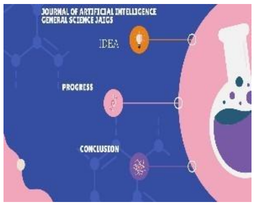
Vol. 1, Issue 1, January 2024

Journal of Artificial Intelligence General Science (JAIGS)
ISSN: 3006-4023
https://ojs.boulibrary.com/index.php/JAIGS



Machine Learning Applications in Healthcare: Current Trends and Future Prospects

Dr. José Gabriel Carrasco Ramírez[1] Md. Mafiqul Islam[2], ASM Ibnul Hasan Even[3],

[3]Lawer graduated at Universidad Católica Andrés Bello. Caracas. Venezuela. / CEO, Quarks Advantage. Jersey City, United States. / Director at Goya Foods Corp., S.A. Caracas. Venezuela
[2]Department of Information Science and Library Management, University of Rajshahi, Bangladesh
[3]Senior Assistant Secretary, Cabinet Division, Bangladesh Secretariat, Dhaka

Corresponding Author: **Md.Mafiqul Islam**

ARTICLE INFO

Article History:
Received:
05.01.2024
Accepted:
10.01.2024
Online: 02.02.2024

Keywords
Machinelearning, diagnis, treatment, healthcare.

ABSTRACT

The integration of machine learning (ML) in healthcare has witnessed remarkable advancements, transforming the landscape of medical diagnosis, treatment, and overall patient care. This article provides a comprehensive review of the current trends and future prospects of machine learning applications in the healthcare domain. The current landscape is characterized by the utilization of ML algorithms for disease diagnosis and risk prediction, personalized treatment plans, and efficient healthcare resource management. Notable applications include image recognition for radiology and pathology, predictive analytics for disease prognosis, and the development of precision medicine tailored to individual patient profiles. This review explores the evolving role of ML in improving patient outcomes, enhancing clinical decision-making, and optimizing healthcare workflows. It delves into the challenges faced in integrating ML into existing healthcare systems, such as data privacy concerns, interpretability of complex models, and the need for robust validation processes. Additionally, the article discusses future prospects and emerging trends in ML healthcare applications, including the potential for predictive analytics to preemptively identify health issues, the integration of wearable devices and remote monitoring for continuous patient care, and the intersection of ML with genomics for personalized medicine. The overarching goal of this article is to provide healthcare professionals, researchers, and policymakers with insights into the current state of ML applications in healthcare, along with an outlook on the transformative potential that machine learning holds for the future of healthcare delivery and patient outcomes.

Introduction

Artificial Intelligence (AI) has been a leading choice for resolving complicated issues in various fields of study in recent times. The use of AI techniques (such as machine learning, expert systems, deep learning, and others) to tackle ordinary issues has grown more and more commonplace.

Using Machine Learning (ML) techniques to tackle issues, various research have focused on healthcare, an area with significant social impact. During the COVID-19 Pandemic, for example, some studies used machine learning techniques to forecast patient outcomes (Malki et al., 2021; Arowolo et al., 2022). Some attempted to estimate the mortality risk for heart failure patients in intensive care units (Luo et al., 2022). The use of ML approaches in healthcare applications confronts obstacles, particularly in modeling, analysis, and validation, given the seriousness of the topics addressed (Ghassemi et al., 2020). In order to ensure that machine learning models are created to address actual problems in the field and are comprehensible and elucidable to the clinical community, resolving them necessitates strong collaboration between data scientists and healthcare professionals.

The data utilized for training and testing machine learning models has a direct impact on the models' performance and results (Gopal, 2019). It therefore becomes challenging to extrapolate findings from data from certain locales and patient characteristics to other contexts. Because healthcare services professionals have rigorous daily schedules, it might be challenging to incorporate the need for constant monitoring and specialist feedback, which further complicates the analysis and validation of model outcomes in healthcare applications. Conventional statistical analysis of data might not be as effective when it comes to

reduces the need for research and development in ML model evaluation and monitoring for healthcare applications, particularly when it comes to models that are in production and applied to scenarios that could imply the difference between a patient's life and death.

In light of this, it is reasonable to conclude that research on the assessment and upkeep of machine learning models used in the healthcare industry is highly pertinent. In spite of this, there aren't many works in the field that address the drawbacks and offer solutions for the given issue. As a result, a Systematic Literature Review (SLR) on assessing and tracking practical machine learning models for healthcare applications is presented in this article.

