções, Tabular $\rightarrow$ Texto e Texto $\rightarrow$ Tabular, e a função objetivo final correspondeu à média entre essas direções. Assim, o par tabular–texto do mesmo paciente permaneceu com peso máximo, enquanto pacientes com doenças clinicamente relacionadas também contribuíram para o alinhamento de acordo com sua proximidade na ICD-10.

6) Configuração do Treinamento: O alinhador foi treinado utilizando o otimizador *AdamW* e um escalonador cosseno para a taxa de aprendizado. Foram utilizados lotes de 64 pacientes, espaço latente de 256 dimensões, taxa de aprendizado de aproximadamente $2{,}5 \times 10^{-5}$, decaimento de peso igual a 0,02 e temperatura $\tau = 0{,}01$.

O treinamento foi configurado para até 200 épocas. Em cada lote, os índices dos pacientes foram utilizados para recuperar a submatriz correspondente da matriz ontológica completa. A *Soft Hierarchical InfoNCE* foi utilizada como única função objetivo nessa configuração.

7) Construção da Representação Multimodal: Após o alinhamento, as projeções tabular e textual foram combinadas por meio de uma média vetorial:

$$\mathbf{z}_i^{mm} = \mathrm{norm}\left(\frac{\mathbf{z}_i^{tab} + \mathbf{z}_i^{text}}{2}\right), \tag{9}$$

em que $\mathrm{norm}(\cdot)$ representa a normalização pela norma L_2. Essa representação foi utilizada como *embedding* multimodal do paciente nas análises posteriores.

IV. Resultados e Discussão

A. Comparação na Tarefa Downstream: Pneumonia

A primeira avaliação *downstream* considerou a tarefa de classificação binária para pneumonia. Essa condição representa um caso desafiador no *SimSUM*, pois possui baixa frequência no conjunto de dados. Por isso, além da acurácia, foram analisadas métricas mais adequadas para dados desbalanceados, como AUROC, AUPRC, *recall*, precisão e *F1-score*.

A Tabela I apresenta a comparação entre diferentes representações: *embeddings* textuais do *ClinicalBERT* congelado, *ClinicalBERT* com ajuste fino, *TabICL* avaliado diretamente, *embeddings* alinhados tabulares (z_{tab}), *embeddings* alinhados textuais (z_{text}) e a concatenação das duas modalidades.

Os resultados mostram que a concatenação das representações alinhadas apresentou o maior AUROC, com valor de 0.9729, indicando melhor capacidade geral de separação entre

Tabela I
COMPARAÇÃO DOS MODELOS NA TAREFA *downstream* DE CLASSIFICAÇÃO DE PNEUMONIA. OS MELHORES VALORES DE CADA MÉTRICA ESTÃO DESTACADOS EM NEGRITO.

Modelo / Representação	AUROC	AUPRC	Acc.@BestTh	Prec.@BestTh	Rec.@BestTh	F1@BestTh
ClinicalBERT puro	0.8718	0.2263	**0.9892**	0.4615	0.3000	0.3636
ClinicalBERT fine-tuning rápido	0.9576	0.3578	**0.9895**	0.5000	0.4762	0.4878
TabICLClassifier direto	0.9477	**0.4267**	0.9885	0.4615	0.5714	**0.5106**
Alinhado z_{tab}	0.9635	0.2741	0.9790	0.2941	**0.7500**	0.4225
Alinhado z_{text}	0.8793	0.1942	0.9769	0.1951	0.4000	0.2623
Concatenação (z_{tab}, z_{text})	**0.9729**	0.3076	0.9851	0.3636	0.6000	0.4528

casos positivos e negativos. No entanto, o melhor *F1-score* foi obtido pelo TabICLClassifier direto, com F1 de 0.5106, além da maior AUPRC, 0.4267. Esse resultado indica que a representação tabular foi especialmente relevante para a identificação de pneumonia.

