

TRAUMA, CIRURGIA E MEDICINA INTENSIVA

EDIÇÃO XIII

Capítulo 36

APLICABILIDADE DO *MACHINE LEARNING* NA MEDICINA DIAGNÓSTICA: DA IMAGEM À DECISÃO CLÍNICA

LUÍS FELIPE GONÇALVES DE SOUZA[1]
JORDANNA PORTO INÁCIO[2]
ALEXANDRE PAIXÃO FRANCO[3]
NIVIA CAROLINE PORFÍRIO FERREIRA[4]
FERNANDA JOYCE RODRIGUES NOLETO VELOSO[5]
DANIEL LOPES DE OLIVEIRA[6]
FRANCIELE CARDOSO[7]

[1]Universidade de Rio Verde (campus Rio Verde); Especialização em Psiquiatria pela Faculdade Cenbrap. E-mail: dr.luisfgs@gmail.com.
[2]Médica; Centro Estadual de Reabilitação e Readaptação Dr. Henrique Santillo (CRER). E-mail: jordannaporto@hotmail.com.
[3]Graduando em Medicina; Universidade Tiradentes (UNIT). E-mail: alexandrepfranco@gmail.com.
[4]Graduada em Medicina; Centro Universitário de Várzea Grande (UNIVAG). E-mail: nniviacarolineporfiriof@gmail.com.
[5]Graduada em Medicina; ITPAC – Porto Nacional. E-mail: nandanoleto87@icloud.com.
[6]Graduado em Medicina; UNIFIMES, campus Trindade. E-mail: daniellopesoliveira001@gmail.com.
[7]Graduada em Medicina; Faculdade Morgana Potrich (FAMP). E-mail: bcfranciele@gmail.com.

Palavras-chave: *Aprendizado de Máquina; Diagnóstico Clínico; Inteligência Artificial.*

DOI 10.59290/2411329000

EDITORA PASTEUR

INTRODUÇÃO

A incorporação de recursos computacionais ao diagnóstico médico passou da automação de tarefas repetitivas para sistemas capazes de identificar padrões complexos em exames, sinais fisiológicos e dados clínicos. Entre essas tecnologias, o aprendizado de máquina (*machine learning*) utiliza exemplos para construir relações estatísticas entre informações de entrada e respostas de interesse. Na prática clínica, isso pode significar localizar uma alteração em uma imagem, estimar a probabilidade de determinada doença ou indicar que um exame merece avaliação prioritária. O ganho potencial não reside apenas na velocidade, mas na oportunidade de apoiar decisões em ambientes de grande volume assistencial.

A ampla disponibilidade de prontuários eletrônicos, equipamentos de imagem digital e métodos laboratoriais favoreceu o desenvolvimento desses modelos. Contudo, a capacidade de encontrar correlações em dados históricos não equivale à compreensão causal de uma doença nem garante utilidade à beira do leito. Um algoritmo pode apresentar desempenho expressivo no conjunto em que foi testado e falhar quando aplicado a outra população, aparelho ou fluxo assistencial. Por isso, a interpretação clínica exige distinguir validação técnica de benefício real ao paciente (COLLINS *et al.*, 2024; MOONS *et al.*, 2025).

No âmbito da Clínica Médica e da Cirurgia, o interesse é concreto: identificação de lesões em radiografias, tomografias e mamografias; apoio à análise de eletrocardiogramas; reconhecimento de padrões em lâminas anatomopatológicas; rastreamento de retinopatia diabética; e alertas associados à deterioração clínica. Este capítulo apresenta fundamentos, exemplos documentados, métricas de desempenho, limites éticos e critérios de implantação do aprendizado de máquina, enfatizando seu papel como recurso de apoio a decisões diagnósticas seguras.

APRENDIZADO DE MÁQUINA E RACIOCÍNIO DIAGNÓSTICO

O aprendizado supervisionado parte de registros identificados por um desfecho ou rótulo, como presença de neoplasia, resultado de biópsia ou diagnóstico confirmado. Um modelo é ajustado para relacionar atributos das observações a esses rótulos e, depois, estimar respostas em casos não utilizados no treinamento. Na aprendizagem não supervisionada, o sistema explora semelhanças, agrupamentos ou estruturas latentes sem depender da mesma rotulagem explícita. Há ainda métodos de aprendizagem por representação e abordagens que utilizam grande volume de dados não rotulados para obter características úteis a tarefas posteriores.

