Capítulo 4

Da Avaliação Estática à Auditoria Contínua: o Grande Desafio da IA Clínica em Dados em Evolução

Marcos André Gonçalves, Leonardo Rocha

Abstract

Current paradigms for evaluating clinical AI remain misaligned with real-world deployment, where healthcare systems evolve through shifts in populations, protocols, and data-generation processes. Models that perform well at deployment may degrade, become miscalibrated, or rely on unstable patterns over time. This paper argues for a shift from static evaluation to continuous auditing of clinical AI under evolving data conditions. We propose a holistic perspective combining strict temporal validation, longitudinal monitoring, drift diagnosis, continuous mitigation, and governance. Interpretability is positioned not as a standalone solution, but as a supporting layer within this broader reliability-oriented framework. We frame this transition as a grand challenge for digital health in Brazil, particularly given the scale, heterogeneity, and longitudinal nature of SUS data.

Resumo

Os paradigmas atuais de avaliação de IA clínica permanecem desalinhados das condições reais de implantação, nas quais os sistemas de saúde evoluem continuamente devido a mudanças nas populações, nos protocolos e nos processos de geração de dados. Modelos com bom desempenho inicial podem se degradar, perder calibração ou apoiar decisões com base em padrões instáveis ao longo do tempo. Este artigo defende a transição da avaliação estática para a auditoria contínua de sistemas de IA clínica com dados em evolução. Propomos uma perspectiva holística que combina validação temporal rigorosa, monitoramento longitudinal, diagnóstico de drifts, mitigação contínua e governança. A interpretabilidade é posicionada não como solução isolada, mas como uma camada de apoio em um arcabouço orientado à confiabilidade. Enquadramos essa transição como um grande desafio para a saúde digital no Brasil, especialmente diante dos dados heterogêneos, longitudinais e em larga escala do SUS.

 ©2026 SBC – Soc. Bras. de Computação

4.1 Desafio Central

A avaliação atual de sistemas de IA clínica, baseada em cenários estáticos, é insuficiente para garantir confiabilidade em ambientes de saúde em constante evolução. Modelos de IA clínica não falham apenas por baixa acurácia, mas também por perderem o alinhamento com a realidade de uso contínuo.

Desafio único proposto: estabelecer, para a próxima década da saúde digital no Brasil, um marco holístico e multidisciplinar de auditoria de sistemas de IA clínica em dados que evoluem, permitindo monitorar, validar, explicar, corrigir e governar continuamente o comportamento de modelos fundacionais em cenários reais de decisão clínica. Ao longo deste trabalho, nos referimos a esse paradigma como *auditoria contínua de IA clínica*, entendido como o conjunto de práticas de monitoramento, diagnóstico e governança de modelos ao longo do tempo em ambientes dinâmicos.

Nossa proposta não se limita a um problema técnico isolado nem a uma agenda incremental com resultados já consolidados. O foco está em uma lacuna estrutural: a ausência de práticas institucionais robustas para garantir que os modelos permaneçam confiáveis à medida que populações, protocolos, fluxos de cuidado e sistemas de registro se transformam ao longo do tempo.

Em termos objetivos, o desafio se desdobra em três perguntas centrais: (i) como auditar o desempenho, a confiabilidade e a estabilidade sob mudanças temporais? (ii) como detectar precocemente degradações com potencial impacto clínico? (iii) como definir mecanismos de mitigação e de governança capazes de preservar a segurança, a utilidade e a legitimidade social ao longo do ciclo de vida do modelo?

4.2 Contexto e Motivação

Modelos fundacionais tabulares, como TabPFN e TabICL, representam um avanço importante para tarefas preditivas em dados estruturados, ao combinar arquiteturas baseadas em *Transformers* com aprendizado em contexto [Vaswani et al. 2017, Hollmann et al. 2025, Qu et al. 2025]. Em paralelo, a literatura recente em aprendizado de dados tabulares vem ampliando a discussão sobre capacidade de generalização, robustez temporal e comparação com métodos clássicos [Gorishniy et al. 2021, Grinsztajn et al. 2022, Borisov et al. 2024, Jiang et al. 2025]. Ainda assim, grande parte das avaliações permanece ancorada em protocolos aproximadamente independentes e identicamente distribuídos (i.i.d.), pouco aderentes à realidade dos ambientes clínicos longitudinais.

