Avaliação do modelo de aprendizagem de máquina em aplicações de saúde: Uma revisão

https://doi.org/10.56238/interdiinovationscrese-013

Cezar Miranda Paula de Souza
Universidade Federal do Rio Grande do Norte, Brasil
ORCID: https://orcid.org/0009-0005-7189-8115
E-mail: cezarmiranda@gmail.com

Cephas Alves da Silveira Barreto
Universidade Federal do Rio Grande do Norte, Brasil
ORCID: https://orcid.org/0000-0002-4756-8571
E-mail: cephasax@gmail.com

Lhayana Vieira de Macedo
Universidade Federal do Rio Grande do Norte, Brasil
ORCID: https://orcid.org/0009-0009-0509-0555
E-mail: lhayana11@gmail.com

Bruna Alice Oliveira de Brito
Universidade Federal do Rio Grande do Norte, Brasil
ORCID: https://orcid.org/0009-0001-8116-495X
E-mail: brna.oliveira03@gmail.com

Victor Vieira Targino
Universidade Federal do Rio Grande do Norte, Brasil
ORCID: https://orcid.org/0000-0002-9036-6537
E-mail: victorvieira.rn@gmail.com

Emanuel Costa Betcel
Universidade Federal do Rio Grande do Norte, Brasil
ORCID: https://orcid.org/0009-0009-6814-4311
E-mail: emanuelbetcel@gmail.com

Fernando Gomes de Almeida
Universidade Federal do Rio Grande do Norte, Brasil
ORCID: https://orcid.org/0009-0006-2185-6969
E-mail: fernandogdalmeida@gmail.com

Arthur Andrade Galvíncio Rodrigues
Universidade Federal do Rio Grande do Norte, Brasil
ORCID: https://orcid.org/0009-0002-7107-742X
E-mail: arthurgalvincio.br@gmail.com

Ramon Santos Malaquias
Universidade Federal do Rio Grande do Norte, Brasil
ORCID: https://orcid.org/0000-0002-8350-2836
E-mail: ramonstmalaquias@gmail.com

Itamir de Morais Barroca Filho
Universidade Federal do Rio Grande do Norte, Brasil
ORCID: https://orcid.org/0000-0003-1694-8237
E-mail: itamir.filho@imd.ufrn.br

RESUMO

Os modelos de Aprendizado de Máquina (AM) têm sido aplicados para resolver problemas em diversos contextos, o que necessariamente envolve a avaliação adequada dos modelos para garantir seu desempenho. Uma vez implantados, os modelos de AM estão sujeitos a problemas de desempenho, como aqueles relacionados a mudanças nos dados (drift). Esse tipo de problema tem motivado esforços na análise e manutenção de modelos, bem como no aprendizado contínuo, que busca a capacidade de aprender continuamente a partir de um fluxo (contínuo) de dados. Portanto, é importante entender e desenvolver metodologias que possam ser utilizadas para avaliar modelos de AM, tornando seu uso em ambientes do mundo real viável. Entre as áreas atuais de aplicação de AM, uma que se destaca, em particular, é o Aprendizado de Máquina para a área da saúde, especialmente em conjunto com Software para Suporte à Decisão em Aplicações Médicas, apresentando desafios específicos para a avaliação e monitoramento de modelos, especialmente considerando que previsões ou classificações incorretas podem levar a situações que ameaçam a vida. Este artigo apresenta uma revisão sistemática da literatura cujo objetivo é identificar técnicas atuais para avaliar e manter modelos de AM aplicados a área da saúde em uso efetivo no mundo real.

Palavras-chave: Validação de modelos de AM, AM para a área da saúde, Monitoramento de modelos de AM.

1 INTRODUÇÃO

Recentemente, a Inteligência Artificial (IA) consolidou-se como uma das principais alternativas para a solução de problemas complexos em qualquer área do conhecimento. Tem se tornado cada vez mais comum ouvir falar ou mesmo encontrar sistemas que fazem uso de técnicas de IA (por exemplo, Aprendizado de Máquina, Expert Systems, Deep Learning, entre outros) para resolver problemas do dia a dia.

A saúde, área de alto impacto social, tem sido objeto de diversos estudos que utilizam técnicas de Aprendizado de Máquina (AM) para resolver problemas. Alguns estudos, por exemplo, aplicaram técnicas de AM para prever os resultados dos pacientes durante a pandemia COVID-19 (Malki et al., 2021; Arowolo et al., 2022). Outros tentaram prever o risco de morte para pacientes de UTI com insuficiência cardíaca (Luo et al., 2022). Dada a gravidade das questões abordadas, o uso de técnicas de AM em aplicações de saúde enfrenta particularmente desafios de modelagem, análise e validação (Ghassemi et al., 2020). Resolvê-los requer uma estreita colaboração entre cientistas de dados e especialistas em saúde para garantir que os modelos de AM sejam projetados para resolver problemas reais no campo e sejam interpretáveis e explicáveis para a comunidade clínica.

Os resultados e o desempenho dos modelos de AM estão intimamente relacionados aos dados usados para treiná-los e testá-los (Gopal, 2019). Portanto, torna-se difícil generalizar os resultados obtidos com dados de locais específicos e características dos pacientes para outros que não aqueles. Outro aspecto que dificulta a análise e validação dos resultados dos modelos em aplicações assistenciais é a necessidade de monitoramento contínuo e feedback de especialistas, o que é difícil de incorporar devido ao cotidiano exigente dos profissionais dos serviços de saúde. A análise estatística tradicional dos resultados pode não ser tão eficiente quando se trata de modelos em produção e aplicados a situações que podem significar vida ou morte para os pacientes, delimitando a necessidade de pesquisa e desenvolvimento em modelos de avaliação e monitoramento de AM para aplicações em saúde.

Com base nesse contexto, é possível afirmar que estudos relacionados à avaliação e manutenção de modelos de AM aplicados à saúde são de grande relevância. Apesar disso, a literatura da área não apresenta muitos trabalhos que discutam as limitações e forneçam caminhos claros para o problema descrito. Assim, este artigo apresenta uma Revisão Sistemática da Literatura (SLR) sobre avaliação e monitoramento de modelos de AM do mundo real para aplicações em saúde.

A revisão segue a metodologia de Resistência Sistemática definida por Kitchenhan (Kitchenham & Charters, 2007) e reflete a literatura atual sobre avaliação e manutenção de modelos de LM em saúde. Esse tipo de estudo apresenta limitações temporais, pois observa trabalhos publicados até a data de sua realização. Por outro lado, é um estudo facilmente reprodutível, pois se baseia em um protocolo formal de revisão da literatura.

É importante mencionar que os trabalhos listados na revisão foram analisados considerando todo o ciclo de vida de um modelo de AM aplicado a um contexto real na área da saúde, que compreende: avaliação de desempenho, monitoramento do modelo e manutenção. Além disso, este trabalho também apresenta, como objetivo secundário, uma abordagem para analisar e avaliar o desempenho de modelos de AM em saúde que será proposta com base nos resultados da revisão e observações realizadas.

As próximas seções estão organizadas da seguinte forma: a seção 2 discutirá conceitos relacionados; A seção 3 apresenta a metodologia de revisão; A seção 4 analisa os resultados da revisão sistemática; A seção 5 apresenta a discussão e delineia uma proposta de pesquisa; e, finalmente, a Seção 6 discute conclusões e trabalhos futuros.