The review incorporates the most recent research on assessing and sustaining machine learning models in health and adheres to Kitchenhan's (Kitchenham & Charters, 2007) Systematic Resistance approach. Since this kind of analysis only looks at works that have been published up until the point of realization, it has restrictions in terms of time. However, because the study is based on a systematic literature review procedure, it may be readily replicated.

It is significant to note that the works included in the review were examined from the perspective of an ML model applied to a real-world setting in the health sector, taking into account model monitoring, maintenance, and performance assessment. A secondary goal of this study is to give a method for assessing and analyzing the effectiveness of machine learning models in the field of health. This method will be suggested in light of the findings of the review and the observations made.

The sections that follow are arranged as follows: Related ideas will be covered in Section 2; review technique will be covered in Section 3; results from the systematic review will be analyzed in Section 4; discussion and a research proposal will be presented in Section 5; and conclusions and future work will be covered in Section 6.a

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 ofan ML model

ML Model Evaluation

The process of creating a machine learning model consists of three stages: pre-processing, which is gathering and handling data; processing, which is applying machine learning techniques to the pre-processed data; and post- processing, which is gathering and analyzing model performance metrics (Mitchell et al., 2007). Post-processing typically involves testing, which is the act of training the model on a sample of data in order to gather performance indicators. As part of the post-processing phase, validation is a different task that is typically carried out following testing. Verifying model performance against various data samples stored for that purpose alone is the task at hand. Subsequently, the model undergoes serialization and embedding within its intended application to effectively address the problem at hand (Gopal, 2019).

A problem is delimited by its context. Can performance monitoring and evaluation during real-world operation (in production) be considered validation as well, if validation of the model takes place prior to delivery and effective usage against real-world data? If so, what distinguishes one from the other? It appears that current literature gives the issue little thought. Generally speaking, validation and evaluation refer to both the last stages of model construction (post-processing) and the assessment of the same model following its implementation. This makes it difficult to do research on model monitoring and assessment because there is disagreement over terminology. Validation, performance evaluation, monitoring, and maintenance in this research pertain to models that have already been constructed and are being used successfully, rather than ones that are still in the creation stages.

Continuous Monitoring and Evaluation

Because of the unique clinical setting, machine learning presents many obstacles for the healthcare industry. Examples include handling massive amounts medical data, handling complex data, handling unstructured data, and managing patient privacy issues. Accuracy is also vital because errors can put patients in danger of dying. These elements have the potential to completely undermine the use and efficacy of ML models. Consequently, it is imperative that Healthcare ML applications have ongoing monitoring and performance review.

In order to enable collaboration amongst individuals to envision, develop, implement, run, monitor, and continuously improve machine learning systems, Machine Learning Operations (MLOps), which applies DevOps principles to the lifecycle of ML models, aims to manage the Intelligence Cycle for ML models (Treveil et al., 2020).

Putting models into production is not the end of the process; it is only the beginning. To close the feedback loop, production data should be gathered and tracked continually once a model is in use. New data can then be chosen and labeled into fresh training datasets, which can then be utilized to enhance machine learning models. According to Maleki et al. (2020), this would enable models to continuously adjust and get better.

The lifecycle of ML models can be impacted by factors that are intrinsic to business and product features, such as model impact and implementation cost (Wiens et al., 2019). Unwanted consequences on model performance may result from a mismatch between the model and business metrics. If a statistically correct model doesn't perform up to business expectations, it will fail. Consequently, research on ongoing model validation and monitoring is crucial. This is particularly valid in scenarios like machine learning for healthcare.

Methodology

A systematic review, as defined by Kitchenham and Charters (2007), is an investigation that seeks to locate research works pertaining to a certain subject and tackles more general inquiries about the progress of research. Thus, this effort, which aims to comprehend the present state-of-the-art related healthcare model evaluation, monitoring, and maintenance, is a good fit for conducting a Systematic Literature Review (SLR). In order to grasp the present model assessment and monitoring landscape, this procedure compares the defined quality standards using a qualitative analysis and a quantitative approach to gather and organize the chosen data. Planning, carrying out, and extracting data are the three phases of the research process, as explained in the parts that follow.

