Research, Society and Development, v. 12, n. 6, e5412642042, 2023
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v12i6.42042

A systematic literature review on Machine Learning Model evaluation on healthcare applications

Uma revisão sistemática da literatura sobre avaliação de Modelos de Aprendizado de Máquina em aplicações de saúde

Una revisión sistemática de la literatura sobre la evaluación de Modelos de Aprendizaje Automático en aplicaciones de salud

Received: 05/24/2023 | Revised: 06/01/2023 | Accepted: 06/02/2023 | Published: 06/14/2023

Cezar Miranda Paula de Souza
ORCID: https://orcid.org/0009-0005-7189-8115
Federal University of Rio Grande do Norte, Brazil
E-mail: cezarmiranda@gmail.com
Cephas Alves da Silveira Barreto
ORCID: https://orcid.org/0000-0002-4756-8571
Federal University of Rio Grande do Norte, Brazil
E-mail: cephasax@gmail.com
Lhayana Vieira de Macedo
ORCID: https://orcid.org/0009-0009-0509-0555
Federal University of Rio Grande do Norte, Brazil
E-mail: lhayana11@gmail.com
Bruna Alice Oliveira de Brito
ORCID: https://orcid.org/0009-0001-8116-495X
Federal University of Rio Grande do Norte, Brazil
E-mail: brna.oliveira03@gmail.com
Victor Vieira Targino
ORCID: https://orcid.org/0000-0002-9036-6537
Federal University of Rio Grande do Norte, Brazil
E-mail: victorvieira.rn@gmail.com
Emanuel Costa Betcel
ORCID: https://orcid.org/0009-0009-6814-4311
Federal University of Rio Grande do Norte, Brazil
E-mail: emanuelbetcel@gmail.com
Fernando Gomes de Almeida
ORCID: https://orcid.org/0009-0006-2185-6969
Federal University of Rio Grande do Norte, Brazil
E-mail: fernandogdalmeida@gmail.com
Arthur Andrade Galvíncio Rodrigues
ORCID: https://orcid.org/0009-0002-7107-742X
Federal University of Rio Grande do Norte, Brazil
E-mail: arthurgalvincio.br@gmail.com
Ramon Santos Malaquias
ORCID: https://orcid.org/0000-0002-8350-2836
Federal University of Rio Grande do Norte, Brazil
E-mail: ramonstmalaquias@gmail.com
Itamir de Morais Barroca Filho
ORCID: https://orcid.org/0000-0003-1694-8237
Federal University of Rio Grande do Norte, Brazil
E-mail: itamir.filho@imd.ufrn.br

Abstract

Machine Learning (ML) models have been applied to solve problems in various fields, which necessarily involves proper evaluation of models to ensure performance. Once deployed, ML models are subject to performance issues, such as those related to changes in data (drift). This type of issue has prompted efforts in model analysis and maintenance, as well as in continual learning, which seeks the ability to continuously learn from a (continuous) stream of data. Therefore, it's important to understand and develop methodologies that can be used to evaluate ML models, making their use in real-world environments feasible. Amongst current areas of application for ML, one that stands out, in particular, is Machine Learning for Healthcare, especially in conjunction with Software for Decision Support of Medical Applications, which presents specific challenges for the evaluation and monitoring of models, particularly given that incorrect prediction or classification can lead to life-threatening situations. This paper presents a systematic literature review that aims at identifying state-of-the-art techniques for evaluating and maintaining ML models for healthcare in effective use in the real world.

Keywords: ML model validation; ML for Healthcare; ML model monitoring.

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.

Resumen

Los modelos de Aprendizaje Automático (AA) se han aplicado para resolver problemas en diversos campos, lo que implica necesariamente una adecuada evaluación de los modelos para garantizar su rendimiento. Una vez implementados, los modelos de AA están sujetos a problemas de rendimiento, como los relacionados con los cambios en los datos (drift). Este tipo de problema ha motivado esfuerzos en el análisis y mantenimiento de modelos, así como en el aprendizaje continuo, que busca la capacidad de aprender de forma continua a partir de un flujo continuo de datos. Por lo tanto, es importante entender y desarrollar metodologías que puedan ser utilizadas para evaluar modelos de AA, lo que permite su uso en entornos del mundo real. Entre las áreas actuales de aplicación del AA, una que destaca en particular es el Aprendizaje Automático para la Salud, especialmente en conjunto con el Software de Soporte de Decisiones para Aplicaciones Médicas, lo que presenta desafíos específicos para la evaluación y monitoreo de modelos, especialmente dado que una predicción o clasificación incorrecta puede conducir a situaciones que ponen en peligro la vida. Este artículo presenta una revisión sistemática de la literatura, que tiene como objetivo identificar técnicas de vanguardia para evaluar y mantener modelos de AA para la salud en un uso efectivo en el mundo real.