Na prática, os sistemas de saúde mudam continuamente. Diretrizes terapêuticas são atualizadas, medicamentos e exames são incorporados, perfis demográficos se alteram, sistemas administrativos mudam de codificação e fluxos assistenciais são reorganizados. Esses fatores produzem diferentes formas de drift temporal e estrutural [Tsymbal 2004, Gama et al. 2014, Lu et al. 2019], com risco de degradação silenciosa em aplicações médicas [Nestor et al. 2019]. Um modelo pode preservar um desempenho médio aparentemente aceitável e, ainda assim, perder calibração, falhar em subgrupos específicos ou passar a explorar sinais contingentes, sem significado clínico estável. Em contextos clínicos, falhas decorrentes da degradação de modelos não são apenas técnicas, mas também podem impactar diretamente decisões assistenciais e desfechos em saúde.

 ©2026 SBC – Soc. Bras. de Computação

No contexto brasileiro, esse problema adquire uma relevância ainda maior. O SUS opera em âmbito nacional, com grande heterogeneidade regional, diversidade de infraestruturas de registro e bases administrativas e epidemiológicas de longa duração [Guerra Junior et al. 2018]. Nessas condições, a robustez temporal deixa de ser um refinamento metodológico e se constitui um requisito de segurança, governança e valor público.

Nesse contexto, o Brasil não apenas exemplifica os desafios descritos, mas também reúne condições particularmente favoráveis para liderar essa agenda em escala global. A combinação entre um sistema público de saúde de abrangência nacional, como o SUS, bases de dados administrativas e epidemiológicas de longa duração e elevada heterogeneidade regional cria um ambiente singular para o desenvolvimento, avaliação e governança de sistemas de IA clínica sob condições reais de mudança ao longo do tempo.

Diferentemente de cenários mais homogêneos ou fragmentados, o contexto brasileiro permite observar, de forma integrada, múltiplas formas de drift (i.e., demográfico, clínico, organizacional e de codificação) ao longo de diferentes regiões, períodos e níveis de atenção à saúde. Essa diversidade, frequentemente tratada como um obstáculo, pode ser reinterpretada como um ativo científico e tecnológico, ao possibilitar a construção de protocolos, métricas e infraestruturas de auditoria contínua de IA clínica que sejam robustos a variações reais de contexto.

Além disso, a escala e a natureza longitudinal dos dados disponíveis viabilizam estudos de estabilidade temporal, calibração dinâmica e análise longitudinal do comportamento dos modelos ao longo de anos ou décadas, o que aproxima a avaliação das condições efetivas de uso. Esse tipo de evidência é essencial para transformar a IA clínica de uma tecnologia experimental em uma infraestrutura confiável de suporte à decisão.

Assim, o enfrentamento desse desafio no contexto brasileiro não deve ser visto apenas como uma necessidade local, mas como uma oportunidade estratégica de posicionar o país como uma das principais referências internacionais em saúde digital confiável, responsável e orientada a valor público. Ao estruturar capacidades institucionais de auditoria contínua de IA clínica em larga escala, o Brasil pode não apenas adaptar soluções existentes, mas contribuir ativamente para a definição de novos padrões globais de avaliação, monitoramento e governança de IA em saúde.

4.3 Por que este é um Grande Desafio

A agenda de IA responsável tem enfatizado transparência, *fairness*, *accountability* e supervisão humana. No entanto, os mecanismos operacionais ainda são insuficientes para verificar se um sistema permanece clinicamente alinhado quando o ambiente em que opera muda. Por isso, defendemos uma transição conceitual e prática: sair de uma cultura de *benchmarking* estático e migrar para uma cultura de auditoria contínua.

Essa transição exige distinguir pelo menos dois fenômenos complementares. O primeiro é o **population drift**, associado a mudanças no perfil dos casos observados. O segundo é o **concept drift**, relacionado à mudança nas relações que sustentam as decisões do modelo [Tsymbal 2004, Gama et al. 2014, Lu et al. 2019]. Ambos são críticos. Entretanto, o segundo é especialmente preocupante, pois pode permanecer invisível em métricas agregadas enquanto altera silenciosamente a lógica de decisão do sistema.