Research Planning

Methodological planning is the first step in the Systematic Literature Review to minimize biases and errors in the selection and analysis of studies. Research objectives, questions, search engine, search string, inclusion, exclusion, and quality standards are all defined during planning. These are essential for the carrying out stage.

Research Objective

The primary goal of this review is to define the state-of-the-art in terms of healthcare model assessment, upkeep, and monitoring. That is explained by the Research Questions (RQ) that follow:

- RQ1: What are the approaches and strategies used to assess the effectiveness of machine learning models in practical settings?
- RQ2: How are their primary attributes articulated, and what are they?
- RQ3: Are there any particulars when evaluating ML models for applications in healthcare?
- RQ4: How does domain data quality assurance take place, and how is model updating managed in light of system operation?

● RQ5: What are the primary obstacles and prospects associated with assessing machine learning models for healthcare applications?

Search Engine, Content and Requirements for Exclusion

Because it indexes the most pertinent resources for computer science and machine learning, including ACM Digital Library, IEEE Explorer, Science Direct, and Springer Link, Elsevier's Scopus search engine was selected as the research platform. The following definitions of the inclusion and exclusion criteria were used to decide which studies should be included or excluded in a systematic review.

Selection standards:

Only research written in English are acceptable. The studies must either suggest or examine how machine learning models are evaluated for use in healthcare settings.

● Exclusion criteria: ○ Grey literature (books, technical reports, non-scientific articles); ○ Duplicate results; ○ Same-author or same-research works; ○ Studies unrelated to healthcare; ○ Studies unrelated to machine learning; ○ Studies that don't deal with real-world operations; ○ Studies that can't be downloaded; ○ Studies that don't address any of the research questions; ○ Studies published before 2010.Standards of Quality

The conformance of the work to the study objectives and research questions is assessed by the Quality Criteria (QC). Stated differently, research questions determine what needs to be explored, while quality standards measure the contributions of the works to the field objectively. The subsequent standards of quality were instituted:

● QC1: Does the work deal with evaluating machine learning models that are already being used in an actual operation (a "production environment")?

● QC2: Does the work explicitly describe how one or more machine learning models in production are evaluated?

● QC3: Are there any unique aspects to the administration of machine learning models in applications for healthcare?

● QC4: Are data-related change management decisions explained in detail, including the reasons behind them?

● QC5: Are model management decisions explained in detail, including their justifications?

QC6: Are opportunities and constraints for evaluating machine learning models in production described?

QC7: Is a systematic and repeatable framework for production model evaluation described or proposed in the work?

QC8: Does the study consider the opinions of domain experts and/or particular procedures for the application area, going beyond statistical methodologies for model evaluation?

A scale is used to measure the quality criteria for every job. Each is given a score based on how well they meet each quality requirement after having their work read. The quality criterion was rated on a scale of 0 for not meeting the criterion, 0.5 for meeting it partially, and 1.0 for fully meeting it.

In order to provide works that are as coherent as feasible to the research subject, Kitchenham and Charters (2007) contend that a search string needs to be developed through an iterative process of trial and error, observation, and reworking. Based on popular terms and research issues in Machine Learning for Healthcare applications, the search string was created. That procedure produced the search string that follows:

"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")

Following the string's definition, the works' title, abstract, and keywords were taken into consideration when conducting the search in the selected search engine. An electronic spreadsheet with the gathered information and notes about the phases of the research execution (to be detailed below) can be accessed at this link: bit.ly/3XktPfB. The year of publishing, the work's title, an author list, keywords, the work type, and a link are among the extracted data.