Palabras clave: Validación de modelos de AA; AA para el sector de la salud; Monitoreo de modelos de AA.

1. Introduction

Recently, Artificial Intelligence (AI) has consolidated itself as one of the go-to alternatives for solving complex problems in any given field of knowledge. It has become increasingly common to hear about or even find systems that make use of AI techniques (e.g. Machine Learning, Expert Systems, Deep Learning, among others) to solve everyday problems.

Healthcare, an area of high social impact, has been the subject of several studies that use Machine Learning (ML) techniques to solve problems. Some studies, for instance, applied ML techniques to predict patient outcomes during the COVID-19 Pandemic (Malki et al., 2021; Arowolo et al., 2022). Others tried to predict risk-of-death for ICU patients with heart failure (Luo et al., 2022). Given the severity of the issues addressed, the usage of ML techniques in healthcare applications faces particularly through modelling, analysis and validation challenges (Ghassemi et al., 2020). Solving them requires close collaboration between data scientists and healthcare experts to make sure that ML models are designed to solve real problems in the field and are interpretable and explainable to the clinical community.

Outcomes and performance of ML models are closely related to the data used for training and testing them (Gopal, 2019). Therefore, it becomes difficult to generalize results obtained with data from specific locations and patient characteristics to those other than those. Another aspect that makes it difficult to analyze and validate model results in healthcare applications is the need for continuous monitoring and specialist feedback, which is difficult to incorporate due to the demanding day-to-day routine of healthcare services professionals. The traditional statistical analysis of results may not be as efficient when it comes to models in production and applied to situations that can mean life or death for patients, delimiting the need for research and development in ML model evaluation and monitoring for healthcare applications.

Based on this context, it is possible to state that studies related to evaluating and maintaining ML models applied to health are of great relevance. Despite this, the literature in the area does not present many works that discuss the limitations and provide clear paths for the described problem. Thus, this article presents a Systematic Literature Review (SLR) on evaluating and monitoring real-world ML models for healthcare applications.

The review follows the Systematic Resistance methodology defined by Kitchenhan (Kitchenham & Charters, 2007) and reflects the current literature on evaluating and maintaining ML models in health. This type of study has limitations related to time, as it observes works published up to the date of their realization. On the other hand, it is an easily reproducible study since it is based on a formal literature review protocol.

It is important to mention that the works listed in the review were analyzed considering the entire life cycle of an ML model applied to a real context in the health area, which comprehends: performance evaluation, model monitoring, and maintenance. In addition, this work also presents, as a secondary objective, an approach for analyzing and evaluating the performance of ML models in health that will be proposed based on the results of the review and observations made.

The next sections are organized as follows: Section 2 will discuss related concepts; Section 3 presents review methodology; Section 4 analyzes outcomes from the systematic review; Section 5 presents the discussion and outlines a research proposal; and, finally, Section 6 discusses conclusions and future work.

2. Related Concepts

This work is related to the monitoring and evaluating of ML models in the healthcare context. In this sense, this section will briefly explain some important aspects for evaluation and continuous observation of the results and performance of an ML model.

2.1 ML Model Evaluation

Building a Machine Learning model involves the following steps: pre-processing, which includes data collection and handling; processing, which amounts to running ML methods over the pre-processed data; and post-processing, with model performance metric collection and analysis (Mitchell et al., 2007). Traditionally, post-processing includes testing, which means training the model over a data sample to collect performance metrics. Another activity, called validation, is usually performed after testing as part of the post-processing step. This activity involves verifying model performance against different data samples kept for that purpose specifically. After that, the model is serialized and embedded in its target application to fulfil its role in solving the proposed problem (Gopal, 2019).

This context delimits a problem. If model validation occurs before delivery and effective use against real-world data, can performance monitoring and evaluation in actual operation (in production) be called validation too? If so, how can one be differentiated from the other? Current literature seems to have little consideration for that matter. Validation and evaluation usually refer to both the final steps of building the model (post-processing) and evaluating that same model after it is effectively in use. That makes researching model monitoring and evaluation challenging, given the lack of consensus on terminology. In this research, validation, performance evaluation, monitoring, and maintenance refer to models already built and effectively in use, not those still under development.