 ©2026 SBC – Soc. Bras. de Computação

Nesse cenário, métodos de interpretabilidade podem desempenhar um papel importante, mas não devem ser confundidos com uma solução completa do problema. Técnicas como SHAP e explicações locais ajudam a entender quais variáveis influenciaram predições específicas [Lundberg and Lee 2017, Lundberg et al. 2020]. Abordagens mais recentes, incluindo métodos baseados em conceitos e auditorias temporais de representações internas, ampliam essa capacidade diagnóstica [Pelosi et al. 2025, Campos et al. 2026]. Ainda assim, o núcleo do desafio não é apenas explicar decisões pontuais, mas também garantir que o comportamento do sistema permaneça clinicamente plausível, estável e governável em horizontes temporais mais longos.

Superar esse desafio exige mais do que melhorias pontuais em desempenho preditivo. Exige estruturar capacidades técnicas, clínicas e institucionais que permitam acompanhar a confiabilidade da IA de forma longitudinal. Nesse sentido, cinco dimensões são particularmente relevantes: **robustez temporal**, para manter desempenho e calibração em múltiplas janelas; **detecção precoce de degradação**, para identificar sinais de instabilidade antes que produzam impacto clínico relevante; **capacidade diagnóstica**, para distinguir causas de degradação, como drift populacional, mudanças no processo assistencial, alterações na codificação ou fragilidade em subgrupos; **efetividade de mitigação**, para restaurar segurança e utilidade por meio de recalibração, atualização, restrição de uso e revalidação; e **utilidade para governança**, para transformar evidências técnicas em decisões rastreáveis e acionáveis por equipes clínicas, gestoras e regulatórias.

É justamente essa combinação de amplitude, profundidade técnica, impacto social e necessidade de coordenação multidisciplinar que caracteriza o tema como um grande desafio para a Computação Aplicada à Saúde no Brasil. A resposta a esse desafio não é um modelo melhor isoladamente, mas a institucionalização da *auditoria contínua de IA clínica* como prática sistemática em ambientes reais.

4.4 Estrutura Geral para Enfrentamento

Enfrentar esse desafio requer uma estrutura de *auditoria contínua de IA clínica*, organizada em três macrofrentes complementares.

(1) Monitoramento temporal. Acompanhar o desempenho preditivo, a calibração, o erro por subgrupos e a estabilidade de comportamento em janelas temporais sucessivas, com validação estrita por tempo e sem vazamento de informação futura. O objetivo não é apenas verificar se o modelo continua acertando, mas se permanece calibrado, consistente e clinicamente seguro em horizontes temporais mais longos.

(2) Diagnóstico de degradação. Investigar as causas de queda ou instabilidade, distinguindo drift populacional, mudanças no processo assistencial, alterações na codificação, dependência de variáveis transitórias e fragilidade em subgrupos específicos. Nessa etapa, análises por coorte, perturbações direcionadas, comparações entre períodos e técnicas de interpretabilidade desempenham papel central como instrumentos diagnósticos para localizar a origem das mudanças.

(3) Mitigação e governança contínuas. Definir protocolos graduais de resposta, incluindo recalibração, atualização do modelo, restrição de uso, ampliação da revisão humana, revalidação obrigatória e eventual descontinuação. Essas ações devem ser acom-

 ©2026 SBC – Soc. Bras. de Computação

panhadas de trilhas de decisão rastreáveis, com responsabilidades claramente distribuídas entre equipes técnicas, clínicas e gestoras. Isso implica tornar explícitos os mecanismos técnicos e as responsabilidades associadas às decisões mediadas por IA.

Como componente transversal, temos que técnicas de interpretabilidade, incluindo métodos baseados em conceitos, análise de suficiência e necessidade e inspeção de representações internas, podem aprofundar o diagnóstico sempre que agregarem valor à decisão clínica/regulatória [Lundberg and Lee 2017, Lundberg et al. 2020, Pelosi et al. 2025, Campos et al. 2026, Marcolino et al. 2025]. Nesse contexto, a interpretabilidade atua como suporte à *auditoria contínua de IA clínica*, especialmente na identificação de mudanças nos padrões de decisão ao longo do tempo.

