Revista Eletrônica de Iniciação Científica em Computação, 2026, 24:1
ISSN: 3085-8461 • doi: 10.5753/reic.2026.8498

Artigo de Pesquisa/Research Paper

Interpretabilidade Mecanística Dinâmica para Modelos Fundacionais Tabulares na Saúde

Dynamic Mechanistic Interpretability for Tabular Foundation Models in Healthcare

João Marcos Campos [Universidade Federal de Minas Gerais (UFMG) | *joaomarcoscampos@dcc.ufmg.br*]
Rafael Martins Gomes [Universidade Federal de Minas Gerais (UFMG)]
Diogo Tuler Chaves [Universidade Federal de Minas Gerais (UFMG)]
Wagner Meira Jr. [Universidade Federal de Minas Gerais (UFMG)]
Mariangela Leal Cherchiglia [Universidade Federal de Minas Gerais (UFMG)]
Hugo André da Rocha [Universidade Federal de Minas Gerais (UFMG)]
Leonardo Rocha [Universidade Federal de São João del-Rei (UFSJ)]
Marcos André Gonçalves [Universidade Federal de Minas Gerais (UFMG)]

Departamento de Ciência da Computação, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brasil.

Resumo. Modelos Fundacionais Tabulares (TFMs) apresentam alto desempenho preditivo, mas permanecem vulneráveis ao desvio temporal (*drift*) na saúde. Propomos um arcabouço de Interpretabilidade Mecanística Dinâmica que os torna auditáveis e sensíveis ao tempo. Nosso pipeline processa representações internas em lotes e usa *Sparse Autoencoders* (SAEs) para separar padrões de ativação sobrepostos. Ao adaptar *Testing with Concept Activation Vectors* com árvores de decisão substitutas, identificamos padrões de risco clínico sem anotações manuais. Em um registro longitudinal de diálise renal, a abordagem mostra que as previsões se apoiam em conceitos latentes estáveis, distinguindo o desvio populacional de mudanças no raciocínio interno do modelo.

Abstract.
Tabular Foundation Models (TFMs) achieve high predictive performance but remain opaque and vulnerable to temporal drift in healthcare. We propose a Dynamic Mechanistic Interpretability framework that makes them auditable and time-aware. Our pipeline processes internal representations in batches and uses Sparse Autoencoders (SAEs) to disentangle overlapping activation patterns. By adapting Testing with Concept Activation Vectors (TCAV) together with surrogate decision trees, we identify clinical risk patterns without manual annotations. Evaluated on a longitudinal renal dialysis registry (1997–2015), the approach shows that predictions rely on stable latent concepts, distinguishing population drift from changes in the model's internal reasoning.

Palavras-chave: Interpretabilidade mecanística, modelos fundacionais tabulares, saúde, inteligência artificial, explicabilidade

Keywords: Mechanistic interpretability, tabular foundation models, healthcare, artificial intelligence, explainability

Recebido/Received: 16 June 2026 • **Aceito/Accepted:** 17 June 2026 • **Publicado/Published:** 10 July 2026

1 Introdução

A tomada de decisão clínica é inerentemente conceitual. Médicos não diagnosticam pacientes com base em variáveis isoladas, mas através da combinação de padrões clínicos de alto nível, como fenótipos, trajetórias de tratamento e síndromes. Por exemplo, um nefrologista pode interpretar conjuntamente frequência de hemodiálise, uso de eritropoetina e presença de internações como evidência de agravamento progressivo da doença renal.

Entretanto, modelos modernos de aprendizado de máquina em saúde, especialmente Modelos Fundacionais Tabulares (TFMs) baseados em Transformers, operam sobre representações latentes densas e altamente distribuídas, dificultando compreender quais conceitos clínicos internos sustentam suas decisões Molnar [2022]. Modelos Fundacionais Tabulares (TFMs) são modelos pré-treinados em grandes coleções de tabelas e capazes de generalizar para novas tarefas por meio de inferência contextual (*in-context learning*), sem necessidade de retreinamento específico.

Esse desafio é agravado pela mudança de distribuição temporal (*drift*). À medida que as populações envelhecem e os protocolos de tratamento evoluem, o raciocínio interno de um modelo pode mudar de maneiras sutis, mesmo quando seu desempenho preditivo global parece estável. Os métodos de interpretabilidade existentes são predominantemente estáticos e focados na importância de atributos individuais, carecendo de mecanismos para auditar como os conceitos latentes do modelo reagem a mudanças longitudinais.

Para mitigar essas limitações, propomos um arcabouço de **Interpretabilidade Mecanística Dinâmica** para TFMs. Nossa premissa central é que a implantação confiável de modelos de aprendizado de máquina em cenários críticos exige monitorar não apenas *o que* o modelo prevê, mas *quais conceitos internos ele utiliza* e *como esses conceitos evoluem no tempo*. Para operacionalizar essa ideia, adaptamos o método *Testing with Concept Activation Vectors* (TCAV) Kim *et al.* [2018] como um instrumento de auditoria dinâmica.