Entre as técnicas clássicas encontram-se regressões penalizadas, árvores de decisão, florestas aleatórias, métodos de gradiente e máquinas de vetores de suporte. O aprendizado profundo (*deep learning*) utiliza redes neurais com múltiplas camadas e tornou-se especialmente importante no processamento de imagens, sons e sinais biológicos. Modelos não devem ser selecionados apenas por sofisticação: um método estatístico simples, bem calibrado e interpretável pode funcionar melhor do que uma rede complexa quando a amostra é limitada, o evento é raro ou o fluxo de dados muda frequentemente (COLLINS *et al.*, 2024).

No diagnóstico, é essencial especificar a pergunta que será respondida. Identificar sinais de pneumonia em radiografia, estimar risco futuro de internação e sugerir hipótese diagnóstica com base em sintomas são três tarefas diferentes, com desfechos, padrões de referência e consequências clínicas distintas. Um

classificador não fornece automaticamente orientação terapêutica: mesmo diante de probabilidade elevada, o profissional deve avaliar prevalência, comorbidades, outras evidências e custo potencial de um falso alarme ou de um caso não reconhecido.

O treinamento costuma ser separado da avaliação em dados não vistos, idealmente com validação externa em hospitais, períodos ou populações independentes. Deve-se impedir vazamento de informação, como a presença de exames realizados após o diagnóstico dentro das variáveis destinadas a prever o diagnóstico anterior. Sem esse cuidado, o sistema pode aparentemente "antecipar" uma doença quando, na realidade, aprendeu pistas produzidas depois da decisão clínica. A literatura de avaliação de modelos destaca essas falhas como fontes relevantes de viés (MOONS *et al.*, 2025).

DA IMAGEM MÉDICA À DETECÇÃO DE ALTERAÇÕES

Radiologia e diagnóstico por imagem reúnem condições favoráveis ao aprendizado de máquina: dados digitais padronizáveis, tarefas visuais repetidas e grande quantidade de exames. Modelos podem sinalizar achados, segmentar estruturas, quantificar volumes e apoiar a classificação de lesões. Em radiografia de tórax, por exemplo, sistemas podem funcionar como segunda leitura ou triagem de exames com suspeita de alteração aguda. Na tomografia, a segmentação permite auxiliar medições de órgãos, hemorragias e lesões; na ressonância, pode contribuir para análise volumétrica e processamento da imagem. A utilidade de cada aplicação depende da indicação específica validada, e não de uma capacidade genérica de "interpretar exames".

A mamografia oferece um dos exemplos mais consistentes de avaliação clínica. O ensaio MASAI, conduzido em programa populacional de rastreamento na Suécia, comparou leitura apoiada por inteligência artificial com leitura dupla convencional. Uma análise publicada em 2023 avaliou a segurança e a carga de trabalho; avaliação posterior descreveu maior detecção de câncer no grupo com apoio de IA em relação ao controle. Na publicação de 2026, a sensibilidade foi de 80,5% no grupo com IA e de 73,8% no grupo controle, com especificidade de 98,5% em ambos, e a taxa de câncer de intervalo atendeu ao critério predefinido de não inferioridade (LÅNG *et al.*, 2023; GOMMERS *et al.*, 2026).

Esses resultados não autorizam afirmar que qualquer software de mamografia produzirá o mesmo benefício. O ensaio examinou um protocolo determinado, em serviço estruturado, com sistema e equipe definidos. A extrapolação para populações de diferentes idades, padrões de risco, disponibilidade de exames e programas nacionais exige avaliação própria. Resultados anteriores de McKinney *et al.* (2020) já haviam demonstrado desempenho promissor em bases internacionais, mas o valor adicional dos estudos prospectivos é observar o funcionamento de um sistema inserido no rastreamento real.

No uso rotineiro de tomografias de urgência, outra hipótese relevante é ordenar a fila de leitura para que exames com suspeita de achados tempo-dependentes sejam vistos mais rapidamente. Trata-se de priorização, não de confirmação diagnóstica independente. Uma classificação errada pode atrasar um caso urgente ou criar interrupções desnecessárias. A implantação precisa medir tempos de atendimento, consequências dos alertas, proporção de exames não classificados e capacidade de manter avaliação humana inclusive quando o sistema estiver indisponível.