2 CONCEITOS RELACIONADOS

Este trabalho está relacionado ao monitoramento e avaliação de modelos de AM no contexto da saúde. Nesse sentido, esta seção explicará brevemente alguns aspectos importantes para a avaliação e observação contínua dos resultados e desempenho de um modelo de AM.

2.1 AVALIAÇÃO DO MODELO AM

A construção de um modelo de Aprendizado de Máquina envolve as seguintes etapas: pré-processamento, que inclui coleta e manuseio de dados; processamento, que equivale a executar métodos de AM sobre os dados pré-processados; e pós-processamento, com coleta e análise de métricas de desempenho do modelo (Mitchell et al., 2007). Tradicionalmente, o pós-processamento inclui testes, o que significa treinar o modelo em uma amostra de dados para coletar métricas de desempenho. Outra atividade, chamada validação, geralmente é realizada após o teste como parte da etapa de pós-processamento. Essa atividade envolve a verificação do desempenho do modelo em relação a diferentes amostras de dados mantidas especificamente para essa finalidade. Depois disso, o modelo é serializado e incorporado em sua aplicação de destino para cumprir seu papel na solução do problema proposto (Gopal, 2019).

Esse contexto delimita um problema. Se a validação do modelo ocorre antes da entrega e do uso efetivo em relação a dados do mundo real, o monitoramento e a avaliação de desempenho na operação real (em produção) também podem ser chamados de validação? Se sim, como diferenciar um do outro? A literatura atual parece ter pouca consideração a esse respeito. Validação e avaliação geralmente se referem tanto às etapas finais da construção do modelo (pós-processamento) quanto à avaliação desse mesmo modelo depois que ele está efetivamente em uso. Isso torna a pesquisa de monitoramento e avaliação de modelos desafiadora, dada a falta de consenso sobre a terminologia.

Nesta pesquisa, validação, avaliação de desempenho, monitoramento e manutenção referem-se a modelos já construídos e efetivamente em uso, e não àqueles ainda em desenvolvimento.

2.2 MONITORAMENTO E AVALIAÇÃO CONTÍNUA

Há consideráveis desafios para a AM para a Saúde inerentes ao contexto clínico. Por exemplo: lidar com grandes volumes de dados, complexidade de dados, dados não estruturados e preocupações com a privacidade do paciente, sem mencionar os requisitos críticos em relação à precisão, uma vez que erros podem resultar em situações de risco de vida para os pacientes. Esses fatores podem se tornar um fator decisivo para a eficácia e utilidade do modelo de AM. Portanto, o monitoramento contínuo e a avaliação de desempenho para aplicativos de AM de saúde são uma necessidade crítica.

As Operações de Aprendizado de Máquina (AMOps), que adaptam os princípios de DevOps ao ciclo de vida do modelo de AM, pretendem gerenciar o Ciclo de Inteligência para modelos de AM para que as pessoas possam trabalhar juntas para imaginar, desenvolver, implantar, operar, monitorar e melhorar sistemas de aprendizado de máquina continuamente (Treveil et al., 2020).

Colocar modelos em produção é apenas parte do processo, não o fim dele. Uma vez que um modelo esteja em operação, os dados de produção devem ser coletados e monitorados continuamente para fechar o ciclo de feedback. Dessa forma, novos dados podem ser selecionados e rotulados em novos conjuntos de dados de treinamento e ser usados para melhorar os modelos de AM. Isso permitiria que os modelos se adaptassem e melhorassem continuamente (Maleki et al., 2020).

Fatores inerentes aos aspectos de negócios e produtos podem afetar o ciclo de vida dos modelos de AM, como o custo de implementação e o impacto do modelo (Wiens et al., 2019). O desalinhamento entre o modelo e as métricas de negócios pode levar a efeitos indesejáveis no desempenho do modelo. Um modelo estatisticamente preciso que não atenda às expectativas de negócios está fadado ao fracasso. Portanto, estudos sobre monitoramento e validação contínua de modelos são essenciais. Isso é especialmente verdadeiro em contextos como o AM para a saúde.

3 METODOLOGIA

De acordo com Kitchenham & Charters (2007), uma Revisão Sistemática é um estudo que visa identificar trabalhos de pesquisa relacionados a um tópico específico e aborda questões mais amplas sobre a evolução da pesquisa. Portanto, a realização de uma Revisão Sistemática da Literatura (SLR) é adequada para este trabalho, que busca compreender o estado da arte atual em relação à avaliação, monitoramento e manutenção de modelos assistenciais. Esse processo utiliza uma abordagem quantitativa para coletar e organizar os dados selecionados e uma análise qualitativa para comparar os critérios de qualidade estabelecidos para entender o panorama atual de avaliação e monitoramento do

modelo. O processo de pesquisa ocorre em três etapas: planejamento, execução e extração de dados, conforme detalhado nas seções a seguir.

3.1 PLANEJAMENTO DE PESQUISA

A Revisão Sistemática da Literatura inicia-se com o planejamento metodológico para reduzir erros e vieses na seleção e análise dos estudos. O planejamento define o objetivo da pesquisa, as perguntas, o mecanismo de busca, a cadeia de pesquisa, a inclusão, a exclusão e os critérios de qualidade. Estes são necessários para a fase de execução.

3.1.1 Objetivo da Pesquisa e Questões de Pesquisa

O objetivo principal desta revisão é estabelecer o estado atual da arte em relação à avaliação, monitoramento e manutenção de modelos de atenção à saúde. As seguintes Questões de Pesquisa (QR) explicam isso:

- **RQ1**: Quais métodos e técnicas avaliam o desempenho de modelos de aprendizado de máquina em aplicações do mundo real?
- **RQ2**: Quais são suas principais características e como são descritas?
- **RQ3**: Existem especificidades para a avaliação do modelo de AM em aplicações de Saúde?
- **RQ4**: Como a atualização do modelo é tratada considerando a operação do sistema e como a garantia de qualidade de dados de domínio acontece?
- **RQ5**: Quais são os principais desafios e oportunidades na avaliação de modelos de AM em aplicações de saúde?

3.1.2 Mecanismo de busca, critérios de inclusão e exclusão

O buscador Scopus, da *Elsevier,* foi escolhido como plataforma para a pesquisa, pois indexa as bases de dados mais relevantes para as áreas de ciência da computação e aprendizado de máquina, como *ACM Digital Library*, *IEEE Explorer*, *Science Direct* e *Springer Link*. Os critérios de inclusão e exclusão, que determinam quais estudos devem ser incluídos ou excluídos em uma revisão sistemática, foram definidos a seguir.

- **Critérios de inclusão**:
  - somente estudos escritos em inglês;
  - Os estudos devem propor ou analisar o processo de avaliação de modelos de aprendizado de máquina em aplicações de saúde.
- **Critérios de exclusão**:
  - Literatura cinzenta (livros, relatórios técnicos, artigos não científicos);
  - Resultados duplicados;

- Trabalhos de mesmo autor ou de mesma pesquisa;
- Trabalhos não relacionados à saúde;
- Trabalhos não relacionados ao Aprendizado de Máquina;
- Obras que não abordam a operação do mundo real;
- Obras indisponíveis para download;
- Trabalhos que não abordam nenhuma das questões de pesquisa;
- Trabalhos publicados antes de 2010.