O modelo alinhado z_{tab} apresentou o maior *recall*, alcançando 0.7500 no melhor limiar. Esse comportamento sugere que a representação tabular alinhada favoreceu a recuperação de casos positivos, ainda que com menor precisão. Já o *ClinicalBERT* congelado teve desempenho inferior, principalmente em AUPRC e F1-score, o que reforça a limitação do uso de *embeddings* textuais congelados para uma classe rara.

Com o ajuste fino do ClinicalBERT, houve melhora em relação ao ClinicalBERT congelado, principalmente em AUROC, AUPRC e *F1-score*. Ainda assim, os melhores resultados gerais permaneceram associados às representações tabulares ou multimodais. Isso sugere que, para tarefas intrinsecamente raras e com maior dificuldade de identificação, como é o caso da pneumonia (cerca de 1% de casos positivos dentro do *dataset*), os atributos estruturados do *SimSUM* carregam sinais importantes para a tarefa *downstream*, enquanto a modalidade textual isolada apresentou maior dificuldade.

B. Análise do Espaço Latente

A organização do espaço latente foi analisada por meio de *PCA*, comparando as projeções multimodais e textuais após o alinhamento. O objetivo dessa análise foi verificar se as representações aprendidas apresentavam estrutura semântica associada às condições clínicas, mesmo sem supervisão explícita direta na visualização.

As figuras 1, 2 e 3 apresentam três casos representativos. O resfriado comum representa uma condição com maior presença no conjunto de dados e apresenta uma região mais ampla e visualmente definida no espaço latente. A asma mostra uma organização intermediária, com presença de agrupamentos, mas também sobreposição com outras regiões. Em contraste, pneumonia aparece com poucos exemplos positivos e maior dispersão, indicando maior dificuldade de formação de uma região semântica bem delimitada.

Os resultados indicam que o alinhamento entre *TabICL* e *ClinicalBERT* produziu um espaço latente com alguma organização clínica. Condições mais frequentes, como resfriado comum, tendem a ocupar regiões mais evidentes, sugerindo que o modelo capturou padrões consistentes entre dados tabulares e textuais. Por outro lado, pneumonia apresenta comportamento mais disperso, compatível com sua baixa frequência

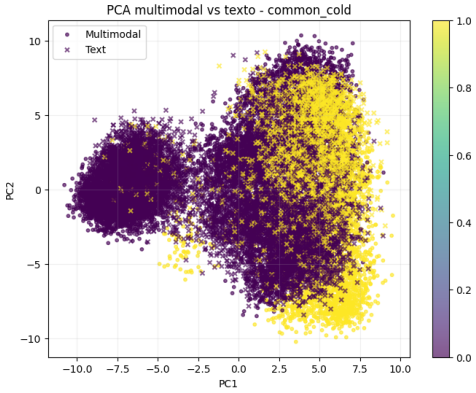

Figura 1. Projeção PCA do espaço latente para resfriado comum. Observa-se maior concentração de exemplos positivos em regiões específicas do espaço.

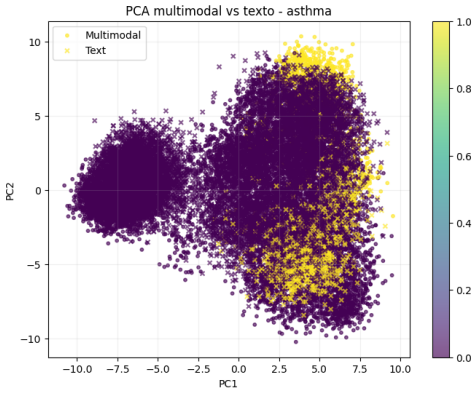

Figura 2. Projeção PCA do espaço latente para asma. A distribuição apresenta uma organização intermediária entre concentração e dispersão dos exemplos positivos.