Imagens adquiridas em diferentes fabricantes, hospitais ou protocolos podem apresentar

diferenças que afetam as previsões. Artefatos, compressão, movimentação, contraste e alterações epidemiológicas também modificam as condições de uso. A validação externa e o monitoramento periódico são indispensáveis para detectar perda de desempenho depois da implantação (SOUNDERAJAH *et al.*, 2025) (**Tabela 36.1**).

Tabela 36.1 Exemplos de aplicação e uso pretendido no diagnóstico

Área	Uso potencial	Limite clínico
Radiologia	Sinalizar achados e organizar leitura	Confirmar imagem, protocolo e contexto
Cardiologia	Apoiar análise de ECG e medidas	Não substituir monitorização e exame clínico
Patologia	Destacar regiões de lâminas	Validar amostra, coloração e laudo
Clínica hospitalar	Alertar para possível sepse	Interpretar sinais e agir conforme protocolos

Fonte: elaboração dos autores a partir das referências apresentadas neste capítulo.

CARDIOLOGIA E INTERPRETAÇÃO DE SINAIS FISIOLÓGICOS

Os eletrocardiogramas constituem sinais estruturados, amplamente disponíveis, com aplicações promissoras. Redes neurais podem aprender características da morfologia das ondas e identificar padrões que não são evidentes à leitura convencional. Em estudo retrospectivo, Attia *et al.* (2019) avaliaram um algoritmo capaz de reconhecer, em eletrocardiogramas registrados durante ritmo sinusal, sinais associados à história de fibrilação atrial. Esse tipo de ferramenta é potencialmente útil para selecionar pacientes que merecem investigação complementar, mas não comprova por si só uma arritmia ativa nem substitui monitorização eletrocardiográfica quando indicada.

A interpretação médica continua dependente de contexto. Em um paciente com palpitações, síncope, dispneia ou sinais de instabilidade hemodinâmica, a decisão sobre atendimento imediato não pode ser adiada em razão de uma pontuação algorítmica. Da mesma forma, a classificação de anomalias no ECG depende de calibração, qualidade dos eletrodos, ritmo, uso de medicamentos e condições clínicas que podem distorcer o sinal. O sistema pode ajudar na triagem e no reconhecimento de alterações, mas não se deve confundir rastreamento, diagnóstico confirmado e definição de tratamento.

Em ecocardiografia e outros exames cardiovasculares, o aprendizado de máquina pode facilitar contorno de câmaras, estimativas quantitativas e padronização de medidas. A oportunidade é reduzir variabilidade e agilizar tarefas; o risco é transmitir segurança indevida quando a janela acústica é inadequada ou existe anatomia fora do padrão. A análise clínica requer revisão das imagens relevantes e coerência entre a medida produzida, a observação do examinador e os sintomas do paciente.

ONCOLOGIA E ANATOMIA PATOLÓGICA DIGITAL

A digitalização de lâminas histológicas permite treinar modelos capazes de localizar regiões suspeitas e reconhecer padrões celulares associados a tumores. Na patologia, sistemas de apoio podem destacar campos para revisão, estimar características morfológicas e auxiliar no controle de qualidade. Campanella *et al.* (2019) estudaram redes profundas aplicadas a lâminas digitalizadas completas, utilizando supervisão relativamente fraca, isto é, diagnós-

ticos associados às lâminas sem exigir anotação minuciosa de cada área pelo patologista.

Em biópsias de próstata, Ström *et al.* (2020) demonstraram a viabilidade de um sistema para detecção e classificação histopatológica, com potencial de reduzir tarefas repetitivas e variabilidade de graduação. Entretanto, o padrão diagnóstico exige correlação com qualidade da amostra, contexto clínico, critérios anatomopatológicos e, quando necessário, exames complementares. Dados obtidos em um laboratório não asseguram desempenho idêntico em outro, especialmente se houver diferenças de coloração, preparação, escaneamento ou prevalência das lesões.

Na oncologia, a distinção entre apoiar o patologista e substituir a decisão médica é particularmente importante: classificações podem influenciar procedimentos invasivos, estadiamento e terapias de alto risco. Por esse motivo, qualquer incorporação deve definir quais classes de achados serão encaminhadas para revisão humana, como serão registradas discordâncias e qual profissional responde pelo laudo final. A tecnologia amplia a capacidade de inspecionar padrões, mas não elimina a responsabilidade pelo diagnóstico.