3.1.3 Critérios de Qualidade

Os Critérios de Qualidade (CQ) avaliam a aderência do trabalho ao objetivo da pesquisa e às questões de pesquisa. Em outras palavras, as questões de pesquisa estabelecem o que deve ser investigado, e critérios de qualidade quantificam objetivamente o valor dos trabalhos para a pesquisa. Foram estabelecidos os seguintes critérios de qualidade:

- **QC1**: O trabalho aborda a avaliação de modelos de aprendizado de máquina já em uso em uma operação do mundo real (ou seja, em um "ambiente de produção")?
- **QC2**: O trabalho detalha claramente o procedimento de avaliação de um ou mais modelos de aprendizado de máquina em produção?
- **QC3**: Existem particularidades relacionadas à gestão de modelos de Aprendizado de Máquina em aplicações de saúde?
- **QC4**: As escolhas de gerenciamento de mudanças relacionadas a dados são detalhadas junto com suas motivações?
- **QC5**: As escolhas de gerenciamento de modelos são detalhadas junto com suas motivações?
- **QC6**: São descritas limitações e oportunidades para avaliação de modelos de aprendizado de máquina na produção?
- **QC7**: O trabalho descreve ou propõe um framework para avaliação de modelos de produção de forma estruturada e reprodutível?
- **QC8**: O trabalho vai além das técnicas estatísticas de avaliação de modelos, levando em consideração opiniões de especialistas do domínio e/ou protocolos específicos para a área de aplicação?

A mensuração dos critérios de qualidade de cada trabalho é feita por meio de uma escala. Após a leitura do trabalho, cada um recebe uma pontuação indicando o quão bem abordam cada critério de qualidade. Utilizou-se a seguinte escala: 0, quando não contempla o critério de qualidade; 0,5, quando atende parcialmente ao critério; e 1.0, quando atende plenamente a ela.

De acordo com Kitchenham & Charters (2007), uma cadeia de busca deve ser refinada em um processo iterativo de tentativa, observação e refatoração que visa retornar trabalhos o mais coerentes possível ao sujeito da pesquisa. A sequência de pesquisa foi baseada nas perguntas de pesquisa e palavras-chave amplamente utilizadas em aplicativos de Aprendizado de Máquina for Healthcare. A seguinte cadeia de caracteres de pesquisa resultou desse processo:

"saúde" E ("aprendizado de máquina" OU "OPERAÇÕES DE AM" OU "AMOPS" OU "operação de aprendizado de máquina") E ("melhoria contínua" OU "implantação contínua" OU "aprendizado contínuo" OU "deriva de modelo" OU "deriva de dados" OU "deriva de alvo" OU "deriva de conceito" OU "decaimento de modelo" OU "loop de feedback" OU "integridade do modelo" OU "integridade de aprendizado de máquina" OU "validação de modelo" OU "avaliação de modelo" OU "avaliação de aprendizado de máquina" OU "validação de aprendizado de máquina")

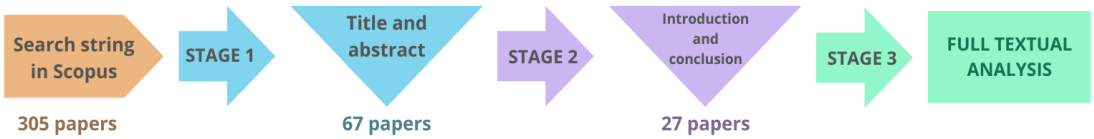
Após a definição da string, a busca foi realizada no buscador escolhido, considerando o título, o resumo e as palavras-chave dos trabalhos. Os dados coletados e as anotações referentes às etapas de execução da pesquisa (a serem descritas a seguir) estão disponíveis em planilha eletrônica acessível através do link: *https://bit.ly/3XktPfB*. Os dados extraídos incluem o ano de publicação; título do trabalho; lista de autores; Keywords; tipo de trabalho; e link (URL).

3.2 EXECUÇÃO

O protocolo de Revisão Sistemática seguido nesta pesquisa divide a execução em três etapas sucessivas: [1] Inicialmente, são lidos o título e resumo de cada trabalho; [2] em seguida, é lida a introdução e conclusão dos selecionados; [3] e, por fim, os filtrados considerados aderentes à pesquisa são lidos na íntegra. Os critérios de inclusão e exclusão são observados durante as leituras das duas primeiras etapas. Quando um artigo não atende a todos os critérios de inclusão ou toca em algum critério de exclusão, ele é removido e não será lido na última etapa. Na etapa final, os artigos remanescentes das etapas 1 e 2 são lidos na íntegra e os critérios de qualidade são mensurados.

A Figura 1 descreve o processo de pesquisa. Dois pesquisadores analisaram cada trabalho para as etapas 1 e 2. Para evitar viés, cada pesquisador indicou separadamente se o trabalho deveria ser excluído ou mantido para a etapa final, com base nos critérios de inclusão e exclusão. Em caso de discordância, uma conversa consensual entre os pesquisadores definiria se o artigo deveria permanecer. Na etapa final, apenas um pesquisador por trabalho foi envolvido. A Tabela 1 detalha a quantidade inicial em cada etapa, quantos foram removidos e quantos permaneceram.

Figura 1: Protocolo de revisão.



Fonte: Autores (2023).

Tabela 1: Trabalhos incluídos e excluídos em cada etapa.

Entrada	Removido	Remanescente

Fonte: Autores (2023).

4 RESULTADOS

Após a execução das duas primeiras iterações (etapas 1 e 2), 27 (vinte e sete) obras foram selecionadas para leitura na íntegra. Na etapa 3, realizou-se a avaliação dos critérios de qualidade de cada um. As questões de pesquisa foram então analisadas por meio de medidas de qualidade e dados extraídos da leitura de cada trabalho. Esta seção detalha parte dessa análise.

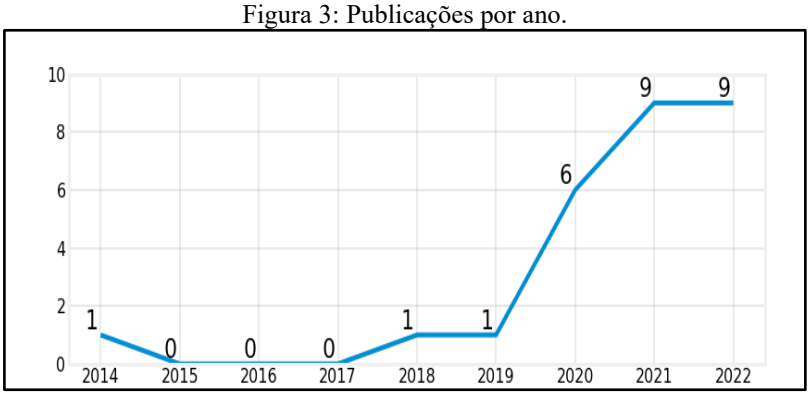
As obras da fase 3 foram categorizadas de acordo com sua editora. A Figura 2 mostra aqueles à esquerda, deixando claro que diversos editores estavam envolvidos. O lado direito da Figura 2 demonstra uma predominância de periódicos quanto ao tipo de publicação, perfazendo cerca de 89% dos trabalhos lidos na íntegra.

Figura 2: Editores e tipo de publicação.

Fonte: Autores (2023).