Estendemos o TCAV em dois eixos fundamentais para auditoria em cenários reais. Primeiro, eliminamos a necessidade de definir conceitos *a priori*, utilizando *Sparse Autoencoders* (SAEs) para descobrir, de forma não supervisionada, fenótipos clinicamente significativos diretamente a partir do espaço latente do modelo. Em segundo lugar, introduzimos uma perspectiva temporal, calculando direções e sensibilidades de conceitos em janelas de tempo deslizantes. Isso permite uma separação clara entre o *desvio populacional* (mudanças na prevalência de características dos pacientes) e o *desvio de conceito* (mudanças na semântica interna e na lógica de decisão do modelo).

Validamos nosso arcabouço utilizando um amplo conjunto de dados longitudinais de registros eletrônicos de saúde referentes a diálise renal, abrangendo quase duas décadas. Em resumo, nossas principais contribuições são:

- **Auditoria de modelos em larga escala:** Introduzimos um *pipeline* baseado em SAEs que processa ativações de *Transformers* em lotes gerenciáveis, viabilizando a inspeção de representações internas com eficiência de memória.
- **Monitoramento sob condições dinâmicas:** Propomos um arcabouço capaz de distinguir mudanças na população de pacientes de alterações no raciocínio do modelo, garantindo que a lógica de decisão clínica permaneça estável.

Esse é um trabalho relacionado ao projeto de iniciação científica de João Marcos Campos. Até o presente momento, esse projeto resultou na publicação em um artigo na *World Conference on eXplainable Artificial Intelligence - XAI 2026* Campos *et al*. [2026] e outro artigo submetido e em processo de avaliação ECML-PKDD 2026.

2 Trabalhos Relacionados

A interpretabilidade em aprendizado de máquina clínico tem sido dominada por técnicas de atribuição de características, como SHAP e LIME Marcolino *et al*. [2025]; Molnar [2022]. Esses métodos fornecem explicações estáticas baseadas em variáveis individuais, falhando em capturar padrões clínicos de alto nível. Já abordagens baseadas em conceitos, como o *Testing with Concept Activation Vectors* (TCAV) Kim *et al*. [2018], que mede a dependência preditiva em relação a conceitos semânticos por meio de derivadas direcionais no espaço de ativação. Para remover a dependência de anotações manuais do TCAV, o *Automated Concept-based Explanations* (ACE) Ghorbani *et al*. [2019] propôs a descoberta automática de conceitos. Em dados tabulares, adaptações anteriores limitaram-se a predicados booleanos rígidos e manuais Pendyala *et al*. [2022].

Paralelamente, a adaptação da arquitetura *Transformer* Vaswani *et al*. [2017] para dados tabulares culminou em Modelos Fundacionais Tabulares (TFMs), como o TabPFN Hollmann *et al*. [2023]. Através do *in-context learning*, o TabPFN infere regras preditivas sem a necessidade de retreinamento. Para lidar com a opacidade dessas representações densas, *Sparse Autoencoders* (SAEs) surgiram como uma ferramenta para desemaranhar (*disentangle*) representações internas em características esparsas e interpretáveis Cunningham *et al*. [2023]; Bricken *et al*. [2023]. Diferentemente de arquiteturas intrinsecamente interpretáveis, como o XNNTab Elhadri *et al*. [2024], aplicamos SAEs de forma *post-hoc* para descobrir fenótipos latentes em modelos pré-treinados.

Finalmente, no contexto de dados dinâmicos, o *Drift-Resilient TabPFN* Helli *et al*. [2024] melhora a robustez preditiva ao treinar com simulações de mudanças temporais. No entanto, embora aprimore *o que* o modelo prevê sob *drift*, não explica *como* o raciocínio interno se adapta. A maioria das estratégias de monitoramento Paiva *et al*. [2024] foca apenas na degradação da acurácia. Nosso trabalho preenche essa lacuna ao combinar SAEs, TCAV e análise temporal, fornecendo uma visão mecanística sobre como a lógica de decisão do modelo evolui ao longo do tempo.

3 Metodologia

Propomos um arcabouço de *Interpretabilidade Mecanística Dinâmica* para analisar Modelos Fundacionais Tabulares (TFMs) em aplicações clínicas. Por TFM, entendemos um modelo pré-treinado para dados tabulares que pode ser aplicado a novas bases com pouco ou nenhum ajuste específico, utilizando exemplos rotulados como contexto para produzir predições para novas observações. No caso deste trabalho, utilizamos o *Drift-Resilient TabPFN*, um modelo fundacional tabular projetado para lidar com mudanças temporais na distribuição dos dados.

Nosso objetivo é responder duas perguntas complementares: (i) *quais padrões clínicos latentes o modelo parece usar para predizer risco?* e (ii) *esses padrões permanecem estáveis ao longo do tempo?* Para isso, combinamos avaliação temporal, extração de representações internas, decomposição não supervisionada de fatores latentes, tradução desses fatores em regras clínicas e auditoria por sensibilidade e intervenção. A Figura 1 resume o fluxo principal.