OFTALMOLOGIA E RASTREAMENTO DE RETINOPATIA DIABÉTICA

O rastreamento de retinopatia diabética ilustra como uma tarefa delimitada pode alcançar validação robusta fora de serviços de alta complexidade. No estudo pivotal de Abràmoff *et al.* (2018), um sistema autônomo foi utilizado em ambientes de atenção primária para identificar retinopatia diabética acima de determinado limiar. O estudo relatou sensibilidade de 87,2% e especificidade de 90,7% segundo a definição adotada e demonstrou a importância de um protocolo de aquisição de imagem e treinamento do operador.

A expressão "autônomo" nessa aplicação deve ser compreendida em seu escopo regulatório específico. Ela não significa que algoritmos possam substituir integralmente consultas oftalmológicas nem avaliar indiscriminadamente qualquer doença ocular. A captação de imagens inadequadas, pacientes com sintomas fora da tarefa do sistema e outras alterações que exigem avaliação clínica precisam de fluxos de encaminhamento. Em atenção primária, o benefício depende tanto de reconhecer pessoas em risco quanto de garantir acesso ao serviço especializado após o alerta.

RECONHECIMENTO PRECOCE DA DETERIORAÇÃO E DAS INFECÇÕES GRAVES

Em unidades de pronto atendimento e enfermarias, sinais vitais, exames laboratoriais e informações evolutivas podem alimentar sistemas que estimam risco de deterioração. Modelos para sepse e outras condições agudas tentam sinalizar combinações de dados que merecem reavaliação da equipe. Em estudo prospectivo multicêntrico, Adams *et al.* (2022) descreveram a implantação do TREWS, um sistema de alerta precoce para sepse, e observaram associações entre confirmação oportuna do alerta e melhores desfechos. O desenho observacional, entretanto, não permite afirmar que o algoritmo isoladamente tenha causado toda a diferença clínica relatada.

O exemplo evidencia que a utilidade depende de uma cadeia assistencial: alerta oportuno, interpretação do profissional, exame do paciente, coleta apropriada de amostras e adoção de condutas em tempo adequado. Alertas frequentes ou mal calibrados podem provocar fadiga, consumo excessivo de recursos e desconfiança. Em serviços com menor infraestru-

tura, sistemas dependentes de exames raramente solicitados podem ser pouco úteis. A pergunta relevante não é apenas "o modelo detectou a sepse?", mas "o fluxo com o modelo melhorou o reconhecimento e o cuidado sem aumentar danos?".

Na cirurgia e no período perioperatório, ferramentas semelhantes podem apoiar a identificação de pacientes em risco de complicações, sem substituir avaliação anestésica, exame físico e reavaliação contínua. A classificação preditiva deve ser diferenciada do diagnóstico de complicação já estabelecida; em situações críticas, os protocolos assistenciais e a avaliação imediata continuam prioritários.

EXAMES LABORATORIAIS, PRONTUÁRIOS E INTEGRAÇÃO DE DADOS

Os dados laboratoriais oferecem variáveis quantitativas passíveis de análise longitudinal: hemograma, biomarcadores cardíacos, função renal, eletrólitos e marcadores inflamatórios. A aprendizagem de máquina pode explorar tendências e relações entre medições sucessivas para apoiar rastreamento de alterações ou estimativas de risco. Entretanto, a probabilidade produzida não substitui limites analíticos, controle de qualidade, intervalo de referência e interpretação fisiopatológica. Exame realizado em tempo inadequado ou amostra contaminada pode induzir erro em qualquer modelo, por mais sofisticado que seja.

Prontuários eletrônicos combinam códigos diagnósticos, prescrições, exames e registros narrativos. Essas bases são úteis, mas não constituem automaticamente a verdade clínica: registros incompletos, códigos inseridos para faturamento, diferenças de linguagem e dados ausentes podem criar associações artificiais. Sistemas capazes de processar texto devem ser avaliados quanto à capacidade de recuperar corretamente achados, contexto temporal e negativas; extrair "sem evidência de pneumonia" como "pneumonia presente" é uma falha de interpretação com consequências diagnósticas. As orientações da Organização Mundial da Saúde destacam problemas adicionais de confiabilidade e privacidade em modelos multimodais de propósito amplo (OMS, 2024).