2.2 Continuous Monitoring and Evaluation

There are considerable challenges to ML for Healthcare inherent to the clinical context. For instance: dealing with large volumes of data, data complexity, unstructured data, and patient privacy concerns, not to mention critical requirements regarding accuracy, since mistakes can result in life-threatening situations for patients. Those factors can become dealbreakers to ML model effectiveness and usefulness. Therefore, continuous monitoring and performance evaluation for Healthcare ML applications is a critical necessity.

Machine Learning Operations (MLOps), which adapts DevOps principles to ML model lifecycle, intends to manage the Intelligence Cycle for ML models so that people can work together to imagine, develop, deploy, operate, monitor, and improve machine learning systems on an ongoing basis (Treveil et al., 2020).

Getting models into production is just part of the process, not the end of it. Once a model is in operation, production data should be collected and monitored continuously to close the feedback loop. That way, new data can be selected and labelled into new training datasets and be used to improve ML models. That would allow models to adapt and improve continuously (Maleki et al., 2020).

Factors inherent to business and product aspects can affect ML models' lifecycle, such as implementation cost and model impact (Wiens et al., 2019). Misalignment between model and business metrics can lead to undesirable effects on model performance. A statistically accurate model that fails to meet business expectations is doomed to failure. Therefore, studies about continuous model monitoring and validation are essential. That is especially true in contexts such as ML for healthcare.

3. Methodology

According to Kitchenham and Charters (2007), a Systematic Review is a study that aims at identifying research works related to a specific topic and addresses broader questions regarding research evolution. Therefore, conducting a Systematic Literature Review (SLR) is a good fit for this work, which seeks to understand current state-of-the-art regarding healthcare model evaluation, monitoring, and maintenance. This process utilizes a quantitative approach to collect and organize the selected data and a qualitative analysis to compare the established quality criteria to understand the current model evaluation and monitoring landscape. The research process occurs in three stages: planning, execution, and data extraction, as detailed in the following sections.

3.1 Research Planning

The Systematic Literature Review begins with methodological planning to reduce errors and biases in study selection and analysis. Planning defines the research objective, questions, search engine, search string, inclusion, exclusion, and quality criteria. Those are necessary for the execution phase.

3.1.1 Research Objective and Research Questions

This review's main objective is to establish current state-of-the-art regarding healthcare model evaluation, monitoring, and maintenance. The following Research Questions (RQ) account for that:

- **RQ1**: Which methods and techniques evaluate machine learning models' performance in real-world applications?
- **RQ2**: What are their main characteristics, and how are they described?
- **RQ3**: Are there specificities for ML model evaluation in Healthcare applications?
- **RQ4**: How is model update handled considering system operation, and how does domain data quality assurance happen?
- **RQ5**: What are the main challenges and opportunities in evaluating ML models in healthcare applications?

3.1.2 Search Engine, Inclusion and Exclusion Criteria

Scopus search engine, from *Elsevier*, was chosen as the platform for the research, as it indexes the most relevant databases for the areas of computer science and machine learning, such as *ACM Digital Library*, *IEEE Explorer*, *Science Direct*, and *Springer Link*. The inclusion and exclusion criteria, which determine which studies should be included or excluded in a systematic review, were defined as follows.

- **Inclusion criteria**:
  - English-written studies only;
  - The studies must propose or analyze the evaluation process of machine learning models in healthcare applications.
- **Exclusion criteria**:
  - Grey literature (books, technical reports, non-scientific articles);
  - Duplicated results;
  - Same-author or same-research works;
  - Studies not related to healthcare;
  - Studies not related to Machine Learning;
  - Studies that do not address real-world operation;
  - Studies unavailable for download;
  - Studies that do not address any of the research questions;
  - Studies published prior to 2010.

3.1.3 Quality Criteria

The Quality Criteria (QC) evaluate the work's adherence to the research objective and research questions. In other words, research questions establish what should be investigated, and quality criteria objectively quantify how valuable the works are to the research. The following quality criteria were established:

- **QC1**: Does the work address the evaluation of machine learning models already in use in a real-world operation (i.e., in a "production environment")?
- **QC2**: Does the work clearly detail the evaluation procedure for one or more machine learning models in production?
- **QC3**: Are there any particularities related to the management of machine learning models in healthcare applications?
- **QC4**: Are data-related change management choices detailed along with their motivations?
- **QC5**: Are model management choices detailed along with their motivations?
- **QC6**: Are limitations and opportunities described for machine learning model evaluation in production?
- **QC7**: Does the work describe or propose a framework for production model evaluation in a structured and reproducible manner?
- **QC8**: Does the work go beyond statistical techniques for model evaluation, taking into account domain experts' opinions and/or specific protocols for the application area?