3.1 Avaliação Temporal e Extração de Representações

Em aplicações clínicas reais, os dados raramente permanecem estáveis ao longo do tempo. Mudanças em protocolos médicos, envelhecimento populacional, introdução de novos medicamentos ou alterações em práticas de codificação podem modificar tanto a frequência quanto o significado clínico dos atributos observados. Chamamos esse fenômeno de *mudança temporal de distribuição* (*temporal drift*). Por exemplo, um aumento no uso de eritropoetina ao longo dos anos pode refletir avanços no manejo da anemia em pacientes renais; nesse caso, a presença desse medicamento pode passar a representar maior adesão ao cuidado, e não apenas maior gravidade clínica.

Cada amostra da base corresponde a um *paciente em um dado ano* (*patient-year*). Assim, um mesmo paciente pode aparecer em múltiplos anos, permitindo observar trajetórias longitudinais de cuidado. Seja $\mathcal{D} = \{(x_i, y_i, t_i)\}_{i=1}^{N}$, onde $x_i \in \mathbb{R}^d$ representa os atributos clínicos de um paciente-ano, $y_i \in \{0, 1\}$ o desfecho e t_i o ano da observação. Para respeitar a natureza temporal dos dados, abandonamos a validação cruzada aleatória tradicional, que pressupõe observações i.i.d., e adotamos uma estratégia *walk-forward*: o modelo é condicionado em uma janela temporal passada e avaliado em

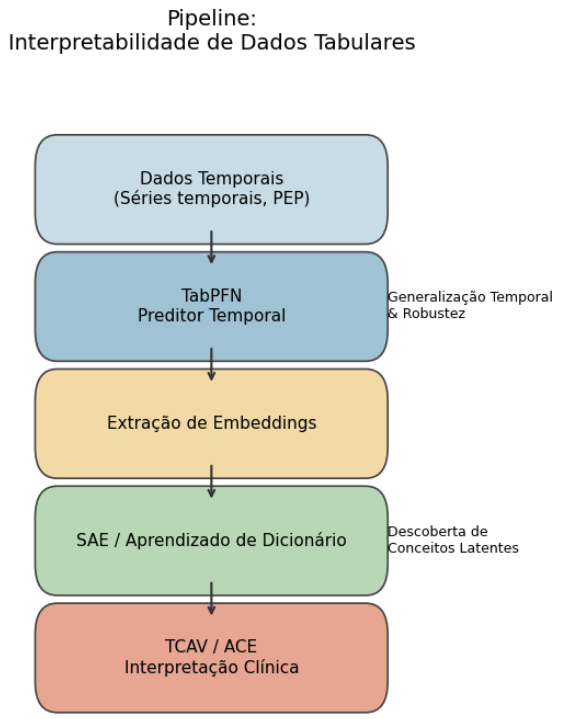

Figura 1. Pipeline de Interpretabilidade Mecanística Dinâmica.

períodos futuros.

O *Drift-Resilient TabPFN* recebe o ano como indicador de domínio temporal, permitindo que o modelo diferencie padrões persistentes de mudanças associadas à população, ao tratamento ou à codificação. Após a predição, extraímos representações internas do modelo no *token* de classificação (`y_token`). Intuitivamente, essas representações são vetores numéricos que resumem como o modelo organiza a evidência clínica de cada paciente-ano antes de produzir a predição. Para cada amostra, obtemos um vetor latente $z_i \in \mathbb{R}^D$, e normalizamos essas representações usando apenas estatísticas estimadas nos dados de treino, evitando vazamento temporal.

3.2 Decomposição Não Supervisionada de Conceitos

As representações latentes extraídas são densas e frequentemente polisemânticas: uma mesma dimensão do vetor pode contribuir para múltiplos padrões clínicos, e um mesmo padrão clínico pode estar distribuído em várias dimensões. Chamamos de *conceito latente* um fator interno que se ativa para um subconjunto de paciente-anos e que pode corresponder, por exemplo, a uma trajetória de cuidado como "hemodiálise regular com suporte medicamentoso", "baixa exposição à diálise com sinais de complicação" ou "uso de acesso vascular associado a internação".

Para decompor esse espaço em fatores mais interpretáveis, avaliamos duas estratégias. Como linha de base linear, utilizamos *Dictionary Learning* (DL), que representa cada embedding como uma combinação esparsa de átomos lineares. Como abordagem principal, empregamos *Sparse Autoencoders* (SAEs), que projetam os embeddings em um novo espaço latente expandido e esparso. Nesse espaço produzido pelo SAE, cada dimensão pode ser interpretada como um fator latente candidato: isto é, uma direção ou unidade de ativação que tende a responder a um padrão recorrente nos embeddings do modelo.

Dado um embedding $\mathbf{z}$, o SAE produz uma codificação esparsa $\mathbf{h}$ por meio de um codificador com ativação ReLU e reconstrói a entrada com um decodificador linear:

$$\mathbf{h} = \mathrm{ReLU}(\mathbf{W}_e\mathbf{z} + \mathbf{b}_e), \tag{1}$$

$$\hat{\mathbf{z}} = \mathbf{W}_d\mathbf{h} + \mathbf{b}_d. \tag{2}$$

O treinamento minimiza o erro de reconstrução e uma penalização de esparsidade:

$$\mathcal{L}_{\mathrm{SAE}} = \|\mathbf{z} - \hat{\mathbf{z}}\|_2^2 + \alpha\|\mathbf{h}\|_1. \tag{3}$$

Cada coordenada h_k da representação esparsa $\mathbf{h}$ corresponde, portanto, à ativação do fator latente k. Quando h_k assume valores altos apenas para um subconjunto específico de paciente-anos, interpretamos essa dimensão como um pré-candidato a conceito clínico. Assim, os conceitos analisados nas etapas seguintes não são definidos manualmente, mas emergem das dimensões do espaço latente aprendido pelo SAE.