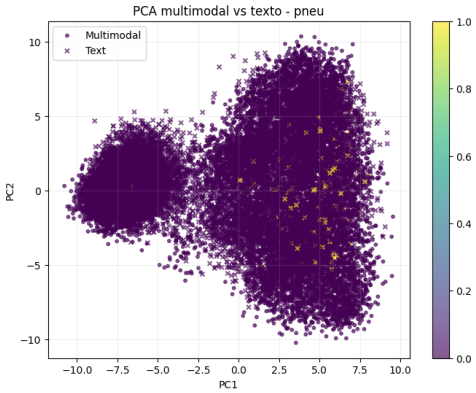

Figura 3. Projeção PCA do espaço latente para pneumonia. Os exemplos positivos aparecem de forma mais esparsa no espaço latente.

no *SimSUM* e com os resultados inferiores observados nas métricas *downstream*.

Essa análise também reforça que o espaço aprendido não se limita a memorizar pares individuais. A presença de regiões associadas a determinadas doenças sugere que o modelo começou a organizar os pacientes de acordo com similaridades

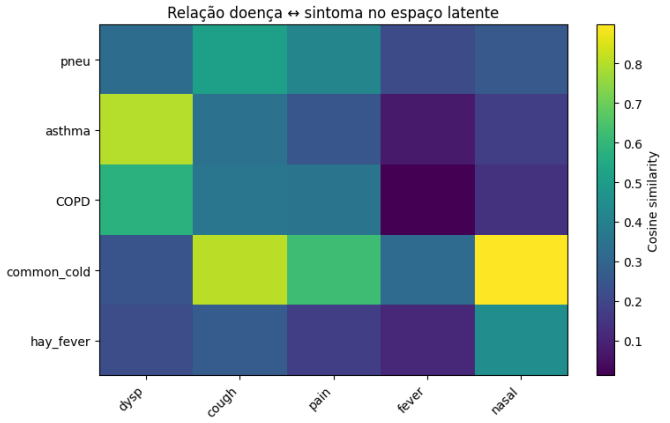

Figura 4. Similaridade cosseno entre doenças e sintomas no espaço latente multimodal. Cores mais intensas indicam maior similaridade.

clínicas.

C. Relação entre Doenças e Sintomas

Além da projeção por *PCA*, foi analisada a relação entre doenças e sintomas no espaço latente aprendido. Para isso, foi calculada a similaridade cosseno entre as representações associadas às condições clínicas e aos sintomas presentes no *SimSUM*. Essa análise permite verificar se o espaço multimodal preserva relações semanticamente coerentes entre doenças respiratórias e seus sintomas característicos.

A Figura 4 mostra que o espaço latente capturou associações clinicamente plausíveis. Asma e DPOC apresentaram maior similaridade com dispneia, o que é coerente com o perfil respiratório dessas condições. O resfriado comum apresentou forte associação com tosse, dor e sintomas nasais, refletindo padrões típicos dessa condição no conjunto de dados. A rinite alérgica apresentou maior relação com sintomas nasais, também compatível com sua manifestação clínica esperada.

Pneumonia apresentou associação moderada com tosse e dor, mas sem uma concentração tão forte quanto as demais condições. Esse comportamento é consistente com os resultados anteriores, nos quais pneumonia apareceu como uma classe mais difícil de representar, possivelmente devido à sua menor frequência no conjunto de dados e à sobreposição de sintomas com outras doenças respiratórias.

De modo geral, a matriz de similaridade sugere que o alinhamento multimodal não apenas aproximou dados tabulares e textuais, mas também organizou o espaço latente de forma clinicamente interpretável.

D. Retrieval Semântico

A avaliação por *retrieval* semântico foi utilizada para verificar se os *embeddings* alinhados eram capazes de recuperar pacientes clinicamente relacionados entre modalidades. Foram analisadas as direções Tabular $\rightarrow$ Texto e Texto $\rightarrow$ Tabular, utilizando as métricas *Recall@K* e *Precision@K*.