AVALIAÇÃO DE DESEMPENHO: ALÉM DA ACURÁCIA

Um sistema com alta acurácia global pode ser inadequado para uma doença rara ou para um cenário em que a consequência de perder casos é grave. Sensibilidade corresponde à proporção de pessoas com a condição corretamente identificadas; especificidade, à proporção das pessoas sem a condição corretamente classificadas. Valores preditivos positivo e negativo dependem da prevalência da condição na população examinada. Assim, a mesma sensibilidade e especificidade podem gerar diferentes proporções de falsos alarmes em serviços de rastreamento e unidades especializadas.

A área sob a curva ROC descreve a capacidade de discriminar, em média, indivíduos com e sem um desfecho ao longo de limiares, mas não informa sozinha quantos pacientes sofrerão erros em um ponto de decisão. A calibração avalia se probabilidades previstas correspondem às frequências observadas. Por exemplo, quando um modelo atribui 30% de risco a um grupo, espera-se, em condições ideais, que aproximadamente 30% tenham o evento previsto. O limiar de ação deve considerar consequências de falso positivo e falso negativo, capacidade de confirmação e recursos disponíveis.

A apresentação transparente de amostra, distribuição de doença, referência diagnóstica, dados faltantes, critérios de exclusão e desempenho por subgrupo é indispensável. O

STARD-AI, publicado em 2025, orienta relatos de acurácia diagnóstica com inteligência artificial; o TRIPOD+AI, de 2024, orienta o relato de modelos de predição clínica, e o PROBAST+AI, de 2025, apoia a avaliação de qualidade, risco de viés e aplicabilidade (SOUNDERAJAH *et al.*, 2025; COLLINS *et al.*, 2024; MOONS *et al.*, 2025).

Quando o sistema se integra à assistência, também devem ser medidos tempo até a interpretação, necessidade de exames adicionais, sobrecarga dos profissionais, diferenças entre grupos de pacientes e impacto sobre desfechos clinicamente importantes. A simples superioridade estatística em um conjunto retrospectivo não comprova redução de mortalidade, internações ou procedimentos desnecessários. A literatura de ensaios com IA destaca a necessidade de descrever precisamente como os resultados do algoritmo influenciaram as decisões humanas (LIU *et al.*, 2020).

LIMITAÇÕES, GENERALIZAÇÃO E INIQUIDADES

A composição do banco de treinamento determina quais padrões o sistema tem oportunidade de aprender. Se determinada faixa etária, condição clínica, grupo populacional ou tipo de equipamento estiver pouco representado, as previsões podem ser menos confiáveis nesses contextos. Também é possível que um modelo aprenda marcadores indiretos do hospital, da qualidade da imagem ou do processo de aquisição, em vez das características da doença. Essas associações podem desaparecer quando a ferramenta é transferida para outro serviço.

Os vieses podem surgir na seleção de pacientes, na confirmação diagnóstica desigual entre grupos e na ausência de resultados negativos. Por exemplo, se apenas pessoas encaminhadas a exames confirmatórios recebem rótulos finais, a base de treinamento poderá superrepresentar casos graves. A avaliação deve contemplar desempenho por idade, sexo e outras características clinicamente pertinentes, respeitando os limites éticos e legais aplicáveis ao uso de dados sensíveis. Não basta afirmar que o modelo é "imparcial": desigualdades precisam ser investigadas empiricamente (MOONS *et al.*, 2025; OMS, 2021).

Outro risco é a mudança de distribuição ao longo do tempo, denominada dataset shift. Novos aparelhos, técnicas de aquisição, mudanças na prevalência de doenças e alterações de protocolo podem reduzir a confiabilidade do sistema. Processos de governança devem incluir monitoramento, investigação de falhas, versionamento do algoritmo, mecanismos de suspensão e revalidação antes de mudanças substanciais. Sem essas medidas, mesmo uma ferramenta adequadamente validada no lançamento pode se tornar inadequada.

INTERPRETABILIDADE, SUPERVISÃO PROFISSIONAL E SEGURANÇA

Os resultados algorítmicos devem ser exibidos de modo compatível com a decisão clínica que pretendem apoiar. Um aviso de "alto risco" sem descrever a doença-alvo, a população estudada, o intervalo temporal e a ação recomendada pode produzir interpretações equivocadas. Mapas de atenção e explicações visuais podem auxiliar revisão, mas não demonstram necessariamente que o sistema utilizou relações causais corretas. A interpretabilidade deve servir à verificação e ao uso seguro, não apenas à aparência de transparência.