3.3 Regras Clínicas e Validação por Conceitos

Os fatores extraídos pelo SAE ainda vivem no espaço latente do modelo, não diretamente no espaço clínico original. Por isso, usamos árvores de decisão como uma etapa de *tradução*: elas verificam se um fator latente pode ser descrito por regras simples envolvendo atributos clínicos observáveis.

Para cada fator k, definimos um alvo binário indicando se a ativação do fator é alta:

$$y_i^{(k)} = \mathbb{I}(a_i^{(k)} > \tau_k),$$

onde $a_i^{(k)}$ é a ativação do fator k para a amostra i, e τ_k é definido como o 50º percentil da distribuição de ativações positivas (ou seja, computada a partir de $a_i^{(k)} : a_i^{(k)} > 0$). Em seguida, treinamos uma árvore de decisão $T_k(x)$ usando os atributos clínicos originais como entrada. A árvore tenta produzir regras do tipo *IF-THEN* em função dos atributos do *dataset*, por exemplo: baixa frequência de hemodiálise, presença de eritropoetina e ausência de código de transplante. Mantemos apenas fatores cujas regras apresentam precisão mínima de 90% e *recall* mínimo de 25% em dados de validação, filtrando, assim, fatores que não possuem uma caracterização clínica confiável.

As regras extraídas passam, então, a definir operacionalmente o conjunto positivo do conceito: amostras que satisfazem a regra são consideradas exemplos do conceito. Com esses exemplos, treinamos um *Concept Activation Vector* (CAV), isto é, um vetor no espaço de embeddings que separa amostras que satisfazem a regra de amostras negativas escolhidas aleatoriamente. Em seguida, aplicamos TCAV para medir se mover a representação na direção desse conceito aumenta ou reduz a saída associada ao risco.

Formalmente, para um conceito k com CAV $\mathbf{v}_k$, o *score* TCAV é:

$$\mathrm{TCAV}_k = \frac{1}{|\mathcal{Z}|}\sum_{\mathbf{z}\in\mathcal{Z}} \mathbb{I}\left[\nabla_{\mathbf{z}} f(\mathbf{z}) \cdot \mathbf{v}_k > 0\right], \tag{4}$$

onde $f(\mathbf{z})$ é a saída do modelo para a classe de risco. Um valor próximo de 1 indica que o conceito aumenta consistentemente a predição de risco; um valor próximo de 0 indica associação protetiva; valores próximos de 0.5 indicam ausência de direção consistente — por isso, fatores com *score* no intervalo [0.4, 0.6] foram removidos. Além disso, a fim de garantir robustez, repetimos o treinamento dos CAVs com diferentes conjuntos negativos aleatórios e aplicamos testes estatísticos para preservar apenas conceitos com significância e tamanho de efeito relevantes.

3.4 Testes de Necessidade, Suficiência e Estabilidade Temporal

Por fim, avaliamos a importância funcional dos conceitos por meio de intervenções no espaço latente, inspiradas nos testes de destruição e suficiência do ACE. No teste de necessidade, removemos da representação a componente associada ao CAV de um conceito e observamos a variação na probabilidade predita. Se a predição cai após a remoção, isso sugere que o modelo utilizava aquela direção conceitual. No teste de suficiência, preservamos apenas a componente alinhada ao conceito e verificamos quanto da predição original é mantida. Esses testes não devem ser interpretados como causalidade clínica, mas como evidência de dependência funcional do modelo em relação ao conceito.

Como as observações são temporais, repetimos esses testes por domínio temporal. Isso permite avaliar se um conceito mantém a mesma influência ao longo dos anos ou se se torna sensível ao *drift*. Na prática, um conceito é considerado mais confiável para interpretação longitudinal quando sua regra clínica, seu *score* TCAV e seus efeitos de intervenção permanecem relativamente estáveis em diferentes períodos. Conceitos cuja influência varia fortemente entre anos são marcados como sensíveis ao *drift*, indicando que podem refletir mudanças em práticas clínicas, composição da coorte ou codificação dos dados.

4 Configuração Experimental

Avaliamos a robustez e a escalabilidade do nosso arcabouço em um amplo conjunto de dados de Registros Eletrônicos de Saúde (EHR) doSUS. Aplicamos o protocolo de avaliação temporal estrito detalhado na Seção 3.1, reportando os resultados ano a ano. O código-fonte está disponível em `https://github.com/joaomarcostomaz/TCAV`.