Execution

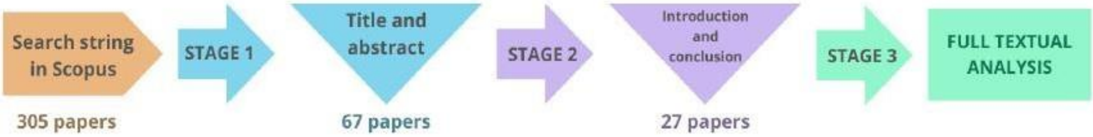
The implementation of this research is divided into three steps by the Systematic Review methodology that was followed: [1] First, each work's title and abstract are read; [2] then, the chosen works' introduction and conclusion are read; [3] and lastly, the entire texts of the filtered works that are thought to be consistent with the research are read. Readings of the first two stages are conducted with consideration for the inclusion and exclusion criteria. An article is eliminated and won't be read in the final stage if it doesn't satisfy all inclusion criteria or touches any exclusion criteria. The last step involves reading every article from phases 1 and 2 in its entirety and measuring the quality standards. Figure 1 outlines how to do a search. Every work was examined by two researchers for stages 1 and 2. In order to prevent bias, each researcher independently decided, based on inclusion and exclusion criteria, whether the work should be retained for the final stage or excluded. If the researchers couldn't agree, they would have a cooperative discussion to decide whether the paper should stay. There was only one researcher working on each project in the end. Table 1 shows the starting number at each stage, the number that was eliminated, and the number that was left.

The implementation of this research is divided into three steps by the Systematic Review methodology that was followed: [1] First, each work's title and abstract are read; [2] then, the chosen works' introduction and conclusion are

read; [3] and lastly, the entire texts of the filtered works that are thought to be consistent with the research are read. Readings of the first two stages are conducted with consideration for the inclusion and exclusion criteria. An article is eliminated and won't be read in the final stage if it doesn't satisfy all inclusion criteria or touches any exclusion criteria. The last step involves reading every article from phases 1 and 2 in its entirety and measuring the quality standards. Figure 1 outlines how to do a search. Every work was examined by two researchers for stages 1 and 2. In order to prevent bias, each researcher independently decided, based on inclusion and exclusion criteria, whether the work should be retained for the final stage or excluded. If the researchers couldn't agree, they would have a cooperative discussion to decide whether the paper should stay. There was only one researcher working on each project in the end. Table 1 shows the starting number at each stage, the number that was eliminated, and the number that was left.

Table 1 - Studies included and excluded at each stage

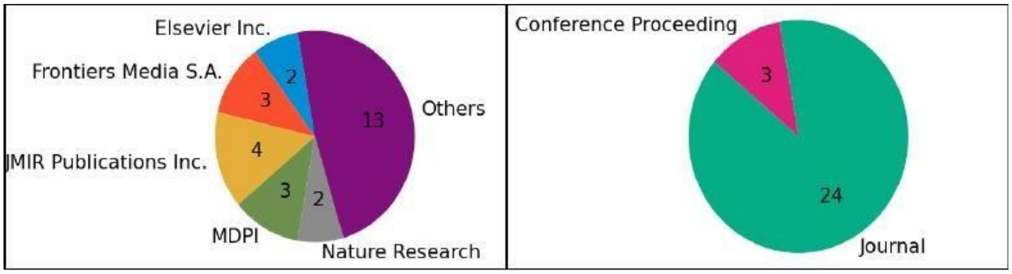
	Input	Removed	Remaining
Stage 1	305	238	67
Stage 2	67	40	27
Stage 3	27	-	-



Results

Following the completion of phases 1 and 2, a total of twenty-seven (27) works were chosen for a comprehensive reading. Each of the quality criteria was evaluated in stage 3. Following that, research topics were examined utilizing data gleaned from reading each work and quality measurements. Some of such analysis is presented in this section.

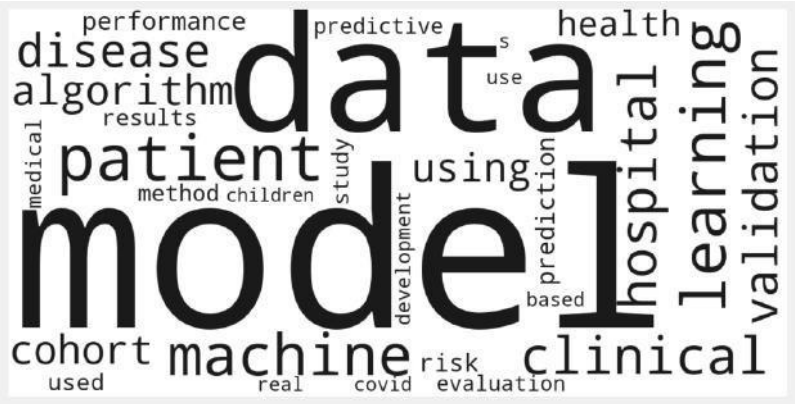
Works in Stage 3 were grouped based on their publisher. Figure 2 illustrates those on the left, demonstrating the involvement of a variety of publishers. Regarding publishing type, Figure 2's right side shows that journals account for the majority of publications read in full—roughly 89%.