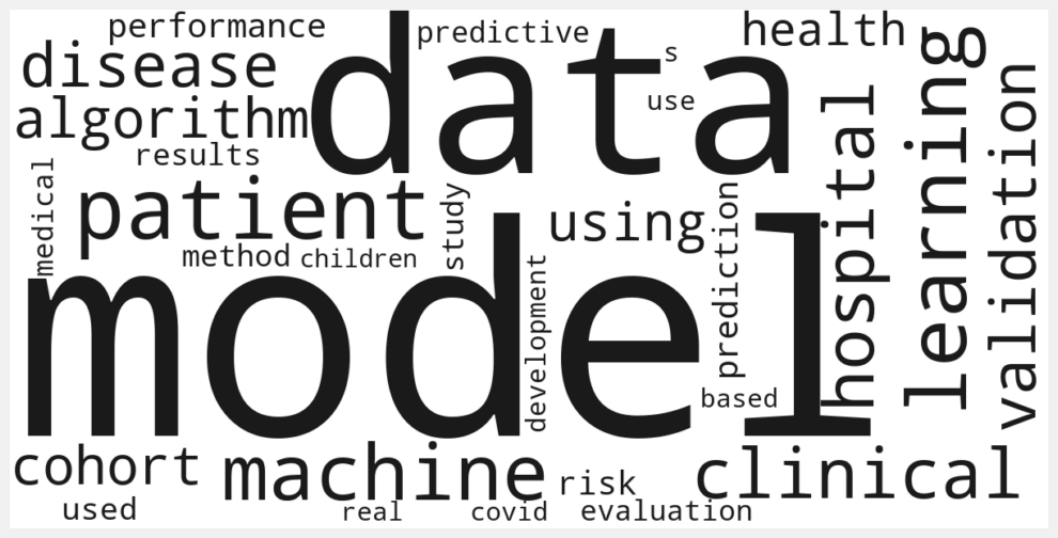
A Figura 3 apresenta a distribuição dos artigos selecionados para leitura na íntegra por ano. Com base na figura, é possível inferir que a Validação de Modelo para aplicações em Saúde vem ganhando relevância, principalmente nos últimos três anos, quando se observa um crescente esforço de pesquisa relacionado ao tema, demonstrando que está se tornando um tema de pesquisa acalorado. Com base nos resumos dos artigos lidos na íntegra, foi construído um gráfico de nuvem de palavras adicional, como mostra a Figura 4, com os termos mais citados dentro dos resumos (quanto mais uma palavra aparece, maior se torna a fonte do texto).

Figura 3: Publicações por ano.



Fonte: Autores (2023).

Figura 4: Nuvem do Word.

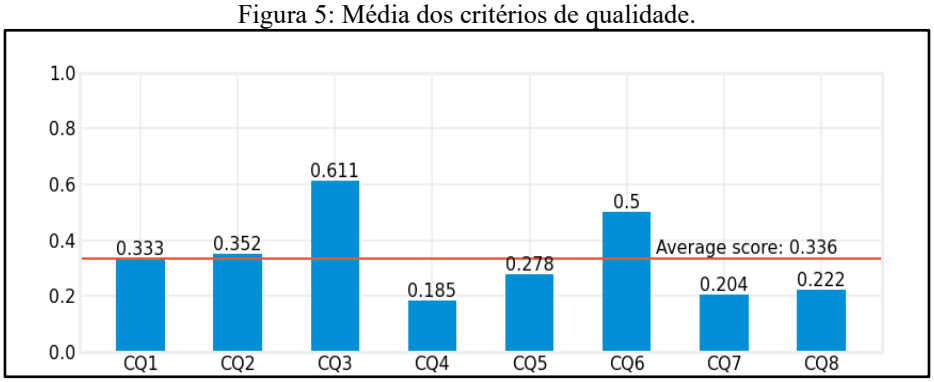


Fonte: Autores (2023).

Pelos valores medidos para os critérios de qualidade, é possível observar, do ponto de vista de cada critério, como os artigos lidos na íntegra geralmente atendiam aos critérios de qualidade. Essa visualização traz uma perspectiva importante sobre a maturidade dos trabalhos em termos de cada critério.

A Figura 5 apresenta os valores médios alcançados pelos artigos lidos na íntegra em cada critério de qualidade. É possível observar que a média geral, traçada em linha tracejada laranja, tem valor de 0,336 e que todos os critérios obtiveram médias abaixo de 0,7, com apenas dois critérios alcançando médias acima de 0,5.

Figura 5: Média dos critérios de qualidade.



Fonte: Autores (2023).

A lista abaixo apresenta os valores médios alcançados por cada critério de qualidade, seguido de uma breve discussão.

- **QC1**: a média obtida nesse critério foi de 0,333. Esse valor indica que a avaliação e o monitoramento dos modelos assistenciais em produção não têm sido consistentemente abordados pelos trabalhos.
- **QC2**: os artigos selecionados obtiveram média de 0,352 para esse critério, o que indica que falta clareza e profundidade na descrição dos procedimentos de avaliação dos modelos em produção.
- **QC3**: este foi o critério de maior média entre as obras lidas, atingindo uma média de 0,611. Esse valor indica que eles podem identificar particularidades da LM para a saúde em algum grau. Apesar disso, nota-se com esse valor que há condições para aprofundar a discussão sobre essas particularidades.
- **QC4**: diferentemente do critério anterior, o valor médio obtido pelos trabalhos neste critério foi de apenas 0,185, o menor valor dentre todos os critérios de qualidade. Com esse resultado, é possível observar que as decisões de gerenciamento de mudanças relacionadas aos dados são relatadas às pressas e podem ser significativamente melhoradas.
- **QC5**: nesse critério, os trabalhos alcançaram média de 0,278, denotando que as escolhas feitas para a gestão do modelo são relatadas apenas superficialmente.
- **QC6**: em relação às limitações e oportunidades na avaliação de um modelo em produção, a nota média alcançada pelos trabalhos, 0,5, indica a abordagem do mesmo, mas que ainda pode haver necessidade de aprofundamento nessa questão.
- **QC7**: os trabalhos atingiram média de 0,204 nesse critério. Assim, observa-se que o esforço de pesquisa para estabelecer estruturas para avaliar modelos de AM é limitado.
- **QC8**: o último critério apresentou 0,222 como média obtida pelos trabalhos. Esse valor indica que aqueles não priorizaram as opiniões de especialistas do domínio ou usaram protocolos específicos da área de aplicação para avaliar e monitorar modelos.

A partir da análise de conteúdo dos trabalhos lidos e dos resultados individuais dos critérios de qualidade, foi possível observar como cada um respondeu às Questões de Pesquisa. A Tabela 2 apresenta esse detalhamento, marcando com um "x" as questões de pesquisa respondidas por cada artigo. Na parte inferior da tabela também pode ser visto um resumo com o número de trabalhos que responderam a cada Pergunta de Pesquisa. No entanto, a tabela não discrimina se as perguntas foram respondidas de forma satisfatória ou superficial. Apenas indica se aquele trabalho se aproxima de uma Questão de Pesquisa.

Tabela 2: Questões de pesquisa tocadas pelo trabalho.