4.1 Base de Dados e Configuração do Modelo

O conjunto de dados *Renal* compreende trajetórias longitudinais de pacientes submetidos a procedimentos médicos no Sistema Único de Saúde (SUS) ao longo de 19 anos (1997–2015) Guerra Junior *et al*. [2018]. A base contém aproximadamente 9,6 milhões de registros clínicos associados a 67.267 pacientes únicos. Os dados possuem alta dimensionalidade, englobando mais de 6.200 tipos de eventos distintos (ex: hemodiálise, diagnósticos de doença renal crônica e administração de eritropoietina). A variável alvo é a ocorrência de ÓBITO (25.072 eventos registrados).

Devido à restrição de tamanho de contexto em modelos baseados em *Transformers*, adotamos uma estratégia de amostragem de Monte Carlo estratificada com um orçamento fixo de linhas por contexto. Essa abordagem preservou fielmente a distribuição de densidade de eventos da população original (teste de Kolmogorov-Smirnov, $p = 0,84$).

Utilizamos o modelo **Drift-Resilient TabPFN** Helli *et al*. [2024] operando com contexto fixo ($N_{ctx} = 128$). Para decompor seus *embeddings* de 192 dimensões, treinamos um **Sparse Autoencoder (SAE)** com pesos atrelados e fator de expansão $F = 1,5$ (gerando 288 fatores latentes), otimizado via Adam com penalidade de esparsidade $\lambda = 0,05$ para equilibrar a fidelidade de reconstrução e a esparsidade das ativações.

4.2 Métricas de Avaliação

A tarefa apresenta desbalanceamento natural entre as classes, uma vez que eventos de óbito representam minoria relativa na coorte longitudinal. Por essa razão, priorizamos métricas sensíveis à classe positiva, especialmente F1-score da classe de risco, em vez de acurácia global.

Avaliamos a qualidade das explicações extraídas em quatro dimensões consolidadas: (1) **Decomposição e Esparsidade**, medidas via Erro Quadrático Médio (MSE) de reconstrução, taxa de quase-zero (ativações $\leq 10^{-5}$), média de neurônios ativos por amostra e ortogonalidade direcional (pares de conceitos com similaridade de cosseno $> 0,5$); (2) **Significância**, filtrando correlações espúrias por meio de *scores* TCAV através de $N = 15$ execuções de *bootstrap* (d de Cohen $> 0,8$ e *p-values* ajustados por FDR); (3) **Fidelidade e Necessidade**, avaliando o mapeamento dos fenótipos via *Readout* Esparso (Lasso R^2) e a pontuação de destruição (Δ de queda preditiva após a ablação do conceito); e (4) **Validação de Regras**, atestando a qualidade das árvores de decisão substitutas.

5 Resultados e Discussão

Avaliamos o pipeline em um registro longitudinal de doença renal (1997–2015). Antes de inspecionar o raciocínio interno do modelo, verificamos seu desempenho preditivo. O *Drift-Resilient TabPFN* superou consistentemente os baselines na identificação de pacientes de alto risco, alcançando F1 da classe positiva de **35%** na janela de avaliação, contra **22%** do TabPFN padrão e **20%** do XGBoost. A auditoria mecanística foi então conduzida em três níveis: qualidade dos conceitos latentes, relevância interpretativa e tradução em fenótipos clínicos sensíveis ao tempo.

5.1 Decomposição do Espaço Latente

O primeiro desafio foi resolver a *superposição* de conceitos médicos em embeddings de alta dimensão. Para isso, comparamos uma linha de base linear, *Dictionary Learning* (DL), com o *Sparse Autoencoder* (SAE). O DL foi treinado com poucos fatores ($K = 8$), priorizando conceitos mais amplos e simples de descrever por regras. Já o SAE foi treinado com 288 dimensões latentes, o que corresponde a uma razão de superposição de 1,5 ($1,5 \times 192$), em linha com a literatura sobre interpretabilidade mecanística. Assim, o DL favorece *interpretabilidade direta*, enquanto o SAE prioriza *completude representacional*.

Como mostrado na Tabela 1, o DL não foi flexível o suficiente para reconstruir adequadamente o espaço de em-

beddings. O SAE apresentou erro quadrático médio muito menor (3.63×10^{-3}, contra 3.81×10^{-2} no DL) e ativações bem mais esparsas: 92,1% dos valores ficaram próximos de zero, contra 49,6% no DL. Embora use mais dimensões, o SAE ativa apenas uma pequena fração delas por amostra: em média, 22,8 de 288 unidades, contra 4 de 8 no DL, o que indica códigos internos compactos e seletivos.

Tabela 1. Comparação entre extração linear (DL) e não-linear (SAE). O SAE atinge reconstrução superior e ortogonalidade direcional, enquanto o DL produz átomos redundantes.

Método	MSE	Esparsidade	Ativos/Amostra	Sim. Direcional
Dictionary Learning	0,0381	49,60%	4 / 8	96% >0,5
Sparse Autoencoder	**0,0036**	**92,09%**	**22,8 / 288**	**0% >0,5**

Das 288 unidades do SAE, 82 permaneceram **inativas** em todo o conjunto de avaliação, sugerindo que o modelo depende de menos padrões do que o tamanho nominal do dicionário indicaria. Também avaliamos redundância e coativação. No DL, 96,6% dos pares de direções apresentaram similaridade cosseno acima de 0,5; no SAE, apenas 6 pares ultrapassaram esse limiar, com máximo de 0,66 e média de 0,06. A coativação também foi limitada: no DL, não houve pares acima de 0,5; no SAE, apenas 3,4% dos pares superaram esse limiar. Esses resultados indicam que o SAE organiza a informação fisiológica em pequenos grupos seletivos, com baixa redundância e baixa ativação simultânea.