The distribution of the articles chosen for a thorough reading by year is shown in Figure 3. The statistic suggests that Model Validation for Healthcare Applications has been becoming more relevant, particularly in the last three years, as evidenced by the increasing amount of research being done on the subject, indicating that it's becoming a hot topic. An additional word cloud chart was constructed using the most cited phrases found in the abstracts of the publications that were read in full. As a word appears more frequently, the text font size increases, as seen in Figure 4.

Figure 3 - Publications per year

Figure Word Cloud



From the perspective of each criterion, it is possible to examine how the articles that are read in their entirety generally met the quality criteria based on the values assessed for the criteria. This graphic offers a crucial viewpoint on the works' maturity with respect to each criterion.

The average values attained by articles read in their whole for each quality criterion are shown in Figure 5. It is evident that all criteria received averages below 0.7, with just two criteria reaching averages above 0.5, and that the overall average, represented by an orange dashed line, has a value of 0.336.

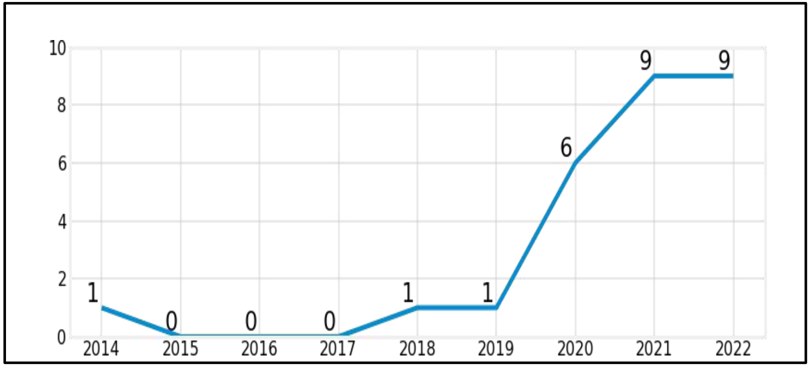
Figure 5 - Quality criteria average

The average values attained by each quality criterion are listed below, with a brief description that follows.

● QC1: Achieved an average of 0.333 in this criterion. This number suggests that the works have not taken a consistent strategy to assessing and tracking healthcare models in use.

● QC2: the average score for the chosen articles was 0.352 for this criterion, indicating a lack of depth and clarity in the description of the evaluation processes for operational models.

● QC3: with an average score of 0.611, this criteria had the highest average across the works read. Their ability to recognize the unique characteristics of machine learning for healthcare is indicated by this value. In spite of this, it is acknowledged with this value that further debate of these particularities can only take place under certain circumstances.



● QC4: in contrast to the preceding criterion, the works in this criterion obtained an average value of only 0.185, which was the lowest value of all the quality criteria. This outcome makes it feasible to see that data-related change management decisions are reported quickly and have a lot of room for improvement.

● QC5: The works only received an average score of 0.278 for this criterion, indicating that the decisions made for model management are only briefly communicated.

● QC6: Concerning the opportunities and constraints for assessing a model in production, the works' average score of 0.5 suggests that these have been addressed, though further research may still be necessary.

QC7: The average score for the works met this criterion, which was 0.204. It is evident, then, that there has been little research done to create frameworks for assessing machine learning models.

● QC8: the final criterion, with an average of 0.222 that the works produced. This value shows that individuals have not employed methods specific to the field of application for model evaluation and monitoring, nor have they given domain experts' opinions precedence.

It was feasible to see how each of the works read addressed the research questions based on the content analysis of those works and the unique outcomes of the quality criteria. This information is displayed in Table 2, where each article's response to each research question is indicated with a "x." Additionally, a summary of the number of papers that addressed each research question is shown at the bottom of the table. However, the table does not distinguish between superficial and satisfactory answers to questions. It just states whether or not a research question is addressed in that study.

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).

Discussion

The review findings are briefly discussed in this section. It uses the information from the read works as well as the outcomes from the prior session to approach several viewpoints for each research issue. Since the measurements for the quality standards originated from the research questions, they will also serve as the foundation for the discussion.