The measurement of the quality criteria for each work is made using a scale. After reading the work, each receives a score indicating how well they address each quality criterion. The following scale was used: 0, when it does not address the quality criterion; 0.5, when it partially meets the criterion; and 1.0, when it fully meets it.

According to Kitchenham and Charters (2007), a search string must be refined in an iterative process of trial, observation, and refactoring that aims at returning works as coherent as possible to the research subject. The search string was based on the research questions and keywords widely used in Machine Learning for Healthcare applications. The following search string resulted from that process:

"health" AND ("machine learning" OR "ML OPS" OR "MLOPS" OR "machine learning operation") AND ("continuous improvement" OR "continuous deployment" OR "continuous learning" OR "model drift" OR "data drift" OR "target drift" OR "concept drift" OR "model decay" OR "feedback loop" OR " model health" OR "machine learning health" OR "model validation" OR "model evaluation" OR "machine learning evaluation" OR "machine learning validation")

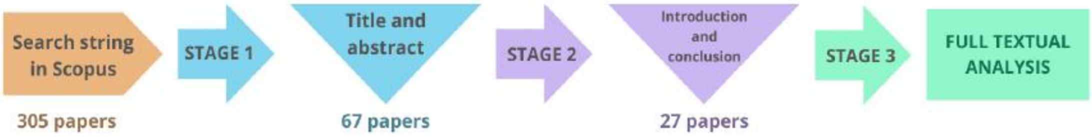
After defining the string, the search was performed in the chosen search engine, considering the works' title, abstract, and keywords. The collected data and notes referring to the stages of the research execution (to be described below) are available in an electronic spreadsheet accessible through the link: *https://bit.ly/3XktPfB*. Extracted data include the year of publication; work title; list of authors; keywords; work type; and link (URL).

3.2 Execution

The Systematic Review protocol followed in this research divides the execution into three successive stages: [1] Initially, the title and abstract of each work are read; [2] then, the introduction and conclusion of the selected ones are read; [3] and finally, the filtered ones deemed adherent to the research are read in full. The inclusion and exclusion criteria are observed during the readings of the first two stages. When an article does not meet all inclusion criteria or touches any exclusion criterion, it is removed and will not be read in the last stage. In the final stage, the articles remaining from stages 1 and 2 are fully read, and quality criteria are measured.

Figure 1 describes the search process. Two researchers analyzed each work for stages 1 and 2. To avoid bias, each researcher separately indicated whether the work should be excluded or kept for the final stage, based on inclusion and exclusion criteria. In case of disagreement, a consensual conversation between the researchers would define whether the article should remain. In the final stage, only one researcher per work was involved. Table 1 details the initial amount at each step, how many got removed, and how many remained.

Figure 1 - Revision protocol.



Source: Authors (2023).

Table 1 - Studies included and excluded at each stage.

	Input	Removed	Remaining
Stage 1	305	238	67
Stage 2	67	40	27
Stage 3	27	-	-

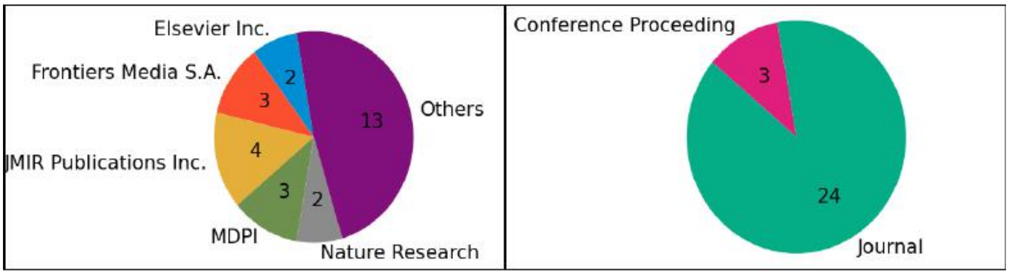
Source: Authors (2023).

4. Results

After the execution of the first two iterations (stages 1 and 2), twenty-seven (27) works got selected for a full reading. In stage 3, quality criteria evaluation took place for each. Research questions were then analyzed using quality measurements and data extracted from reading each work. This section details some of that analysis.

Stage 3 works got categorized according to their publisher. Figure 2 shows those on the left, making it clear that diverse publishers were involved. The right side of Figure 2 demonstrates a predominance of journals in terms of publication type, amounting to about 89% of the works read in full.