Todas essas macrofrentes operacionalizam a *auditoria contínua de IA clínica* como uma prática sistemática, contínua e institucionalizável em ambientes de saúde dinâmicos.

4.5 Agenda para 2026–2035

Para tornar a *auditoria contínua de IA clínica* exequível em escala nacional, propomos cinco frentes de ação articuladas.

F1 – Padronização de validação temporal em IA clínica. Definir protocolos mínimos para o particionamento temporal, a avaliação prospectiva e retrospectiva, a análise por coortes e o reporte de estabilidade.

F2 – Infraestrutura de observabilidade e monitoramento contínuo. Desenvolver capacidades institucionais para acompanhar o desempenho, a calibração, o drift e o comportamento por subgrupos após a implantação.

F3 – Protocolos de mitigação orientados por risco e impacto clínico. Estabelecer critérios objetivos para recalibrar, atualizar, restringir ou descontinuar modelos, com níveis de resposta proporcionais ao risco.

F4 – *Benchmarking* multicêntrico com reprodutibilidade. Criar bases, tarefas e protocolos comuns que permitam comparar modelos em cenários reais de mudança temporal, com diversidade regional e institucional.

F5 – Integração entre pesquisa, operação hospitalar e regulação. Aproximar a comunidade científica, os serviços de saúde, os gestores públicos, as instâncias regulatórias e a sociedade civil na definição de requisitos, indicadores e responsabilidades.

Em conjunto, essas frentes deslocam a discussão de "qual modelo tem maior acurácia hoje" para "qual sistema permanece seguro, útil e confiável em operação contínua".

Mais do que propor direções, esse desafio exige mecanismos claros de avaliação de progresso. Nesse sentido, para que essas frentes possam ser acompanhadas e avaliadas de forma contínua, é importante explicitar critérios iniciais de sucesso. Entre eles, destacam-se: (i) a adoção sistemática de validação temporal em estudos de IA clínica; (ii) a definição e o reporte padronizado de métricas de estabilidade temporal e calibração longitudinal; (iii) a implementação de infraestruturas de monitoramento contínuo em ambientes reais de uso; (iv) o estabelecimento de benchmarks multicêntricos com cenários explícitos de drift; e (v) a consolidação de protocolos institucionais de mitigação e governança orientados por risco clínico.

 ©2026 SBC – Soc. Bras. de Computação

4.6 Ilustração de Potencial Clínico na Interpretação de Predições

Como ilustração concreta do desafio proposto, tomamos como estudo de caso o trabalho aceito na XAI World Conference 2026 [Campos et al. 2026]. O objetivo de trazer esse exemplo não é apenas destacar um avanço metodológico em interpretabilidade, mas mostrar como esse tipo de abordagem pode contribuir, na prática, para a auditoria de sistemas clínicos em dados dinâmicos.

O estudo foi conduzido em uma coorte longitudinal de doença renal crônica, derivada de um registro nacional construído pela integração de bases de dados do SUS, incluindo sistemas ambulatoriais, hospitalares e de mortalidade [Guerra Junior et al. 2018]. O recorte analisado abrange 19 anos (1997–2015), com aproximadamente 9,6 milhões de registros e 67.267 pacientes únicos, e reúne mais de 6.200 tipos de eventos clínico-administrativos. Essa escala aproxima o experimento do cenário real de implantação, no qual mudanças de codificação, prática assistencial e perfil populacional são esperadas, e não excepcionais. A modelagem foi estruturada em unidade paciente-ano (*patient-year*), com treino em janelas históricas e teste em janelas futuras, sem vazamento temporal. O desfecho analisado foi a ocorrência de óbito, totalizando 25.072 eventos no período, em um contexto de estratificação de risco com impacto direto sobre acompanhamento e priorização assistencial. Além das métricas usuais de predição, o estudo examinou a estabilidade das explicações ao longo dos anos, a coerência entre subgrupos e a sensibilidade às mudanças de contexto clínico-operacional.

O valor clínico do caso está menos em explicar um exemplo isolado e mais em permitir uma leitura longitudinal da lógica do modelo. Explicações locais, como as derivadas de SHAP, permanecem úteis para identificar variáveis influentes em predições individuais [Lundberg and Lee 2017, Lundberg et al. 2020]. Contudo, em ambientes dinâmicos, a pergunta mais importante deixa de ser apenas "qual variável pesou neste caso?" e passa a ser "o padrão que sustenta essa decisão continua clinicamente plausível e estável ao longo de seu ciclo de vida?".