Trabalho	**RQ1**	**RQ2**	**RQ3**	**RQ4**	**RQ5**
(2020)			x		x
(2022)			x		
Johri, Sen Saxena, & Kumar (2021)			x		
(2022)	x	x	x	x	x
(2020)			x	x	x
Wojtusiak (2021)	x	x		x	
(2022)			x		
(2022)	x	x	x		x
Risman, Trelles, & Denning (2021)	x	x	x		x
(2021)			x		
(2022)	x	x	x	x	
(2020)	x	x	x		x
(2021)			x		x
(2020)			x	x	
Rafiq, Modave, Guha, Albert (2020)	x	x			x
Harris et al. (2022)		x	x	x	x
(2020)					x
Vieira, Fernandes, Lucena, & Lifschitz (2021)				x	
Consórcio RADAR-CNS et al (2021)	x	x	x		x
Huda et al. (2021)	x	x	x		
(2022)	x	x	x	x	x
(2022)	x	x	x	x	x
(2021)	x	x			
(2022)	x	x	x	x	x
Yang, Zou, Liu, & Mulligan (2014)					x
(2018)	x	x	x		
Fries et al. (2019)		x	x		x
Total	**14**	**16**	**21**	**10**	**16**

Fonte: Autores (2023).

5 DISCUSSÃO

Esta seção apresenta uma breve discussão dos achados da revisão. Aborda algumas perspectivas para cada questão de pesquisa, utilizando tanto os resultados da sessão anterior quanto o conteúdo dos trabalhos lidos. Medidas de critérios de qualidade também serão utilizadas como base para a discussão, uma vez que partiram das questões de pesquisa.

Em relação ao **RQ1** que delimita uma investigação sobre quais métodos e técnicas são usados para avaliar o desempenho do modelo de AM em aplicações do mundo real. Os valores médios obtidos pelos artigos nos critérios de qualidade 1, 2 e 8, respectivamente 0,333, 0,352 e 0,222, indicando que o detalhamento das técnicas utilizadas para validar modelos de AM no mundo real é superficial. Isso se torna uma questão ainda maior em um contexto como o da saúde, onde erros podem levar a situações de risco de vida para os pacientes, o que pode acabar sendo uma barreira para a adoção de aprendizado de máquina em ambientes clínicos e contextos de saúde em geral.

Observa-se nos trabalhos que faltam dados concretos, métricas e melhores práticas para avaliar modelos em produção, ou seja, modelos de AM já implantados e em operação em sistemas do mundo real. A maioria dos artigos revisados apresentou apenas relatos experimentais, focando principalmente na avaliação estatística do desempenho do modelo durante sua construção, como é o caso de (Van Helvoort et al., 2020; Johri, Sen Saxena, & Kumar, 2021; Qasim et al., 2021; Sol et al., 2022; Maleki et al., 2020). Alguns artigos relataram testes realizados em ambientes reais com pacientes. No entanto, eles não detalharam seus procedimentos de avaliação em modelos de produção (Lam et al., 2022; Birkenbihl et al., 2020; Kamran et al., 2022; Consórcio RADAR-CNS et al., 2021). Também é perceptível que há pouca informação sobre as métricas e as melhores práticas para avaliação de modelos na produção para aplicações de saúde.

A análise do RQ1 está altamente relacionada ao RQ2, que trata das características dos métodos e técnicas utilizados para avaliar modelos de AM no mundo real. Portanto, dada a escassez de respostas relacionadas às práticas de avaliação de modelos de AM na produção, há pouca documentação sobre as características dos métodos e técnicas utilizados. Apesar disso, alguns trabalhos mencionam a necessidade de cuidados especiais na avaliação estatística dos dados de treinamento dos modelos. Especialmente quando os grupos que originam os dados de treinamento (pacientes de um hospital específico ou pessoas de determinadas regiões geográficas, por exemplo) têm características distintas (em termos de dados), aplicar esse mesmo modelo a outros grupos pode levar a um baixo desempenho do modelo (Sun et al., 2022; Rafiq, Modave, Guha, Alberto, 2020). Há também comentários sobre a necessidade de profissionais especializados participarem da construção e validação do modelo para promover melhor confiabilidade (Wojtusiak, 2021; Risman, Trelles, Denning, 2021; Harris et al., 2022; Rojas et al., 2022). Os especialistas podem ajudar tanto no processamento quanto na compreensão dos dados, nos testes de desempenho dos modelos e na

definição de métodos de avaliação, garantindo assim que os modelos resultantes sejam precisos e confiáveis.

Outra questão apontada por alguns trabalhos é a necessidade de uma boa interpretabilidade do modelo (Rafiq, Modave, Guha, & Albert, 2020; Harris et al., 2022; Li et al., 2022; Duckworth et al., 2021). A interpretabilidade e a explicabilidade do modelo de AM podem ajudar a garantir que os aplicativos habilitados para AM forneçam decisões coerentes e confiáveis. A explicabilidade é especialmente importante na área da saúde, pois permite a interpretação dos resultados do modelo e facilita a coleta de dados para avaliação do modelo ou processos como a auditoria. Nesse contexto, a comunicação e a colaboração também devem ser priorizadas na validação de modelos de Aprendizado de Máquina em produção, corroborando a necessidade de aprimoramento e aprofundando o assunto, uma vez que as respostas apresentadas para os critérios de qualidade 1, 2 e 8 são superficiais.

O RQ3 busca especificidades do processo de avaliação de modelos de AM em saúde. Está diretamente relacionado ao QC3, no qual os trabalhos obtiveram média de 0,611, a maior pontuação entre todos os critérios de qualidade. Percebe-se, na leitura dos artigos, que parte relevante deles menciona problemas ou especificidades relacionadas à avaliação de modelos em aplicações em saúde (Shickel et al., 2020; Rafiq, Modave, Guha, Alberto, 2020; Rojas et al., 2022; Fries et al., 2019). Uma das questões mais críticas mencionadas é a necessidade de manter os dados atualizados para fornecer informações para a atualização contínua e consistente dos modelos de AM. Portanto, é necessário estabelecer métricas que possam identificar mudanças na distribuição dos dados e desencadear a reciclagem do modelo quando estas forem detectadas (Birkenbihl et al., 2020; Rojas et al., 2022).

Um segundo aspecto diz respeito às preocupações regulatórias e éticas, questões críticas para a gestão do modelo de AM em aplicações de saúde (Carolan et al., 2022; Wojtusiak, 2021). Na área da saúde, questões éticas e regulatórias relativas à confidencialidade, rastreabilidade e explicabilidade do processo decisório (modelo) já estavam fortemente presentes muito antes dos recentes impulsos por direitos de acesso a dados e leis de privacidade de dados por iniciativas como a Lei Geral de Proteção de Dados (LGPD), no Brasil, ou a Lei de Privacidade do Consumidor da Califórnia (CCPA), nos EUA, entre outras (Harris et al., 2022; Maleki et al., 2020; Rojas et al., 2022). Embora essas preocupações regulatórias não sejam específicas do contexto da saúde, elas afetam dramaticamente essa área, uma vez que muitas das melhores práticas de saúde estão relacionadas à personalização das decisões clínicas e à humanização dos processos. Finalmente, embora haja uma discussão razoável sobre as particularidades relevantes para a gestão de modelos de AM em aplicações de saúde, há apenas uma discussão superficial sobre possíveis soluções para os problemas enfrentados pela gestão de modelos devido a essas particularidades. Ou seja, é observável que os trabalhos descrevem problemas existentes, mas não discutem soluções estruturadas para eles (ou apenas o fazem superficialmente).