With reference to RQ1, which establishes the parameters of an inquiry into the strategies and tactics employed in assessing the performance of ML models in practical settings. The articles in quality criterion 1, 2, and 8 obtained average scores of 0.333, 0.352, and 0.222, respectively. These values suggest that the information provided about the methods used to evaluate machine learning models in real-world scenarios is superficial. This becomes much more problematic in the healthcare industry, where mistakes can put patients in danger of death. As a result, machine learning adoption in clinical settings and in the healthcare industry as a whole may be hampered.

It is evident from the works that there is a deficiency of specific metrics, data, and best practices for assessing models in production—that is, machine learning models that have already been implemented and are functioning in actual systems. Only experimental reports were provided by the majority of the examined studies, which mostly concentrated on the statistical assessment of model performance at the time of model creation (Van Helvoort et al., 2020; Johri et al., 2021; Qasim et al., 2021; Sun et al., 2022; Maleki et al., 2020). A few articles described patient tests conducted in real-world settings. The RADAR-CNS Consortium et al., 2021; Kamran et al., 2022; Lam et al., 2022; Birkenbihl et al., 2020; did not, however, provide specifics about their review processes for production models. It is also apparent that there is a dearth of knowledge on measurements and industry best practices for model evaluation in real-world healthcare applications.

RQ2 analysis, which addresses the features of approaches and strategies used to assess machine learning models in practice, is closely tied to RQ1 analysis. Therefore, there is minimal documentation on the features of the approaches and techniques used, given the dearth of responses pertaining to processes for testing ML models in production. Nevertheless, several studies point out that extra caution should be taken while evaluating the models' training data statistically. Low model performance can result from applying the same model to other groups, particularly when the groups that provide the training data (e.g., patients from a particular hospital or individuals from a particular geographic area) have different characteristics (data-wise) (Sun et al., 2022; Rafiq et al., 2020). In order to improve reliability, remarks have also been made regarding the necessity of having experts involved in the development and validation of models (Wojtusiak, 2021; Risman et al., 2021; Harris et al., 2022; Rojas et al., 2022). Experts can assist with data processing and interpretation, model performance testing, and evaluation technique definition, assuring the accuracy and dependability of the final models.

The requirement for strong model interpretability has also been brought up by various works (Rafiq et al., 2020; Harris et al., 2022; Li et al., 2022; Duckworth et al., 2021). Applications using machine learning (ML) can make conclusions that are consistent and trustworthy by ensuring that the models are interpretable and explainable. Explainability is particularly crucial in the healthcare industry since it makes it easier to analyze model output and gather information for procedures like auditing or model review. Since the responses provided for quality criteria 1, 2, and 8 are cursory, it is important to dive deeper into this issue and confirm the need for improvement through collaboration and communication when verifying machine learning models in production.

RQ3 looks for particulars in the way healthcare machine learning models are evaluated. It has a direct bearing on QC3, where works received the highest possible score of 0.611 out of all the quality criteria. Upon reading the publications, it is evident that a significant portion of them address issues or details pertaining to model evaluation in healthcare applications (Shickel et al., 2020; Rafiq et al., 2020; Rojas et al., 2022; Fries et al., 2019). Among the most important points raised is the necessity of maintaining current data in order to supply input for ML models that are updated continuously and consistently. Consequently, it's essential to establish measures capable of detecting shifts in the distribution of data and, upon detection, initiating retraining of the model (Birkenbihl et al., 2020; Rojas et al., 2022).

Another facet is ethical and regulatory matters, which are crucial for managing machine learning models in healthcare applications (Carolan et al., 2022; Wojtusiak, 2021). Long before the recent push for data access rights and data privacy laws by initiatives like the General Data Protection Law (LGPD) in Brazil or the California Consumer Privacy Act

(CCPA) in the US, among others, ethical and regulatory questions concerning data confidentiality, traceability, and explainability of (model) decision process were already strongly present in the healthcare industry (Harris et al., 2022; Maleki et al., 2020; Rojas et al., 2022). Despite not being unique to the healthcare setting, these regulatory issues have a significant impact on this field because many of the greatest practices in healthcare include humanizing procedures and personalizing clinical judgments. Lastly, while the particularities pertinent to the management of machine learning models in healthcare applications are reasonably discussed, there is a scant attention to potential remedies for the issues raised by these particularities in model management. That is to say, it is apparent that while the books outline current issues, they do not address organized answers (or just address them briefly).