O *Recall@K* mede se ao menos um exemplo semanticamente relevante foi recuperado entre os K vizinhos mais próximos. Já a *Precision@K* mede a proporção de exemplos relevantes entre os K resultados recuperados. Dessa forma, o *recall* avalia a capacidade de encontrar algum paciente clinicamente relacionado, enquanto a precisão avalia a qualidade geral da lista de vizinhos recuperados.

A Figura 9 mostra que a direção Tabular $\rightarrow$ Texto apresentou desempenho consistente para a maioria das doenças. Asma, DPOC e rinite alérgica alcançaram valores elevados de *Recall@K* e *Precision@K*, próximos de 1 em diferentes valores de K. Esse resultado indica que, para essas condições, os *embeddings* tabulares foram capazes de recuperar textos clinicamente relacionados com alta consistência.

O resfriado comum apresentou desempenho intermediário, com *recall* elevado, mas precisão inferior às condições mais bem definidas. Isso sugere que, embora o modelo consiga recuperar exemplos relevantes, há maior mistura com outras condições respiratórias. Pneumonia foi o caso mais difícil, apresentando os menores valores de *Precision@K* e *Recall@K*, especialmente para $K = 1$. Esse comportamento é coerente com as análises anteriores, nas quais pneumonia também apresentou menor separação no espaço latente e desempenho mais limitado na tarefa *downstream*.

Na direção Texto $\rightarrow$ Tabular, observou-se comportamento semelhante, com bom desempenho para asma, DPOC e rinite alérgica, mas maior dificuldade para pneumonia. De modo geral, os resultados indicam que o alinhamento multimodal foi capaz de preservar relações clínicas relevantes entre as modalidades, embora a recuperação de classes raras permaneça como um desafio.

E. Resultados da Segunda Etapa: Alinhamento Ontológico

Após a identificação das limitações da *InfoNCE* tradicional, foi avaliada uma segunda estratégia de alinhamento baseada na *Soft Hierarchical InfoNCE*. Nesta etapa, os embeddings tabulares pós-ICL do TabICL foram alinhados aos embeddings textuais do ClinicalBERT ainda congelado, incorporando pesos ontológicos derivados da hierarquia ICD-10.

A avaliação foi realizada com a métrica Ontology Weighted Recall@K (*OWR@K*), que considera não apenas correspondências exatas, mas também recuperações clinicamente próximas segundo os pesos ontológicos. Assim, um exemplo recuperado com uma doença relacionada recebe crédito parcial, de acordo com sua proximidade na hierarquia ICD-10.

Para uma consulta q, a métrica foi definida como:

$$OWR@K(q) = \frac{\sum_{r \in TopK(q)} w(D_q, D_r)}{\sum_{r \in BestK(q)} w(D_q, D_r)}. \qquad (10)$$

Nessa equação, $TopK(q)$ representa os K exemplos recuperados pelo modelo, enquanto $BestK(q)$ representa os K exemplos ideais segundo os pesos ontológicos.

Os resultados da Tabela II mostram que a direção Tabular $\rightarrow$ Texto apresentou desempenho elevado em todos os valores de K, com OWR@K próximo de 1 desde $K = 1$. Esse comportamento indica que, quando a consulta parte dos embeddings tabulares, o modelo consegue recuperar textos semanticamente próximos segundo a estrutura ontológica definida pela ICD-10.