QC4 e QC5, nos quais os artigos obtiveram médias (respectivamente) de 0,185 e 0,278, estão intimamente relacionados ao RQ4, que busca descrever formas de atualização do modelo de AM durante a operação do sistema e as premissas de qualidade observadas nos dados do domínio. Os valores obtidos para os CQ indicam que detalhes sobre as decisões tomadas em relação às atualizações de modelos são escassos. Vale ressaltar que, dados os requisitos críticos de desempenho das aplicações de saúde, é vital entender como gerenciar as atualizações do modelo de AM quando a distribuição de dados de entrada muda, os conceitos se desviam ou o próprio modelo não é mais uma solução viável para o problema em questão (Vieira, Fernandes, Lucena, & Lifschitz, 2021).

O RQ5 e o QC6 relacionado abordam os desafios e oportunidades relacionados à avaliação e monitoramento do modelo de AM em aplicações de saúde. Os trabalhos obtiveram média de 0,500 no QC6. Esse valor indica algum nível de profundidade na discussão de desafios e oportunidades. Os desafios mencionados incluem a obtenção de dados em tempo real, escassez de dados, manutenção de sistemas existentes, quantificação da comparabilidade dos dados de validação (de novos pacientes) com dados de treinamento, acessibilidade e continuidade dos dados, padronização dos modelos, desequilíbrio dos dados, rotina clínica e disponibilidade de especialistas. Por exemplo, modelos treinados em dados derivados de uma única instituição de saúde podem não generalizar bem em cenários multi-institucionais. Uma variação desse problema são os vieses de seleção de pacientes (regionais, socioeconômicos e institucionais) (Van Helvoort et al., 2020; Carolan et al., 2022; Lam et al., 2022; Birkenbihl et al., 2020; Kamran et al., 2022; Risman, Trelles, Denning, 2021; Shickel et al., 2020; Bellocchio et al., 2021; Rafiq, Modave, Guha, Alberto, 2020; Harris et al., 2022; Maleki et al., 2020; Consórcio RADAR-CNS et al., 2021; Li et al., 2022; Lin et al., 2022; Rojas et al., 2022; Yang, Zou, Liu, Mulligan, 2021; Fries et al., 2019).

Tais desafios podem afetar a viabilidade da avaliação e monitoramento do modelo de AM para aplicações em saúde. Apesar disso, as discussões em curso sobre esses temas podem favorecer o surgimento de abordagens que possam fornecer soluções ou formas de mitigar riscos, bem como novos negócios e serviços de saúde. Outros desafios estão relacionados à Aprendizagem Contínua em saúde, que apresenta diferentes limitações.

Em relação às oportunidades apresentadas nos trabalhos selecionados, há menções à criação de padrões e guias internacionais para lidar com os desafios regulatórios da AM em aplicações de saúde. (2022) descreve a necessidade de melhores tecnologias de automação para melhorar a eficiência dos algoritmos. Há também oportunidades para gestão e monitoramento especializado (Algorithmic Stewardship), com projeções de criação em futuro próximo de departamentos de AMOps para serviços de saúde e hospitais (Harris et al., 2022). Outras possibilidades incluem integrar a equidade no ciclo de vida do AM, removendo vieses, bem como coletar *feedback de* especialistas e outras partes interessadas para trazer o conhecimento humano para o processo de aprendizagem (Human-in-the-

Loop Learning), e ir além das métricas estatísticas na avaliação do desempenho do modelo, usando abordagens orientadas a domínio para medir a utilidade e o valor comercial dessas (Rojas et al., 2022; Yang, Zou, Liu, & Mulligan, 2021). Finalmente, há oportunidades para aplicativos do mundo real suportados por dados ao vivo, onde as equipes podem construir e testar iterativamente à beira do leito, plataformas AMOps de entrega contínua (CD), design e supervisão por pessoas com experiência em segurança de IA, avaliação contínua usando randomização para evitar viés e uso de fluxos de dados com o protocolo HL7-FHIR (Harris et al., 2022).

Com base nessas observações, percebe-se a necessidade de aprimoramento e aprofundamento das pesquisas relacionadas à avaliação e monitoramento do modelo de AM em aplicações de saúde. O QC7 busca trabalhos que discutam e proponham soluções para avaliar modelos de AM de forma estruturada e reprodutível. A média geral nesse critério foi de 0,204. Além disso, dos 27 artigos lidos, apenas três (3) atendem plenamente a esse critério (Carolan et al., 2022; Kamran et al., 2022; Fries et al., 2019), o que reforça a necessidade de pesquisas que definam, discutam e aprimorem os métodos de avaliação e manutenção do modelo de AM, especialmente em aplicações críticas como a saúde. Portanto, a principal observação para o QC7 é a necessidade de uma abordagem metodológica para o modelo de AM avaliando, monitorando e mantendo em aplicações de saúde uma vez em operação (produção) no mundo real.

6 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho apresenta o resultado de uma revisão sistemática da literatura que buscou compreender o estado atual da avaliação, monitoramento e manutenção de modelos de Aprendizado de Máquina em aplicações de saúde. Seguindo o protocolo de Kitchenhan (Kitchenham & Charters, 2007), vinte e sete (27) artigos foram submetidos à análise completa. Os resultados obtidos e as discussões que se seguiram (apresentadas nas seções anteriores) indicam a necessidade de mais pesquisas envolvendo avaliação, monitoramento e manutenção de modelos de AM em aplicações de saúde do mundo real. Dito isso, documentação razoável de problemas e limitações está disponível, o que pode fornecer um ponto de partida para pesquisas futuras.

A dificuldade em encontrar estudos que vão além do relato experimental e avaliam efetivamente modelos de AM em operação no mundo real sugere que uma ênfase considerável tem ocorrido na construção de modelos e validação experimental. No entanto, a continuidade desses esforços não parece acontecer quando os modelos entram em operação do sistema. Como resultado, a contabilização da operação do modelo em dados do mundo real não foi abordada de forma consistente. As aplicações de saúde exigem monitoramento, validação e manutenção contínuos dos modelos devido à própria criticidade do domínio e dos serviços envolvidos.

Portanto, embora se reconheça a importância da avaliação e monitoramento contínuo de modelos, a literatura ainda necessita de estudos práticos e metodologias detalhadas para avaliação contínua de modelos de AM em aplicações em saúde. É essencial continuar pesquisando e desenvolvendo métodos eficazes para avaliar, monitorar e manter modelos de AM para garantir que eles sejam seguros, confiáveis e úteis para aplicações de saúde.

Os resultados da revisão sistemática sugerem a necessidade de um fluxo de trabalho de gerenciamento de mudanças para desenvolvedores e gerentes de modelos de AM. Esse processo, a ser proposto em trabalhos futuros, deve incluir as seguintes atividades: [1] Obtenção de documentação disponível (por exemplo, desempenho do modelo de linha de base, decisões de planejamento experimental), [2] Definição de critérios e parâmetros de avaliação com base na opinião de especialistas, desempenho estatístico do mundo real de modelos (métricas quantitativas) e protocolos específicos de produtos, negócios e áreas de aplicação (métricas qualitativas); [3] Prototipagem de avaliação com especialistas de negócios e domínio; [4] Operacionalização e monitoramento dos critérios de medição; [5] Avaliação de critérios de medição (por exemplo, vieses, deriva, resultados atrasados, desempenho estatístico e empresarial); e [6] refatoração de modelos, que pode incluir subatividades como [a] Janela deslizante de coleta e armazenamento de dados do mundo real; [b] Treinamento de modelos com dados clínicos do mundo real; [c] Validação estatística; [d] Ajuste de hiperparâmetros; [e] Retreinamento do modelo sempre que a distribuição de dados mudar; [f] Padronização de modelos.