5.2 Conceitos após Filtragem por Árvore de Decisão e TCAV

A interseção entre o filtro por escore TCAV e as regras extraídas por árvores de decisão resultou em **13 conceitos interpretáveis** (Tabela 2) entre os 288 fatores aprendidos pelo SAE. Esses fatores são simultaneamente relevantes para o comportamento do modelo e caracterizáveis por regras simples no espaço dos atributos. A filtragem preserva apenas 4,5% dos fatores originais, priorizando confiabilidade em detrimento de cobertura.

A Figura 2 ilustra como a combinação entre árvores de decisão e TCAV produz conceitos ao mesmo tempo *coerentes* e *relevantes*. No painel esquerdo, o Fator 270 ocupa uma região no espaço de embeddings; no painel direito, a regra extraída seleciona praticamente o mesmo subconjunto. Como o Fator 270 também apresenta um escore TCAV elevado, a visualização reforça que os conceitos retidos correspondem a padrões definíveis por regras e efetivamente utilizados nas predições.

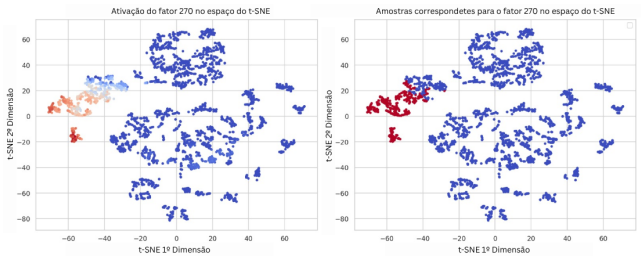

Figura 2. Fator 270 no espaço de embeddings e sua coorte definida por regra. À esquerda, projeção t-SNE dos embeddings coloridos pela ativação do fator. À direita, amostras que satisfazem a regra destacadas em vermelho.

5.3 Dos Vetores aos Fenótipos: Tradução Clínica

As regras detectadas foram utilizadas para traduzir os vetores conceituais validados em assinaturas clinicamente significativas, que foram posteriormente revisadas por **dois nefrologistas seniores**. Os especialistas confirmaram a plausibilidade clínica dos padrões extraídos. Os 13 fatores interpretáveis revelam *estados de tratamento* organizados em torno da intensidade da hemodiálise e do cuidado de suporte, podendo ser agrupados em *tratamento ativo*, *doença subtratada* e *vias associadas a complicações*. Interpretamos TCAV ≈ 0 como *protetivo* e TCAV ≈ 1 como *associado a risco*.

Tratamento ativo sob controle (protetivo). Diversos fatores com TCAV ≈ 0 correspondem a regimes de hemodiálise mais frequentes, como os Fatores 274 e 131. Esses fatores sugerem que os anos-paciente acima de certos limiares de diálise tendem a associar-se a menor risco previsto. Outros fatores protetivos incluem terapia de suporte: o Fator 94 combina hemodiálise muito frequente com uso de eritropoetina (`EVENT_c2MED_ERITRO`) e um limite para sevelâmer (`EVENT_c2MED_SEVEL`), enquanto o Fator 75 combina alta frequência de diálise com eritropoetina e medicação relacionada à calcificação (`EVENT_c2MED_CALCI`). Segundo os especialistas, esses padrões são compatíveis com trajetórias mais bem manejadas.

Monitoramento misto em pacientes não transplantados (alto risco). Em contraste, o Fator 225 (`EVENT_C1DIALISE_HD <= 1.5 AND EVENT_c2MED_ERITRO > 1.5 AND DIAGN_Z940 <= 0.5`) apresenta forte associação com risco. Esse padrão descreve anos-paciente com exposição mínima à hemodiálise, evidência de manejo de anemia e ausência de documentação de transplante (`Z94.0`), o que os especialistas consideraram plausível em estados mais graves de DRC, com complicações e baixa terapia substitutiva renal. O indicador de transplante aparece apenas como *ausência*, sugerindo que, nesta coorte, esse é o sinal mais saliente relacionado ao transplante.

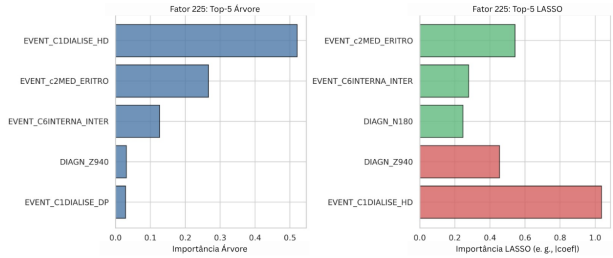

Figura 3. Fator 225: principais atributos segundo a árvore de decisão substituta (esquerda) e o readout linear esparso via LASSO (direita).