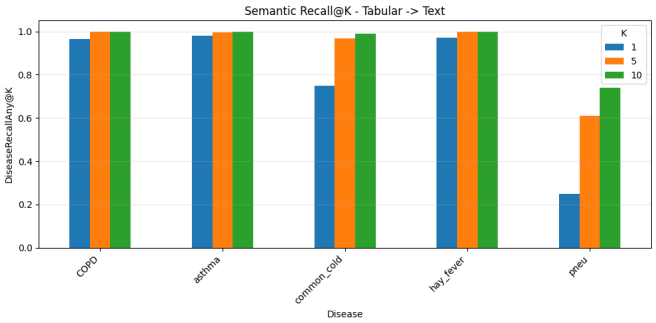

Figura 5. (a) Recall@K: Tabular → Texto

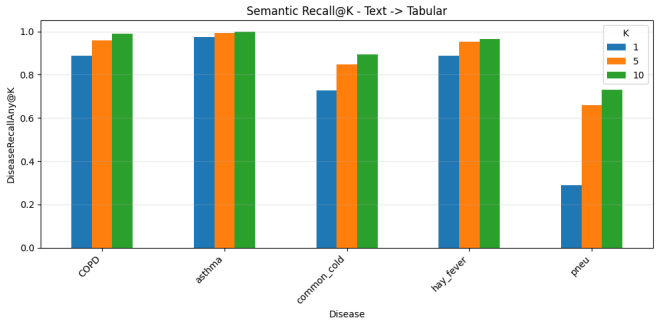

Figura 6. (b) Recall@K: Texto → Tabular

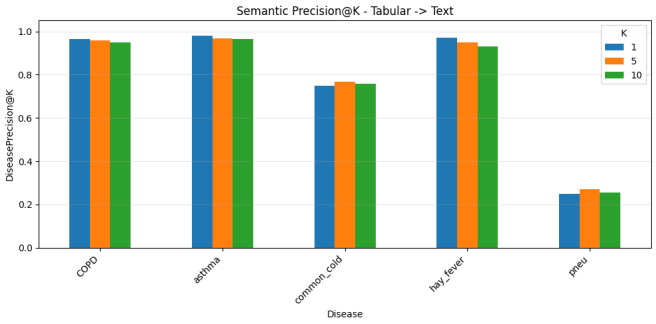

Figura 7. (c) Precision@K: Tabular → Texto

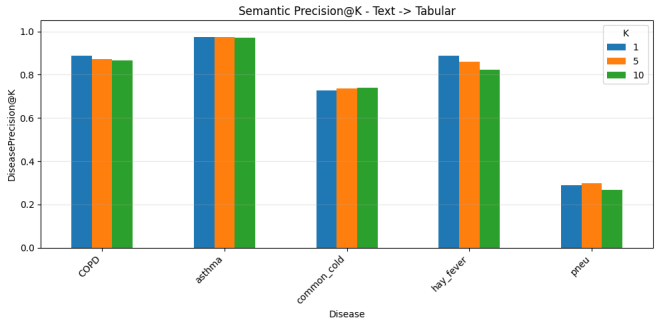

Figura 8. (d) Precision@K: Texto → Tabular

Figura 9. Avaliação de retrieval semântico nas direções Tabular → Texto e Texto → Tabular.

Tabela II
RESULTADOS DE OWR@K PARA O ALINHAMENTO ONTOLÓGICO COM CLINICALBERT CONGELADO.

K	Tabular → Texto	Texto → Tabular	Média
1	**0.9849**	0.8599	0.9224
5	**0.9856**	0.9171	0.9514
10	**0.9840**	0.9404	0.9622
20	**0.9822**	0.9579	0.9700
50	**0.9775**	0.9695	0.9735
100	0.9715	**0.9718**	**0.9716**

Na direção Texto → Tabular, o desempenho inicial foi inferior, com OWR@1 de 0.8599. Entretanto, os valores aumentaram progressivamente conforme K cresceu, alcançando 0.9718 em $K = 100$. Esse resultado sugere que os *embeddings* textuais ainda apresentavam menor precisão no topo do *ranking*, mas conseguiam recuperar exemplos clinicamente relevantes quando uma vizinhança maior era considerada.

Comparando com a primeira etapa, a métrica *OWR@K* evidencia um comportamento importante: a incorporação da ontologia permite avaliar e favorecer recuperações clinicamente próximas, reduzindo a penalização rígida entre doenças relacionadas. Dessa forma, a *Soft Hierarchical InfoNCE* representa um avanço metodológico em relação à *InfoNCE* tradicional, pois torna o alinhamento mais compatível com relações graduais entre condições clínicas.