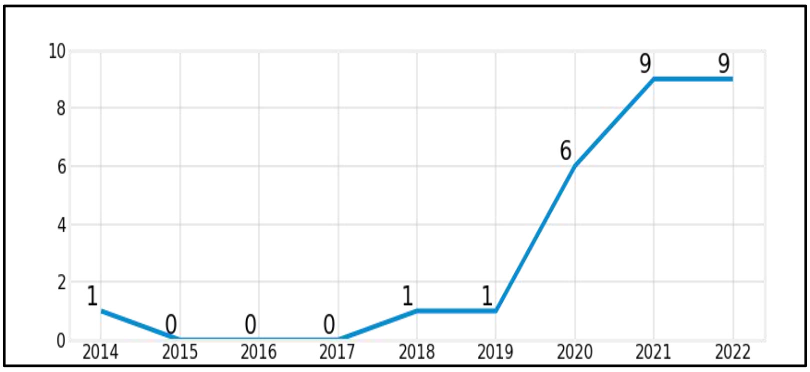
Figure 2 - Publishers and publication type.



Source: Authors (2023).

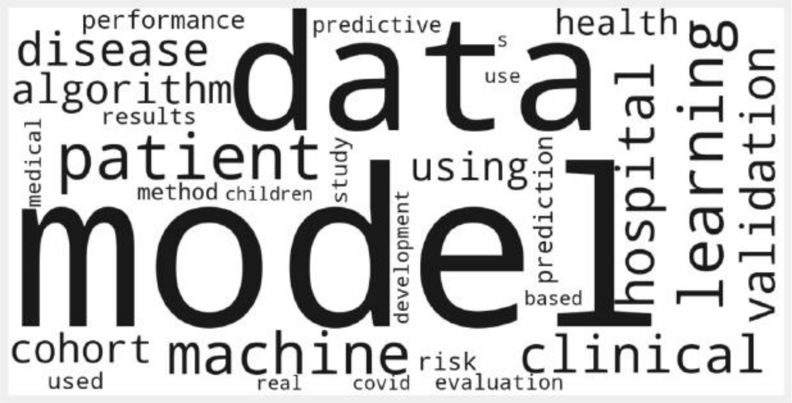
Figure 3 presents the distribution of the articles selected for full reading by year. Based on the figure, it is possible to infer that Model Validation for Healthcare applications has been gaining relevance, especially in the last three years, when a growing research effort related to the topic can be observed, demonstrating it's becoming a heated research topic. Based on the abstracts of the papers read in full, an additional word cloud chart was built, as shown in Figure 4, with the most cited terms within the abstracts (the more a word appears, the larger the text font becomes).

Figure 3 - Publications per year.



Source: Authors (2023).

Figure 4 - Word cloud.

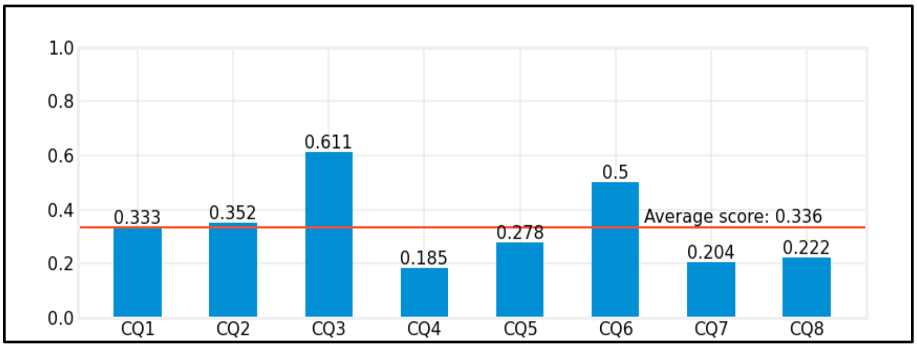


Source: Authors (2023).

By the values measured for the quality criteria, it is possible to observe, from the point of view of each criterion, how the articles read in full generally met the quality criteria. This visualization brings an important perspective on the maturity of the works in terms of each criterion.

Figure 5 presents the average values reached by articles read in full in each quality criterion. It is possible to observe that the overall average, drawn in an orange dashed line, has a value of 0.336 and that all criteria obtained averages below 0.7, with only two criteria reaching averages above 0.5.

Figure 5 - Quality criteria average.



Source: Authors (2023).

The list below presents the average values achieved by each quality criterion, followed by a brief discussion.