A Figura 3 resume a caracterização desse fator no espaço de entrada. Embora a árvore e o LASSO diferenciem seus ranqueamentos, ambos concordam quanto ao papel central de `EVENT_C1DIALISE_HD` e `EVENT_c2MED_ERITRO`. Os especialistas também observaram que a eritropoetina pode assumir um papel ambíguo: em contextos de baixa diálise, pode sinalizar complicações refratárias ou atraso na intensificação do cuidado.

Tabela 2. Fatores interpretáveis do SAE. Regras simplificadas extraídas por árvore de decisão, métricas de qualidade da regra (precisão/*recall*) e *score* TCAV médio para os fatores filtrados.

Fator	Condições simplificadas	Prec.	Rec.	TCAV
4	`10.5 < EVENT_C1DIALISE_HD <= 13.5 AND EVENT_c5TX_EXTX <= 0.5`	1.00	0.688	0.000
64	`0.5 < EVENT_C1DIALISE_HD <= 2.5 AND EVENT_c3ACESSO_CT <= 1.0`	1.00	0.487	0.001
75	`10.5 < EVENT_C1DIALISE_HD <= 12.5 AND EVENT_c2MED_ERITRO > 7.5 AND EVENT_c2MED_CALCI > 0.5`	1.00	0.277	0.124
83	`EVENT_C1DIALISE_HD > 11.5 AND EVENT_c5TX_EXTX <= 0.5 AND DIAGN_N189 <= 0.5 AND DIAGN_N180 > 11.5`	1.00	0.348	0.000
87	`EVENT_C1DIALISE_HD > 5.5`	1.00	0.869	1.000
94	`EVENT_C1DIALISE_HD > 13.5 AND EVENT_c2MED_ERITRO > 2.5 AND EVENT_c2MED_SEVEL <= 8.5`	1.00	0.580	0.067
131	`EVENT_C1DIALISE_HD > 15.5`	0.97	0.349	0.000
189	`EVENT_c3ACESSO_CT > 0.5 AND EVENT_c2MED_ERITRO <= 0.5 AND EVENT_C1DIALISE_HD <= 11.5 AND DIAGN_N180 > 1.5`	0.90	0.432	0.846
225	`EVENT_C1DIALISE_HD <= 1.5 AND EVENT_c2MED_ERITRO > 1.5 AND DIAGN_Z940 <= 0.5`	1.00	0.427	1.000
251	`EVENT_c3ACESSO_CT > 0.5 AND EVENT_C1DIALISE_HD > 5.5 AND EVENT_C6INTERNA_INTER > 0.5`	1.00	0.276	0.871
270	`0.5 < EVENT_C1DIALISE_HD <= 6.5`	1.00	0.682	1.000
274	`9.5 < EVENT_C1DIALISE_HD <= 12.5`	1.00	0.451	0.000
280	`0.5 < EVENT_C1DIALISE_HD <= 10.5`	1.00	0.504	1.000

Anos-paciente associados a complicações (alto risco). Outro grupo de fatores de alto risco reflete exposição intermediária à diálise e marcadores de cuidado agudo. O Fator 189 (`EVENT_c3ACESSO_CT > 0.5 AND EVENT_c2MED_ERITRO <= 0.5 AND EVENT_C1DIALISE_HD <= 11.5 AND DIAGN_N180 > 1.5`) está fortemente associado a risco e caracteriza um subgrupo com procedimentos de acesso, marcadores de estadiamento da DRC e baixa exposição à eritropoetina. Os médicos ressaltaram que acessos temporários ou por cateter se associam clinicamente a maior risco de infecção e a inícios não planejados de terapia. De forma consistente, outros fatores associados à exposição intermediária à diálise, como o Fator 270, também apresentam TCAV elevado, sugerindo trajetórias assistenciais instáveis. Em contraste, frequências muito elevadas de hemodiálise tendem a associar-se a conceitos protetivos.

Em conjunto, os fatores refinados sustentam a hipótese de que o modelo fundacional codifica *estados de tratamento* e *trajetórias de cuidado*, e não diagnósticos isolados.

5.4 Ablação Conceitual e Estabilidade Temporal

Para avaliar se os fenótipos descobertos desempenham um papel funcional nas predições, realizamos testes inspirados no ACE, baseados na remoção e no isolamento controlado de conceitos individuais. Como os conceitos são definidos no espaço interno de representação e o *Drift-Resilient TabPFN* depende de contexto, aplicamos as intervenções na interface de decisão do modelo. Para testar *necessidade*, removemos a contribuição do conceito subtraindo sua direção do embedding final. Para testar *suficiência*, preservamos apenas a componente alinhada a essa direção e suprimimos as demais. Em ambos os casos, medimos a mudança no risco previsto.

Em média, a remoção de um único conceito gerou pequenas alterações na predição ($|\Delta_{\text{destroy}}| < 0.02$), sugerindo que o modelo depende de múltiplos padrões em interação. Ainda assim, para alguns conceitos, o efeito foi substancialmente maior nas amostras em que esses conceitos estavam fortemente ativados, indicando relevância para subgrupos específicos. Nos testes de suficiência, a maioria dos conceitos não foi capaz de sustentar sozinha a predição, mas um pequeno subconjunto, como o Fator 4, preservou boa parte da predição original quando isolado.