Foi nesse ponto que a auditoria contínua de IA clínica agregou valor. A análise permitiu distinguir padrões mais estáveis de cuidado e de risco de sinais transitórios associados a trajetórias clínicas ou operacionais menos consistentes. Em termos práticos, isso se traduz em apoio a decisões de governança: manter o modelo em uso, recalibrá-lo, restringir sua aplicação a determinados contextos ou exigir nova validação antes de sua continuidade operacional.

No contexto do SUS, a implicação é direta. A auditoria longitudinal não deve ser tratada como etapa pontual de pesquisa, mas como capacidade permanente de gestão de risco algorítmico. Isso envolve definir indicadores mínimos de estabilidade temporal, gatilhos explícitos de intervenção e trilhas auditáveis de decisão entre equipes técnicas, clínicas e gestoras. Sob essa perspectiva, explicações do modelo deixam de ser apenas artefatos analíticos e passam a funcionar como evidências para decisões clínicas, operacionais e regulatórias.

Esse estudo ilustra, na prática, elementos centrais de uma abordagem de *auditoria contínua de IA clínica*, conectando monitoramento longitudinal, diagnóstico de degradação e decisões de governança.

 ©2026 SBC – Soc. Bras. de Computação

4.7 Conclusão

Em aderência direta à chamada do I Workshop de Grandes Desafios da Computação Aplicada à Saúde, esta proposta sustenta que o grande desafio da próxima década, no Brasil, não é apenas desenvolver modelos mais precisos, mas também construir capacidade institucional permanente de *auditoria de IA clínica* em uso real.

Trata-se de um desafio de natureza estrutural, de alto impacto social, científico e econômico. Ele articula inteligência artificial, ciência de dados, saúde digital, operação hospitalar, governança pública, regulação, ética e inovação. Seu enfrentamento exige mais do que avanços algorítmicos isolados: exige protocolos, infraestrutura, observabilidade, responsabilidades organizacionais e mecanismos de resposta proporcionais ao risco — elementos centrais de uma abordagem de *auditoria contínua de IA clínica*.

Nossa posição é direta. Não basta demonstrar acurácia em condições estáticas. É necessário comprovar confiabilidade em uso real de forma contínua, por meio de práticas sistemáticas de *auditoria contínua de IA clínica*, incluindo monitoramento contínuo, diagnóstico de degradação, protocolos de mitigação, responsabilização rastreável e capacidade de revisão ao longo de todo o ciclo de vida do sistema. Sem isso, a promessa da IA em saúde permanece tecnicamente promissora, mas institucionalmente frágil.

Se esse desafio for enfrentado de forma coordenada entre 2026 e 2035, o Brasil poderá se consolidar como uma das principais referências internacionais em sistemas de saúde digital orientados por valor público: mais seguros para pacientes, mais previsíveis para gestores, mais transparentes para profissionais e mais legítimos para a sociedade. Nesse cenário, modelos fundacionais deixam de ser apenas ferramentas de predição e passam a compor infraestrutura estratégica para prevenção, priorização assistencial e tomada de decisão em escala.

Em síntese, o verdadeiro grande desafio da Computação Aplicada à Saúde no país não é apenas construir modelos melhores, mas consolidar a *auditoria contínua de IA clínica* como capacidade institucional permanente de governança em dados em evolução. É essa capacidade que transforma desempenho pontual em confiança sustentável em sistemas de IA clínica.