Apesar dos resultados positivos, a diferença entre as direções indica que a modalidade textual ainda era o ponto mais frágil do alinhamento. Como o *ClinicalBERT* permaneceu congelado nesta etapa, suas representações textuais não foram adaptadas diretamente ao domínio do *SimSUM* nem à tarefa de alinhamento ontológico. Essa observação motiva os próximos passos, no qual as últimas camadas do *ClinicalBERT* podem ser descongeladas para ajuste fino antes do treinamento com a *Soft Hierarchical InfoNCE*.

V. CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho investigou o alinhamento multimodal entre dados tabulares e textos clínicos no contexto de registros eletrônicos de saúde. Para isso, foram utilizadas representações tabulares extraídas com *TabICL* e representações textuais obtidas com *ClinicalBERT*, posteriormente projetadas para um espaço latente compartilhado.

Na primeira etapa, os resultados mostraram que a *InfoNCE* tradicional foi capaz de produzir um alinhamento consistente entre as modalidades. A análise por *PCA* indicou a formação de regiões associadas a diferentes condições clínicas, enquanto a matriz de similaridade doença–sintoma mostrou relações semanticamente coerentes, como a proximidade entre asma, DPOC e dispneia, e entre resfriado comum, tosse e sintomas nasais. As métricas de retrieval semântico também indicaram desempenho elevado para doenças como asma, DPOC e rinite alérgica.

Apesar desses resultados, a primeira etapa revelou limitações importantes. Pneumonia apresentou menor desempenho na tarefa *downstream*, maior dispersão no espaço latente e

piores valores de retrieval, o que sugere dificuldade na representação de classes raras. Além disso, a *InfoNCE* tradicional trata todos os exemplos não correspondentes como negativos, mesmo quando representam doenças clinicamente próximas.

Para lidar com essa limitação, a segunda etapa incorporou conhecimento ontológico da *ICD-10* por meio da *Soft Hierarchical InfoNCE*. Essa abordagem permitiu substituir a separação rígida entre positivos e negativos por pesos contínuos baseados na proximidade clínica entre doenças. Os resultados com *Ontology Weighted Recall@K* indicaram alto desempenho na direção Tabular $\rightarrow$ Texto e melhora progressiva na direção Texto $\rightarrow$ Tabular conforme o valor de K aumentou.

Conclui-se que o alinhamento multimodal entre *TabICL* e *ClinicalBERT* é viável e produz representações clinicamente interpretáveis. Além disso, a incorporação de informação ontológica mostrou-se uma estratégia promissora para tornar o espaço latente mais coerente com relações médicas, especialmente em cenários nos quais doenças distintas compartilham sintomas ou pertencem a grupos clínicos próximos.

Como trabalhos futuros, pretende-se aprofundar em algumas direções principais. Primeiro, investigar a limitação observada na direção Texto $\rightarrow$ Tabular, uma vez que o uso do ClinicalBERT congelado restringe a adaptação da representação textual ao domínio específico do SimSUM. Avaliar ontologias que possam enriquecer o processo de treinamento, como por exemplo, hierárquias que envolvam, também, sintomas relacionados às doenças.

Além disso, aplicar *fine-tuning* parcial nas últimas camadas do *transformer*, buscando melhorar a qualidade dos embeddings textuais e reduzir a assimetria entre as direções de *retrieval*. Comparar de forma sistemática a *InfoNCE* tradicional e a *Soft Hierarchical InfoNCE*, avaliando se a inclusão da ontologia altera a organização do espaço latente e preserva melhor relações clínicas. Buscar equalizar o desempenho entre as direções Tabular $\rightarrow$ Texto e Texto $\rightarrow$ Tabular, maximizando a coerência semântica global dos *embeddings* multimodais, e, por último, permitir que o trabalho seja aplicado à cenários e datasets reais e de larga escala.