Outros trabalhos futuros poderiam estabelecer uma abordagem metodológica para avaliar o nível de maturidade dos modelos de AM, uma vez em uso no mundo real, com base em boas práticas e preocupações que permeiam todo o ciclo de vida dos modelos.

AGRADECIMENTOS

Os autores agradecem ao Ministério da Ciência, Tecnologia e Inovações (MCTI) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo financiamento da pesquisa por meio da Chamada RHAE - Recursos Humanos em Áreas Estratégicas.

REFERÊNCIAS

Arowolo, M. O., Ogundokun, R. O., Misra, S., Kadri, A. F., & Aduragba, T. O. (2022). Machine Learning Approach Using KPCA-SVMs for Predicting COVID-19. In Garg, L., Chakraborty, C., Mahmoudi, S., Sohmen, V. S. (Eds.), *Healthcare Informatics for Fighting COVID-19 and Future Epidemics* (pp. 193–209). Springer International Publishing. https://doi.org/10.1007/978-3-030-72752-9_10

Bellocchio, F., Lonati, C., Ion Titapiccolo, J., Nadal, J., Meiselbach, H., Schmid, M., Baerthlein, B., Tschulena, U., Schneider, M., Schultheiss, U. T., Barbieri, C., Moore, C., Steppan, S., Eckardt, K.-U., Stuard, S., & Neri, L. (2021). Validation of a Novel Predictive Algorithm for Kidney Failure in Patients Suffering from Chronic Kidney Disease: The Prognostic Reasoning System for Chronic Kidney Disease (PROGRES-CKD). *International Journal of Environmental Research and Public Health*, *18 (23)*. https://doi.org/10.3390/ijerph182312649

Birkenbihl, C., Emon, M. A., Vrooman, H., Westwood, S., Lovestone, S., AddNeuroMed Consortium, Hofmann-Apitius, M., Fröhlich, H., & Alzheimer's Disease Neuroimaging Initiative (2020). Differences in Cohort Study Data Affect External Validation of Artificial Intelligence Models for Predictive Diagnostics of Dementia - Lessons for Translation Into Clinical Practice. *The EPMA Journal, 11 (3)*, 367–376. https://doi.org/10.1007/s13167-020-00216-z

Carolan, J. E., McGonigle, J., Dennis, A., Lorgelly, P., & Banerjee, A. (2022). Technology-Enabled, Evidence-Driven, and Patient-Centered: The Way Forward for Regulating Software as a Medical Device. *JMIR Med Inform, 10 (1)*, e34038. https://doi.org/10.2196/34038

Collin, C. B., Gebhardt, T., Golebiewski, M., Karaderi, T., Hillemanns, M., Khan, F. M., Salehzadeh-Yazdi, A., Kirschner, M., Krobitsch, S., consortium, E.-S., & Kuepfer, L. (2022). Computational Models for Clinical Applications in Personalized Medicine-Guidelines and Recommendations for Data Integration and Model Validation. *Journal of Personalized Medicine, 12 (2)*. https://doi.org/10.3390/jpm12020166

Duckworth, C., Chmiel, F. P., Burns, D. K., Zlatev, Z. D., White, N. M., Daniels, T. W. V., Kiuber, M., & Boniface, M. J. (2021). Emergency Department Admissions During COVID-19: Explainable Machine Learning to Characterise Data Drift and Detect Emergent Health Risks. *MedRxiv*. https://doi.org/10.1101/2021.05.27.21257713

Fries, J. A., Varma, P., Chen, V. S., Xiao, K., Tejeda, H., Saha, P., Dunnmon, J., Chubb, H., Maskatia, S., Fiterau, M., Delp, S., Ashley, E., Ré, C., & Priest, J. R. (2019). Weakly Supervised Classification of Aortic Valve Malformations Using Unlabeled Cardiac MRI Sequences. *BioRxiv*. https://doi.org/10.1101/339630

Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., & Ranganath, R. (2020). A Review of Challenges and Opportunities in Machine Learning for Health. *AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science Proceedings, 2020,* 191–200.
https://doi.org/10.48550/arXiv.1806.00388

Gopal, M. (2019). *Applied Machine Learning.* McGraw-Hill Education.

Harris, S., Bonnici, T., Keen, T., Lilaonitkul, W., White, M. J., & Swanepoel, N. (2022). Clinical Deployment Environments: Five Pillars of Translational Machine Learning for Health. *Frontiers in Digital Health, 4*. https://doi.org/10.3389/fdgth.2022.939292

Van Helvoort, E. M., van Spil, W. E., Jansen, M. P., Welsing, P. M., Kloppenburg, M., Loef, M., Blanco, F. J., Haugen, I. K., Berenbaum, F., Bacardit, J., & others. (2020). Cohort Profile: The Applied Public-Private Research Enabling Osteoarthritis Clinical Headway (IMI-APPROACH) Study: A 2-Year, European, Cohort Study to Describe, Validate and Predict Phenotypes of Osteoarthritis Using Clinical, Imaging and Biochemical Markers. *BMJ Open, 10 (7),* e035101. https://doi.org/10.1136/bmjopen-2019-035101

Huda, A., Castaño, A., Niyogi, A., Schumacher, J., Stewart, M., Bruno, M., Hu, M., Ahmad, F., Deo, R., & Shah, S. (2021). A Machine Learning Model for Identifying Patients at Risk for Wild-type Transthyretin Amyloid Cardiomyopathy. *Nature Communications, 12,* 2725. https://doi.org/10.1038/s41467-021-22876-9

Iakovakis, D., Hadjidimitriou, S., Charisis, V., Bostantjopoulou, S., Katsarou, Z., Klingelhoefer, L., Reichmann, H., Dias, S. B., Diniz, J. A., Trivedi, D., Chaudhuri, K. R., & Hadjileontiadis, L. J. (2018). Motor Impairment Estimates via Touchscreen Typing Dynamics Toward Parkinson's Disease Detection From Data Harvested In-the-Wild. *Frontiers in ICT, 5.* https://doi.org/10.3389/fict.2018.00028

Johri, P., Saxena, V. S., & Kumar, A. (2021). Rummage of Machine Learning Algorithms in Cancer Diagnosis. *International Journal of E-Health and Medical Communications (IJEHMC), 12 (1),* 1–15. http://doi.org/10.4018/IJEHMC.2021010101

Kamran, F., Tang, S., Otles, E., McEvoy, D. S., Saleh, S. N., Gong, J., Li, B. Y., Dutta, S., Liu, X., Medford, R. J., Valley, T. S., West, L. R., Singh, K., Blumberg, S., Donnelly, J. P., Shenoy, E. S., Ayanian, J. Z., Nallamothu, B. K., Sjoding, M. W., & Wiens, J. (2022). Early Identification of Patients Admitted to Hospital for COVID-19 at Risk of Clinical Deterioration: Model Development and Multisite External Validation Study. *BMJ*, *376*. https://doi.org/10.1136/bmj-2021-068576

Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering version 2.3. Engineering, 45(4ve), 1051.