The articles in QC4 and QC5, which received averages of 0.185 and 0.278, respectively, are closely associated with RQ4, which aims to explain how to update the ML model while the system is operating and the quality assumptions noted on the domain data. The results collected for the QCs show that there is a lack of information on the decisions made about model changes. Notably, considering the crucial performance demands of healthcare applications, it is essential to comprehend the management of machine learning model updates when the distribution of input data shifts, ideas diverge, or the model itself is no longer a workable solution for the issue at hand (Vieira et al., 2021).

The issues and possibilities surrounding the assessment and tracking of machine learning models in healthcare applications are covered by RQ5 and the associated QC6. In QC6, the works received an average score of 0.500. This number suggests that opportunities and difficulties have been discussed in some detail. Real-time data acquisition, data scarcity, data maintenance, comparing training data to validation data (from new patients), data continuity and accessibility, model standardization, data imbalance, and issues with clinical routine and specialist availability are some of the challenges that have been mentioned. Models developed using data from a single healthcare facility, for instance, could not perform well in scenarios involving multiple institutions. Patient selection biases (regional, socioeconomic, and institutional) represent a variation on this issue (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).

The viability of ML model monitoring and evaluation for healthcare applications may be impacted by these difficulties. Nevertheless, continuing conversations about these subjects may encourage the formation of new companies and healthcare services, as well as approaches that can offer answers or strategies to reduce dangers. Additional difficulties arise from the healthcare industry's use of continuous learning, which has various restrictions.

The establishment of worldwide standards and guidelines to address the regulatory problems of machine learning in healthcare applications is mentioned in relation to the opportunities given in the chosen works. According to Carolan et al. (2022), more advanced automation technologies are required to increase algorithmic efficiency. With predictions of the establishment of MLOps departments for healthcare services and hospitals in the near future, there are also chances for expert management and monitoring (Algorithmic Stewardship) (Harris et al., 2022). Other options include going beyond statistical metrics in evaluating the model performance, using domain-oriented approaches to measure the usefulness and commercial value of these; integrating equity into the ML lifecycle; eliminating biases; and gathering feedback from experts and other stakeholders to bring human knowledge into the learning process (Human-in-the-Loop Learning) (Rojas et al., 2022; Yang et al.,

2021). The use of data flows with the HL7-FHIR protocol, continuous delivery (CD) MLOps platforms, design and supervision by AI security experts, continuous assessment using randomization to prevent bias, and real-world applications supported by live data where teams can iteratively build and test at the bedside are all possibilities (Harris et al., 2022).

These results highlight the necessity for further development and improvement of the research on ML model monitoring and evaluation in healthcare applications. QC7 looks for publications that address and provide methods for conducting an organized and repeatable evaluation of machine learning models. In this criterion, the overall average was 0.204. Furthermore, out of the 27 articles that were read, only three (3) completely satisfied this criterion (Carolan et al., 2022; Kamran et al., 2022; Fries et al., 2019). This highlights the necessity for more research that clarifies, explores, and enhances the methods for evaluating and maintaining machine learning models, particularly in crucial applications like healthcare. The primary takeaway for QC7 is that, once in real-world operation (production), ML model evaluation, monitoring, and maintenance in healthcare applications require a methodical approach.

Conclusions and Upcoming Studies

The goal of this work was to understand the current status of machine learning model evaluation, monitoring, and maintenance in healthcare applications by a systematic review of the literature. The protocol developed by Kitchenhan and Charters (2007) was followed in the comprehensive analysis of twenty-seven (27) publications. The collected data and the ensuing debates (described in earlier sections) suggest that more research is required to evaluate, monitor, and maintain machine learning models in practical healthcare settings. Nevertheless, there is adequate documentation of issues and constraints that can serve as a foundation for further study.

It appears that a great deal of focus has been placed on model creation and experimental validation because it is difficult to locate research that go beyond the experimental report and efficiently evaluate ML models in real-world operation. However, it appears that when models are included into system operations, these efforts do not continue. Consequently, the issue of accounting for model activity on actual data has not received consistent attention. Because the healthcare industry and the services it provides are so vital, models must be continuously monitored, validated, and maintained.