REFERÊNCIAS

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[2] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021. [Online]. Available: https://arxiv.org/abs/2108.07258

[3] Y. Si, J. Du, Z. Li, X. Jiang, T. Miller, F. Wang, W. Jim Zheng, and K. Roberts, "Deep representation learning of patient data from electronic health records (ehr): A systematic review," *Journal of Biomedical Informatics*, vol. 115, p. 103671, 2021. [Online]. Available: http://dx.doi.org/10.1016/j.jbi.2020.103671

[4] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj *et al.*, "Scalable and accurate deep learning with electronic health records," *npj Digital Medicine*, vol. 1, no. 18, 2018. [Online]. Available: https://www.nature.com/articles/s41746-018-0029-1

[5] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep ehr: A survey of recent advances in deep learning techniques for electronic health record analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1589–1604, 2017. [Online]. Available: https://arxiv.org/abs/1706.03446

[6] T. Bonnier, "Revisiting multimodal transformers for tabular data with text fields," in *Findings of the Association for Computational Linguistics*, 2024.

[7] Arazi et al., "Multabench: Benchmarking multimodal tabular learning with text and image," 2026.

[8] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," *Scientific Reports*, vol. 6, no. 26094, 2016. [Online]. Available: https://www.nature.com/articles/srep26094

[9] P. Rabaey, S. Heytens, and T. Demeester, "Simsum: Simulated benchmark with structured and unstructured medical records," 2024. [Online]. Available: https://arxiv.org/abs/2409.08936

[10] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *KDD*, 2016. [Online]. Available: https://arxiv.org/abs/1603.02754

[11] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," *Advances in Neural Information Processing Systems*, 2021. [Online]. Available: https://arxiv.org/abs/2106.11959

[12] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, "Tabtransformer: Tabular data modeling using contextual embeddings," 2020. [Online]. Available: https://arxiv.org/abs/2012.06678

[13] J. Qu, D. Holzmüller, G. Varoquaux, and M. Le Morvan, "Tabicl: A tabular foundation model for in-context learning on large data," 2025. [Online]. Available: https://arxiv.org/abs/2502.05564

[14] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott, "Publicly available clinical bert embeddings," in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 2019, pp. 72–78. [Online]. Available: https://aclanthology.org/W19-1909

[15] H. Li, X. Zhou, A. T. Luu, and C. Miao, "Rethinking negative pairs in code search," 2023. [Online]. Available: https://arxiv.org/abs/2310.08069

APÊNDICE

Tabela III
EXEMPLOS REAIS DE REGISTROS ESTRUTURADOS E TEXTUAIS DO SIMSUM.

ID	Condição	Estação	Disp.	Tosse	Febre	Nasal	Dias	Trecho do texto clínico
6	Asma	Inverno	Sim	Sim	Baixa	Não	5	*"The patient presents with a low grade fever and reports experiencing significant dyspnea characterized as air hunger."*
16	DPOC	Inverno	Sim	Não	Nenhuma	Não	2	*"The patient presents with dyspnea, particularly noticeable when attempting to sleep, leading to frequent nighttime awakenings."*
3	Rinite alérgica	Verão	Não	Não	Baixa	Sim	2	*"Patient presents with low-grade fever and significant nasal symptoms, including congestion and runny nose."*
276	Pneumonia	Inverno	Não	Sim	Alta	Não	1	*"The patient reports a persistent light cough over the past three days. There is a notable presence of high fever."*
5	Resfriado comum	Inverno	Não	Sim	Alta	Sim	4	*"The patient reports having a light respiratory pain and nasal symptoms, including nasal congestion and a runny nose."*

Nota: "Disp." corresponde a dispneia e "Dias" ao número de dias de afastamento domiciliar. Foi selecionado um registro representativo de cada condição, sem presença simultânea das demais doenças consideradas.