Também investigamos a estabilidade desses efeitos ao longo do tempo. Segundo nosso critério de *drift*, apenas um número limitado de conceitos, como os Fatores 270 e 280, manteve comportamento consistente ao longo dos períodos. Muitos outros variaram substancialmente, possivelmente refletindo mudanças na população, nas práticas clínicas e nas convenções de codificação. Por isso, para fins de implantação, priorizamos conceitos temporalmente estáveis e tratamos conceitos sensíveis a *drift* com maior cautela e monitoramento longitudinal.

6 Conclusão

A implantação de Modelos Fundacionais Tabulares (TFMs) em ambientes críticos de saúde exige ir além das métricas estáticas de acurácia, demandando a auditoria sistemática de como esses modelos raciocinam sob condições reais e dinâmicas. Neste trabalho, introduzimos um arcabouço de ponta a ponta para **Auditoria Mecanística Dinâmica** que torna observáveis e clinicamente significativos os processos de decisão internos de TFMs, mesmo na presença de desvio temporal (*drift*).

Aplicado ao *Drift-Resilient TabPFN* em uma extensa coorte renal longitudinal, demonstramos que *Sparse Autoencoders* (SAEs) resolvem a sobreposição de representações de forma muito superior aos métodos lineares. Ao integrar árvores de decisão substitutas à adaptação temporal do TCAV, isolamos padrões de decisão internos, tanto estáveis quanto sensíveis ao tempo, comprovando a relevância funcional desses conceitos. A validação independente por nefrologistas especialistas atestou que nossa abordagem recupera conhecimento médico coerente e acionável em vez de correlações espúrias. Apesar dos resultados promissores, o trabalho possui limitações. Os conceitos descobertos dependem da representação interna aprendida pelo modelo e podem variar conforme arquitetura, janela temporal e distribuição populacional. Além disso, as intervenções realizadas no espaço latente não estabelecem causalidade clínica estrita. Trabalhos futuros focarão no direcionamento ativo (*model steering*) para corrigir conceitos em desvio sem a necessidade de retreinamento completo, consolidando uma base robusta para uma IA clínica transparente, mecanisticamente alinhada e confiável.

Declarações complementares

Agradecimentos

Este trabalho foi apoiado pela Fapemig, FAPESP, CNPq, CAPES. Também contou com o apoio do Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR) (bolsa # 408490/2024-1) e também do Centro de Inovação e Inteligência Artificial em Saúde (CI-IA Saúde), parcialmente apoiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) – Bolsa nº 2020/09866-4, Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG) – Bolsa nº PPE-00030-21, e UNIMED Belo Horizonte.

Contribuições dos autores

JMC contribuiu para a concepção, metodologia, desenvolvimento dos experimentos e redação do manuscrito. RMG e DTC contribuiram para desenvolvimento dos experimentos e redação do manuscrito. WMJ, MLC, HAR, LR e MAG contribuíram para a análise dos resultados e revisão crítica do manuscrito. Todos os autores leram e aprovaram o manuscrito final.

Conflitos de interesse

Os autores declaram que não têm nenhum conflito de interesses.

Disponibilidade de dados e materiais

Os códigos utilizados neste estudo estão disponíveis em: `https://github.com/joaomarcostomaz/TCAV`

Referências

Bricken, T. *et al*. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread.

Campos, J. M., Gomes, R. M., Chaves, D. T., Meira Jr., W., Cherchiglia, M. L., Rocha, H. A., Rocha, L., and Gonçalves, M. A. (2026). Mechanistic dynamic interpretability for tabular foundation models in healthcare. In *XAI World Conference 2026*. Accepted paper.

Cunningham, H. *et al*. (2023). Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Elhadri, H. *et al*. (2024). XNNTab: Interpretable neural networks for tabular data using sparse autoencoders. *arXiv preprint arXiv:2512.13442*.

Ghorbani, A. *et al*. (2019). Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32.

Guerra Junior, A. A. *et al*. (2018). Building the national database of health centred on the individual: administrative and epidemiological record linkage—brazil, 2000–2015. *International Journal of Population Data Science*, 3.

Helli, K. *et al*. (2024). Drift-resilient TabPFN: In-context learning temporal distribution shifts on tabular data. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Hollmann, N. *et al*. (2023). TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations (ICLR)*.

Kim, B. *et al*. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning (ICML)*, pages 2668–2677.

Marcolino, M. S. *et al*. (2025). Explainable artificial intelligence for predicting cardiovascular events in hospitalised COVID-19 patients. *BMC Infectious Diseases*, 25:1569.

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2 edition.

Paiva, B. *et al*. (2024). A new natural language processing–inspired methodology to investigate temporal drifts in health care data. *JMIR Medical Informatics*, 12:e54246.

Pendyala, S. *et al*. (2022). Concept-based explanations for tabular data. *arXiv preprint arXiv:2209.05690*.

Vaswani, A. *et al*. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.