Em serviços assistenciais, a supervisão humana precisa ser concreta: profissionais treinados, acesso aos exames originais, possibilidade de discordar da sugestão e regras de escalonamento quando o algoritmo falha. A automa-

ção pode produzir dois efeitos opostos: confiança excessiva em uma resposta incorreta ou desconsideração de alertas relevantes devido a experiências anteriores de baixa qualidade. Protocolos de implantação devem investigar ambos os comportamentos e acompanhar consequências para o paciente.

Ferramentas gerativas capazes de redigir laudos e respostas clínicas merecem atenção adicional. Fluência textual não prova fidelidade aos dados. Sistemas podem omitir informações decisivas, inventar achados ou produzir recomendações incompatíveis com o caso. A OMS recomenda governança, responsabilidade e avaliações apropriadas ao contexto de aplicação de modelos multimodais na saúde (OMS, 2024). O uso seguro exige que o médico revise a informação original, cheque a correspondência dos achados e não confunda texto automatizado com interpretação independente validada.

ÉTICA, PROTEÇÃO DE DADOS E REGULAÇÃO

Dados de saúde envolvem informações sensíveis, frequentemente longitudinalmente associadas a pessoas identificáveis. O armazenamento, o compartilhamento com fornecedores, a reutilização para treinamento e a transmissão entre serviços exigem base legal apropriada, segurança da informação e políticas claras de acesso. A proteção do paciente requer avaliar o ciclo completo de dados, inclusive registros inseridos por profissionais em ferramentas externas. A Organização Mundial da Saúde recomenda que os sistemas respeitem autonomia humana, transparência, responsabilidade, inclusão e sustentabilidade (OMS, 2021).

No Brasil, softwares com finalidade médica podem estar sujeitos à regularização sanitária conforme sua classificação e função pretendida. A Resolução da Diretoria Colegiada da Anvisa nº 657/2022 trata da regularização de software como dispositivo médico (SaMD), distinguindo aplicações com finalidade clínica de sistemas meramente administrativos e de determinadas outras categorias excluídas. A conformidade regulatória deve ser examinada para cada produto, e não presumida apenas porque utiliza inteligência artificial (BRASIL, 2022).

Nos Estados Unidos, a *Food and Drug Administration* mantém lista pública de dispositivos médicos habilitados por IA autorizados para comercialização, útil como referência de categorias e usos declarados. A presença na lista não equivale à aprovação automática em outros países nem garante validade para indicações diferentes das registradas. Boas práticas de aprendizado de máquina defendem desenvolvimento multidisciplinar, dados representativos, avaliação humana e monitoramento durante todo o ciclo de vida da tecnologia (FDA, 2026).

INTEGRAÇÃO AO FLUXO CLÍNICO E CRITÉRIOS PARA ADOÇÃO

A implementação começa pela identificação de um problema assistencial concreto: excesso de tempo para laudar exames urgentes, variabilidade de triagem, baixa cobertura de rastreamento ou dificuldade de detectar deterioração. Depois, devem ser definidos público-alvo, exame ou sinal utilizado, ação acionada pelo alerta, responsável pela decisão e indicador de benefício. Adquirir software sem um fluxo clínico definido tende a gerar notificações pouco úteis e custo adicional.

O serviço deve comparar o desempenho local do modelo à prática existente, com análise prospectiva quando apropriada. Durante uma fase inicial, pode ser empregado acompanhamento silencioso das previsões para investigar discordâncias sem alterar a assistência. Após testes de segurança, a equipe deve ser capaci-

tada para conhecer limites, situações de não utilização e resposta a problemas técnicos. Avaliações periódicas precisam medir não apenas métricas algorítmicas, mas resultados relevantes ao paciente, como atrasos, eventos adversos e exames redundantes.

A adoção em hospitais menores e na rede pública requer ponderar conectividade, disponibilidade de equipamentos, custos de manutenção, qualificação profissional e acesso ao cuidado subsequente. Um sistema que identifica provável doença, mas não dispõe de caminho para confirmação e tratamento, pode aumentar ansiedade e desigualdade sem melhorar desfechos. Por outro lado, ferramentas cuidadosamente integradas podem ampliar rastreamento e reduzir atrasos em territórios com menor oferta de especialistas, desde que acompanhadas de rede de referência e avaliação de equidade.