Lam, J., Shimizu, C., Tremoulet, A., Bainto, E., Roberts, S., Sivilay, N., Gardiner, M., Kanegaye, J., Hogan, A., Salazar, J., Mohandas, S., Szmuszkovicz, J., Mahanta, S., Dionne, A., Newburger, J., Ansusinha, E., Debiasi, R., Hao, S., Ling, B., & Sykes, M. (2022). A Machine-Learning Algorithm for Diagnosis of Multisystem Inflammatory Syndrome in Children and Kawasaki Disease in the USA: A Retrospective Model Development and Validation Study. *The Lancet Digital Health, 4,* e717–e726. https://doi.org/10.1016/S2589-7500(22)00149-2

Li, J., Liu, S., Hu, Y., Zhu, L., Mao, Y., & Liu, J. (2022). Predicting Mortality in Intensive Care Unit Patients With Heart Failure Using an Interpretable Machine Learning Model: Retrospective Cohort Study. *J Med Internet Res, 24 (8)*, e38082. https://doi.org/10.2196/38082

Lin, W., Gan, W., Feng, P., Zhong, L., Yao, Z., Chen, P., He, W., & Yu, N. (2022). Online Prediction Model for Primary Aldosteronism in Patients With Hypertension in Chinese Population: A Two-Center Retrospective Study. *Frontiers in Endocrinology, 13*. https://doi.org/10.3389/fendo.2022.882148

Luo, C., Zhu, Y., Zhu, Z., Li, R., Chen, G., & Wang, Z. (2022). A Machine Learning-Based Risk Stratification Tool for In-Hospital Mortality of Intensive Care Unit Patients With Heart Failure. *Journal of Translational Medicine, 20 (1),* 136. https://doi.org/10.1186/s12967-022-03340-8

Maleki, F., Muthukrishnan, N., Ovens, K., Reinhold, C., & Forghani, R. (2020). Machine Learning Algorithm Validation: From Essentials to Advanced Applications and Implications for Regulatory

Certification and Deployment. *Neuroimaging Clinics of North America, 30 (4)*, 433–445. https://doi.org/10.1016/j.nic.2020.08.004

Maleki, F., Muthukrishnan, N., Ovens, K., Md, C., & Forghani, R. (2020). Machine Learning Algorithm Validation. *Neuroimaging Clinics of North America, 30*, 433–445. https://doi.org/10.1016/j.nic.2020.08.004

Malki, Z., Atlam, E.-S., Ewis, A., Dagnew, G., Ghoneim, O. A., Mohamed, A. A., Abdel-Daim, M. M., & Gad, I. (2021). The COVID-19 Pandemic: Prediction Study Based on Machine Learning Models. *Environmental Science and Pollution Research, 28,* 40496–40506. https://doi.org/10.1007/s11356-021-13824-7

Mitchell, T. M., & others. (2007). *Machine Learning* (Vol. 1). McGraw-hill New York.

Qasim, H. M., Ata, O., Ansari, M. A., Alomary, M. N., Alghamdi, S., & Almehmadi, M. (2021). Hybrid Feature Selection Framework for the Parkinson Imbalanced Dataset Prediction Problem. *Medicina, 57 (11),* 1217. https://doi.org/10.3390/medicina57111217

Rafiq, R., Modave, F., Guha, S., & Albert, M. (2020). Validation Methods to Promote Real-world Applicability of Machine Learning in Medicine. *2020 3rd International Conference on Digital Medicine and Image Processing*, 13–19. https://doi.org/10.1145/3441369.3441372

Risman, A., Trelles, M., & Denning, D. W. (2021). Evaluation of Multiple Open-Source Deep Learning Models for Detecting and Grading COVID-19 on Chest Radiographs. *Journal of Medical Imaging, 8 (6)*, 064502. https://doi.org/10.1117/1.JMI.8.6.064502

Rojas, J. C., Fahrenbach, J., Makhni, S., Cook, S. C., Williams, J. S., Umscheid, C. A., & Chin, M. H. (2022). Framework for Integrating Equity Into Machine Learning Models: A Case Study. *Chest, 161 (6)*, 1621–1627. https://doi.org/10.1016/j.chest.2022.02.001

Sengupta, P. P., Shrestha, S., Berthon, B., Messas, E., Donal, E., Tison, G. H., Min, J. K., D'hooge, J., Voigt, J.-U., Dudley, J., Verjans, J. W., Shameer, K., Johnson, K., Lovstakken, L., Tabassian, M., Piccirilli, M., Pernot, M., Yanamala, N., Duchateau, N., & others. (2020). Proposed Requirements for Cardiovascular Imaging-Related Machine Learning Evaluation (PRIME): A Checklist: Reviewed by the American College of Cardiology Healthcare Innovation Council. J*ACC: Cardiovascular Imaging, 13 (9),* 2017–2035. https://doi.org/10.1016/j.jcmg.2020.07.015

Shickel, B., Siegel, S., Heesacker, M., Benton, S., & Rashidi, P. (2020). Automatic Detection and Classification of Cognitive Distortions in Mental Health Text. *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE),* 275–280. https://doi.org/10.1109/BIBE50027.2020.00052

Sun, H., Depraetere, K., Meesseman, L., Cabanillas Silva, P., Szymanowsky, R., Fliegenschmidt, J., Hulde, N., von Dossow, V., Vanbiervliet, M., De Baerdemaeker, J., Roccaro-Waldmeyer, D. M., Stieg, J., Domínguez Hidalgo, M., & Dahlweid, F.-M. (2022). Machine Learning–Based Prediction Models for Different Clinical Risks in Different Hospitals: Evaluation of Live Performance. *J Med Internet Res, 24 (6)*, e34295. https://doi.org/10.2196/34295

The RADAR-CNS Consortium, Böttcher, S., Bruno, E., Manyakov, N. V., Epitashvili, N., Claes, K., Glasstetter, M., Thorpe, S., Lees, S., Dümpelmann, M., van Laerhoven, K., Richardson, M. P., & Schulze-Bonhage, A. (2021). Detecting Tonic-Clonic Seizures in Multimodal Biosignal Data From Wearables: Methodology Design and Validation. *JMIR MHealth and UHealth, 9 (11)*. https://doi.org/10.2196/27674

Treveil, M., Omont, N., Stenac, C., Lefevre, K., Phan, D., Zentici, J., Lavoillotte, A., Miyazaki, M., & Heidmann, L. (2020). *Introducing MLOps*. O'Reilly Media.

Vieira, D. M., Fernandes, C., Lucena, C., & Lifschitz, S. (2021). Driftage: A Multi-Agent System Framework for Concept Drift Detection. *GigaScience, 10 (6)*. https://doi.org/10.1093/gigascience/giab030

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., & others. (2019). Do No Harm: A Roadmap for Responsible Machine Learning for Health Care. *Nature Medicine, 25 (9)*, 1337–1340. https://doi.org/10.1038/s41591-019-0548-6

Wojtusiak., J. (2021). Reproducibility, Transparency and Evaluation of Machine Learning in Health Applications. *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies - HEALTHINF*, 685–692. https://doi.org/10.5220/0010348306850692

Yang, C., Zou, Y., Liu, J., & Mulligan, K. (2014). Predictive Model Evaluation for PHM. *International Journal of Prognostics and Health Management*, 5. https://doi.org/10.36001/ijphm.2014.v5i2.2238