- **QC1**: the average achieved in this criterion was 0.333. That value indicates that evaluating and monitoring healthcare models in production have not been consistently approached by the works.
- **QC2**: the selected articles obtained an average of 0.352 for this criterion, which indicates that clarity and depth are lacking in the description of evaluation procedures for models in production.
- **QC3**: this was the highest average criterion amongst the works read, reaching an average of 0.611. This value indicates that they can identify particularities of ML for healthcare to some degree. Despite this, it is noted with this value that there are conditions for deepening the discussion on these particularities.
- **QC4**: unlike the previous criterion, the average value obtained by the works in this criterion was only 0.185, the lowest value amongst all quality criteria. With this result, it is possible to observe that the data-related change management decisions are reported hastily and can be significantly improved.
- **QC5**: in this criterion, the works reached an average of 0.278, denoting that the choices made for model management only are reported superficially.
- **QC6**: related to the limitations and opportunities in evaluating a model in production, the average score reached by the works, 0.5, indicates the addressing of such, but that there may still be a need for going deeper into this matter.
- **QC7**: works reached an average of 0.204 in this criterion. Thus, it is observable that research effort for establishing frameworks for evaluating ML models is limited.
- **QC8**: the last criterion presented 0.222 as an average obtained by the works. This value indicates that those have not prioritized the opinions of domain experts or used area-of-application-specific protocols for evaluating and monitoring models.

From the content analysis of the works read and the individual results of the quality criteria, it was possible to observe how each answered the Research Questions. Table 2 presents this detailing, marking with an “x” the research questions answered by each article. At the bottom of the table can also be seen a summary with the number of works that answered each Research Question. The table does not discriminate whether questions were answered satisfactorily or superficially, though. It only indicates whether that work approaches a Research Question.

Table 2 - Research questions touched by work.

Work	RQ1	RQ2	RQ3	RQ4	RQ5
Van Helvoort et al. (2020)			x		x
Carolan et al. (2022)			x		
Johri, Sen Saxena, & Kumar (2021)			x		
Lam et al. (2022)	x	x	x	x	x
Birkenbihl et al. (2020)			x	x	x
Wojtusiak (2021)	x	x		x	
Collin et al. (2022)			x		
Kamran et al. (2022)	x	x	x		x
Risman, Trelles, & Denning (2021)	x	x	x		x
Qasim et al. (2021)			x		
Sun et al. (2022)	x	x	x	x	
Shickel et al. (2020)	x	x	x		x
Bellocchio et al. (2021)			x		x
Sengupta et al. (2020)			x	x	
Rafiq, Modave, Guha, & Albert (2020)	x	x			x
Harris et al. (2022)		x	x	x	x
Maleki et al. (2020)					x
Vieira, Fernandes, Lucena, & Lifschitz (2021)				x	
The RADAR-CNS Consortium et al. (2021)	x	x	x		x
Huda et al. (2021)	x	x	x		
Li et al. (2022)	x	x	x	x	x
Lin et al. (2022)	x	x	x	x	x
Duckworth et al. (2021)	x	x			
Rojas et al. (2022)	x	x	x	x	x
Yang, Zou, Liu, & Mulligan (2014)					x
Iakovakis et al. (2018)	x	x	x		
Fries et al. (2019)		x	x		x
Total	**14**	**16**	**21**	**10**	**16**

Source: Authors (2023).

5. Discussion

This section presents a brief discussion of the review findings. It approaches some perspectives for each research question, using both results from the previous session and the content of the works read. Quality criteria measurements will also be used as a basis for the discussion since they came from the research questions.

Regarding **RQ1** which delimits an investigation into which methods and techniques are used to evaluate ML model performance in real-world applications. The average values obtained by the articles in quality criteria 1, 2, and 8, respectively 0.333, 0.352, and 0.222, indicating that detailing the techniques used to validate ML models in the real world is superficial. That becomes an even bigger issue in a context such as healthcare, where errors can lead to life-threatening situations for patients, which can end up being a barrier to machine learning adoption in clinical environments and overall healthcare contexts.

It is observable in the works that there is a lack of concrete data, metrics, and best practices for evaluating models in production, that is, ML models already deployed and in operation in real-world systems. Most of the articles reviewed only presented experimental reports, focusing mainly on the statistical evaluation of model performance during their construction, as is the case of (Van Helvoort et al., 2020; Johri et al., 2021; Qasim et al., 2021; Sun et al., 2022; Maleki et al., 2020). Some articles reported tests carried out in real-world environments with patients. However, they didn't detail their evaluation procedures on production models (Lam et al., 2022; Birkenbihl et al., 2020; Kamran et al., 2022; The RADAR-CNS Consortium et al., 2021). It's also noticeable that there is little information about the metrics and best practices for model evaluation in production for healthcare applications.