PERSPECTIVAS PARA A PRÁTICA CLÍNICA

Os avanços em modelos multimodais permitem combinar imagem, sinais biológicos e texto clínico em uma única arquitetura, ampliando possibilidades de apoio longitudinal ao diagnóstico. Essa integração também aumenta a necessidade de validar cada função e de testar como o profissional interpreta suas respostas. Não há fundamento para considerar que uma ferramenta generalista ofereça automaticamente desempenho superior ao de sistemas especializados, sobretudo em doenças raras, populações sub-representadas ou situações de urgência.

O aprendizado de máquina já demonstrou aplicabilidade em tarefas delimitadas, como rastreamento de retinopatia diabética, apoio à leitura de mamografias, análise de ECG e patologia digital. O avanço com maior relevância clínica, entretanto, não será medido apenas pela quantidade de algoritmos disponíveis, mas pela capacidade de demonstrar benefícios consistentes, evitar erros e inserir a tecnologia nos fluxos assistenciais. A relação entre evidência técnica, decisão médica e contexto do paciente deve permanecer central.

No cenário atual, o uso responsável pode ser entendido como cooperação entre métodos computacionais e julgamento profissional. O sistema identifica padrões e organiza informações; a equipe clínica confirma hipóteses, interpreta consequências e assume as decisões. Com validação externa, governança de dados, monitoramento pós-implantação e rotas claras de revisão humana, o machine learning pode contribuir para diagnósticos mais oportunos e consistentes, preservando a segurança como condição de seu emprego.

REFERÊNCIAS BIBLIOGRÁFICAS

ABRÀMOFF, M. D. *et al*. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digital Medicine, v. 1, art. 39, p. 1-8, 2018. DOI: https://doi.org/10.1038/s41746-018-0040-6.

ADAMS, R. *et al*. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nature Medicine, v. 28, n. 7, p. 1455-1460, 2022. DOI: https://doi.org/10.1038/s41591-022-01894-0.

ATTIA, Z. I. *et al*. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm. The Lancet, v. 394, n. 10201, p. 861-867, 2019. DOI: https://doi.org/10.1016/S0140-6736(19)31721-0.

BRASIL. Agência Nacional de Vigilância Sanitária. Resolução RDC nº 657, de 24 de março de 2022. Dispõe sobre a regularização de software como dispositivo médico (SaMD). Diário Oficial da União, Brasília, DF, 25 mar. 2022.

CAMPANELLA, G. *et al*. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine, v. 25, n. 8, p. 1301-1309, 2019. DOI: https://doi.org/10.1038/s41591-019-0508-1.

COLLINS, G. S. *et al*. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, v. 385, art. e078378, p. 1-11, 2024. DOI: https://doi.org/10.1136/bmj-2023-078378.

FDA. U.S. Food and Drug Administration. List of Artificial Intelligence-Enabled Medical Devices. Silver Spring: FDA, [s. d.]. Disponível em: https://www.fda.gov/medical-devices/artificial-intelligence-enabled-medical-devices/list-artificial-intelligence-enabled-medical-devices. Acesso em: 29 out. 2026.

GOMMERS, J. *et al*. Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study. The Lancet, v. 407, n. 10526, p. 505-514, 2026. DOI: https://doi.org/10.1016/S0140-6736(25)02464-X.

LÅNG, K. *et al*. Artificial intelligence-supported screen reading versus standard double reading in the MASAI trial: a clinical safety analysis. The Lancet Oncology, v. 24, n. 8, p. 936-944, 2023. DOI: https://doi.org/10.1016/S1470-2045(23)00298-X.

LIU, X. *et al*. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nature Medicine, v. 26, n. 9, p. 1364-1374, 2020. DOI: https://doi.org/10.1038/s41591-020-1034-x.

MCKINNEY, S. M. *et al*. International evaluation of an AI system for breast cancer screening. Nature, v. 577, n. 7788, p. 89-94, 2020. DOI: https://doi.org/10.1038/s41586-019-1799-6.

MOONS, K. G. M. *et al*. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ, v. 388, art. e082505, p. 1-10, 2025. DOI: https://doi.org/10.1136/bmj-2024-082505.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization, 2021. ISBN 9789240029200.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva: World Health Organization, 2024. ISBN 9789240084759.

SOUNDERAJAH, V. *et al*. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nature Medicine, v. 31, n. 10, p. 3283-3289, 2025. DOI: https://doi.org/10.1038/s41591-025-03953-8.

STRÖM, P. *et al*. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. The Lancet Oncology, v. 21, n. 2, p. 222-232, 2020. DOI: https://doi.org/10.1016/S1470-2045(19)30738-7.