Sobre os Autores

Marcos André Gonçalves possui graduação em Ciência da Computação pela Universidade Federal do Ceará (1995), mestrado pela Universidade Estadual de Campinas (1997), doutorado em Computer Science pela Virginia Tech, pós-doutorado pela UFMG (2006) e pelo Politecnico di Torino (2024), e atualmente é professor titular da UFMG. Atua em Recuperação de Informação, Aprendizado de Máquina e Processamento de Linguagem Natural, com mais de 400 artigos publicados, índice h 61 no Google Scholar, mais de 15.000 citações e presença no Ranking de Stanford entre os cientistas mais influentes. Recebeu diversos prêmios, incluindo o Prêmio CAPES de Tese (2024, 1º lugar como orientador; 2020, Menção Honrosa como coorientador), foi *General Chair* da ACM/IEEE JCDL 2018, integra *senior committees* de conferências como SIGIR, ACL, RecSys, CIKM, WSDM e ECIR, é *Senior Editor* do Journal of the Brazilian Computer Society, membro editorial do TACL, ex-Membro Afiliado da Academia Brasileira de

 ©2026 SBC – Soc. Bras. de Computação

Ciências, Bolsista de Produtividade CNPq 1A, ex-membro titular da CEX/FAPEMIG e Coordenador do Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional e Tratamento e Disseminação de Informação (INCT-TILDIAR).

Leonardo Rocha possui graduação (2003), mestrado (2005) e doutorado (2009) em Ciência da Computação pela Universidade Federal de Minas Gerais, com período de pós-doutorado na Ohio State University. É professor titular da UFSJ, com produção científica consistente em Recuperação de Informação, Sistemas de Recomendação e Aprendizado de Máquina, totalizando mais de 300 publicações na área, além de ampla experiência na formação de recursos humanos. É bolsista de produtividade do CNPq desde 2016 e atua também em temas como IA, Ciência de Dados, Deep Learning e Mineração de Dados aplicados à saúde. Atualmente, é Coordenador de Comunicação do INCT-TILDIAR.

Agradecimentos

Este trabalho foi apoiado por: CNPq, CAPES, Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR), FAPEMIG, AWS, Google, NVIDIA, CIIASaúde e FAPESP.

Referências

[Borisov et al. 2024] Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., and Kasneci, G. (2024). Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):7499–7519.

[Campos et al. 2026] 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.

[Gama et al. 2014] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37.

[Gorishniy et al. 2021] Gorishniy, Y., Rubachev, I., Khrulkov, V., and Babenko, A. (2021). Revisiting deep learning models for tabular data. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21.

[Grinsztajn et al. 2022] Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22.

[Guerra Junior et al. 2018] Guerra Junior, A. A., Pereira, R. G., Gurgel, E. I., Cherchiglia, M., Dias, L. V., Ávila, J. D., Santos, N., Reis, A., Acurcio, F. A., and Meira Junior, W. (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(1):446.

 ©2026 SBC – Soc. Bras. de Computação

[Hollmann et al. 2025] Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. (2025). Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326.

[Jiang et al. 2025] Jiang, J.-P., Liu, S.-Y., Cai, H.-R., Zhou, Q.-L., and Ye, H.-J. (2025). Representation Learning for Tabular Data: A Comprehensive Survey . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–20.

[Lu et al. 2019] Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., and Zhang, G. (2019). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12):2346–2363.

[Lundberg et al. 2020] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67.

[Lundberg and Lee 2017] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777.

[Marcolino et al. 2025] Marcolino, M. S., Schettini, I., do Nascimento, G., da Rocha, L., Lana, F., and Gonçalves, M. A. (2025). Explainable artificial intelligence for predicting cardiovascular events in hospitalised covid-19 patients. *BMC Infectious Diseases*, 25(1):1569.

[Nestor et al. 2019] Nestor, B., McDermott, M. B., Boag, W., Berner, G., Naumann, T., Hughes, M. C., Zajac, A., and Szolovits, P. (2019). Feature robustness in non-stationary health records: caveats to deployable model performance in rapidly changing clinical environments. In *Machine Learning for Healthcare Conference*, pages 114–150. PMLR.

[Pelosi et al. 2025] Pelosi, D., Cacciagrano, D., and Piangerelli, M. (2025). Explainability and interpretability in concept and data drift: A systematic literature review. *Algorithms*, 18:443.

[Qu et al. 2025] Qu, J., Holzmüller, D., Varoquaux, G., and Le Morvan, M. (2025). TabICL: A tabular foundation model for in-context learning on large data. *arXiv:2502.05564*.

[Tsymbal 2004] Tsymbal, A. (2004). The problem of concept drift: definitions and related work.

[Vaswani et al. 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *NeuriPS*, volume 30.

 ©2026 SBC – Soc. Bras. de Computação