RQ1 analysis is highly related to **RQ2**, which deals with the characteristics of methods and techniques used to evaluate ML models in the real world. Therefore, given the scarcity of responses related to practices for evaluating ML models in production, there is little documentation on the characteristics of the methods and techniques used. Despite that, some works mention the need for special care in the statistical evaluation of the training data of the models. Especially when the groups that originate the training data (patients from a specific hospital or people from certain geographical regions, for instance) have distinct characteristics (data-wise), applying that same model to other groups can lead to low model performance (Sun et al., 2022; Rafiq et al., 2020). There are also comments about the need for specialized professionals to participate in model construction and validation to promote better reliability (Wojtusiak, 2021; Risman et al., 2021; Harris et al., 2022; Rojas et al., 2022). Specialists can help both in processing and making sense of the data, model performance testing, and defining evaluation methods, thus ensuring that the resulting models are accurate and reliable.

Another issue pointed out by some works is the need for good model interpretability (Rafiq et al., 2020; Harris et al., 2022; Li et al., 2022; Duckworth et al., 2021). ML model interpretability and explainability can help ensure that ML-enabled applications provide coherent and reliable decisions. Explainability is especially important in healthcare, as it allows the interpretation of model results and facilitates data collection for model evaluation or processes such as auditing. In this context, communication and collaboration also should be prioritized when validating machine learning models in production, corroborating the need for improvement and going deeper into this matter as the answers presented for quality criteria 1, 2, and 8 are superficial.

RQ3 searches for specificities of the evaluation process for healthcare ML models. It is directly related to QC3, in which works obtained an average of 0.611, the highest score among all quality criteria. It's noticeable when reading the articles that a relevant part of them mentions problems or specificities related to model evaluation in healthcare applications (Shickel et al., 2020; Rafiq et al., 2020; Rojas et al., 2022; Fries et al., 2019). One of the most critical issues mentioned is the need to keep data up-to-date to provide input for continuous and consistent updating of ML models. Therefore, it is necessary to establish metrics that can identify changes in data distribution and trigger model retraining when those are detected (Birkenbihl et al., 2020; Rojas et al., 2022).

A second aspect pertains to regulatory and ethical concerns, critical issues for ML model management in healthcare applications (Carolan et al., 2022; Wojtusiak, 2021). In healthcare, ethical and regulatory questions concerning data confidentiality, traceability, and explainability of (model) decision process were already strongly present long before the recent pushes for data access rights and data privacy laws by initiatives such as the General Data Protection Law (LGPD) in Brazil, or the California Consumer Privacy Act (CCPA) in the US, among others (Harris et al., 2022; Maleki et al., 2020; Rojas et al., 2022). Though these regulatory concerns are not specific to the healthcare context, they affect this area dramatically, given many of the best healthcare practices relate to the personalization of clinical decisions and the humanization of processes. Finally, although there is a reasonable discussion about the particularities relevant to the management of ML models in healthcare applications, there is only a superficial discussion about possible solutions to the problems faced by model management due to these particularities. That is, it is observable that the works describe existing problems but do not discuss structured solutions to them (or only do it superficially).

QC4 and QC5, in which the articles obtained averages (respectively) of 0.185 and 0.278, are tightly related to **RQ4**, which seeks to describe ways to update the ML model during system operation and the quality assumptions observed on the domain data. The values obtained for the QCs indicate that details on the decisions taken regarding model updates are scarce. It is worth mentioning that, given the critical performance requirements of healthcare applications, it is vital to understand how to manage ML model updates when input data distribution changes, concepts deviate, or the very model is no longer a feasible solution for the problem at hand (Vieira et al., 2021).

RQ5 and the related QC6 address the challenges and opportunities related to ML model evaluation and monitoring in healthcare applications. The works obtained an average of 0.500 in QC6. This value indicates some level of depth in discussing challenges and opportunities. Challenges mentioned include data obtention in real-time, data scarcity, maintenance of existing systems, quantifying the comparability of validation data (from new patients) against training data, data accessibility and continuity, standardization of models, data imbalance, and those about the clinical routine and specialist availability. For example, models trained on data derived from a single health institution may not generalize well on multi-institutional scenarios. A variation on this problem is patient selection biases (regional, socioeconomic, and institutional) (Van Helvoort et al., 2020; Carolan et al., 2022; Lam et al., 2022; Birkenbihl et al., 2020; Kamran et al., 2022; Risman et al., 2021; Shickel et al., 2020; Bellocchio et al., 2021; Rafiq et al., 2020; Harris et al., 2022; Maleki et al., 2020; The RADAR-CNS Consortium et al., 2021; Li et al., 2022; Lin et al., 2022; Rojas et al., 2022; Yang et al., 2021; Fries et al., 2019).

Such challenges may impact the feasibility of ML model evaluation and monitoring for healthcare applications. Despite that, the ongoing discussions about these topics can favor the emergence of approaches that can provide solutions or ways to mitigate risks, as well as new businesses and healthcare services. Other challenges are related to Continuous Learning in healthcare, which presents different limitations.

Regarding the opportunities presented in the selected works, there are mentions of the creation of international standards and guides to deal with the regulatory challenges of ML in healthcare applications. Carolan et al. (2022) describes the need for better automation technologies to improve the efficiency of algorithms. There are also opportunities for expert management and monitoring (Algorithmic Stewardship), with projections of the near-future creation of MLOps departments for healthcare services and hospitals (Harris et al., 2022). Other possibilities include integrating equity in the ML lifecycle, removing biases, as well as collecting *feedback* from experts and other stakeholders to bring human knowledge into the learning process (Human-in-the-Loop Learning), and going beyond statistical metrics in evaluating the model performance, using domain-oriented approaches to measure the usefulness and commercial value of these (Rojas et al., 2022; Yang et al., 2021). Finally, there are opportunities for real-world applications supported by live data where teams can iteratively build and test at the bedside, continuous delivery (CD) MLOps platforms, design and oversight by people with AI security expertise, continuous assessment using randomization to avoid bias, and use of data flows with the HL7-FHIR protocol (Harris et al., 2022).

Based on those observations, it is noticeable that there is a need for improvement and deepening of research related to ML model evaluation and monitoring in healthcare applications. QC7 searches for works that discuss and propose solutions for evaluating ML models in a structured and reproducible way. The general average in this criterion was 0.204. In addition, of the 27 articles read, only three (3) fully meet this criterion (Carolan et al., 2022; Kamran et al., 2022; Fries et al., 2019), which reinforces the need for research that defines, discusses, and improves the ML model's evaluation and maintenance methods, especially in critical applications such as healthcare. Therefore, the main observation for QC7 is the need for a methodological approach to ML model evaluating, monitoring, and maintaining in healthcare applications once in real-world operation (production).

6. Conclusions and Future Researches

This work presents the result of a systematic literature review that sought to understand the current state of Machine Learning model evaluation, monitoring, and maintenance in healthcare applications. Following Kitchenhan's protocol (Kitchenham & Charters, 2007), twenty-seven (27) papers underwent complete analysis. The gathered results and the discussions that ensued (presented in previous sections) indicate the need for further research involving ML model evaluation, monitoring, and maintenance in real-world healthcare applications. That said, reasonable documentation of problems and limitations is available, which can provide a starting point for future research.

The struggle to find studies that go beyond the experimental report and effectively evaluate ML models in real-world operation suggests that considerable emphasis has occurred on model construction and experimental validation. Though, continuity of these efforts does not seem to happen when models enter system operation. As a result, accounting for model operation on real-world data has not been consistently addressed. Healthcare applications demand continuous monitoring, validation, and maintenance of the models due to the very criticality of the domain and the services involved.

Therefore, although the importance of ongoing model evaluation and monitoring is acknowledged, the literature still needs practical studies and detailed methodologies for continuous ML model evaluation in healthcare applications. It is essential to continue researching and developing effective methods for evaluating, monitoring, and maintaining ML models to guarantee that they are safe, reliable, and useful for healthcare applications.

The results of the systematic review suggest the need for a change management workflow for developers and managers of ML models. This process, to be proposed in future work, should include the following activities: [1] Obtaining available documentation (for example, baseline model performance, experimental design decisions), [2] Definition of evaluation criteria and parameters based on expert opinion, real-world statistical performance of models (quantitative metrics), and product, business, and area-of-application-specific protocols (qualitative metrics); [3] Evaluation prototyping with business and domain specialists; [4] Operationalization and measurement criteria monitoring; [5] Evaluation of measurement criteria (for example, biases, drift, delayed results, statistical and business performance); and [6] model refactoring, which may include sub-activities such as [a] Sliding-window real-world data collection and storage; [b] Model training with real-world clinical data; [c] Statistical validation; [d] Hyperparameter tuning; [e] Model retraining whenever data distribution change; [f] Standardization of models.

Other future work could establish a methodological approach for assessing the level of maturity of ML models, once in real-world use, based on good practices and concerns that permeate the entire lifecycle of the models.

References

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