Thus, even if the literature recognizes the value of continuous model evaluation and monitoring, there is still a need for empirical research and thorough approaches for continuous ML model evaluation in healthcare applications. To make sure that ML models are secure, dependable, and practical for healthcare applications, research into and development of efficient techniques for assessing, tracking, and maintaining ML models is vital.

The systematic review's findings point to the necessity of a change management procedure for ML model creators and managers. The following tasks ought to be part of this procedure, which will be suggested in later work: [1] acquiring the documentation that is currently available (e.g., baseline model performance, experimental design decisions); [2] defining evaluation criteria and parameters based on professional judgment, real-world statistical model performance (quantitative metrics), and protocols specific to a product, business, or area of application (qualitative metrics); Evaluation prototyping with business and domain specialists is step three. Operationalization and monitoring of measurement criteria is step four. Evaluation of measurement criteria (e.g., biases, drift, delayed results, statistical and business performance) is step five. Finally, model refactoring, which may involve sub-activities like sliding-window real-world data collection and storage, model training with real-world clinical data, statistical validation, hyperparameter tuning, model retraining whenever data distribution changes, and model standardization, are some of the sub-activities that may be included in model refactoring. Further research could develop a methodological framework based on best practices and considerations that pervade the model's whole lifetime for evaluating the maturity of machine learning models once they are put to use in real-world scenarios.

References

[1] Wu, K., & Chen, J. (2023). Cargo operations of Express Air. Engineering Advances, 3(4), 337–341. https://doi.org/10.26855/ea.2023.08.012

[2] Wu, K. (2023). Creating panoramic images using ORB feature detection and RANSAC-based image alignment. Advances in Computer and Communication, 4(4), 220–224. https://doi.org/10.26855/acc.2023.08.002

[3] Liu, S., Wu, K., Jiang, C. X., Huang, B., & Ma, D. (2023). Financial Time-Series Forecasting: towards synergizing performance and interpretability within a hybrid machine learning approach. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2401.00534

[4] Wu, K., & Chi, K. (2024). Enhanced E-commerce Customer Engagement: A Comprehensive Three-Tiered Recommendation System. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 2(3), 348-359. https://doi.org/10.60087/jklst.vol2.n2.p359

[5] hasan, M. R. (2024). Revitalizing the Electric Grid: A Machine Learning Paradigm for Ensuring Stability in the U.S.A. *Journal of Computer Science and Technology Studies*, 6(1), 142-154. https://doi.org/10.32996/jcsts.2024.6.1.15

[6] MD Rokibul Hasan, &Janatul Ferdous. (2024). Dominance of AI and Machine Learning Technique in Hybrid Movie Recommendation System Applying Text-to-number Conversion and Cosine Similarity Approaches. Journal of Computer

Science and Technology Studies, 6(1), 94-102. https://doi.org/10.32996/jcsts.2024.6.1.10

[7]. Naveen Vemuri, N. V. (2024). DevOps for Telehealth Services: Accelerating Deployment and Scalability. *International Journal on Recent and Innovation Trends in Computing and Communication*, *12*(1), 160-163. **DOI:** https://doi.org/10.17762/ijritcc.v12i1.9894

[8] Naveen Vemuri, Naresh Thaneeru, and Venkata Manoj Tatikonda. "Ai-optimized Devops for Streamlined Cloud CI/CD". Ai-optimized Devops for Streamlined Cloud CI/CD 9, no. 2 (February 17, 2024): 7. **https://doi.org/10.5281/zenodo.10673085**.

[7] Msekelwa, P. Z. (2023). Beyond The Borders Global Collaboration in Open Distance Education through Virtual Exchanges. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 2(2), 1-13. https://doi.org/10.60087/jklst.vol2.n2.p12

[8] Msekelwa, P. Z. (2023). DATA DRIVEN PEDAGOGY: LEVERAGING ANALYTICS FOR EFFECTIVE E-LEARNING STRATEGIES. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 1(1), 55-68. https://doi.org/10.60087/jklst.vol1.n.p12

[9] Ahmed, M. T., Islam, M., & Rana, . M. S. . (2023). Climate Change and Environmental Security in Bangladesh: A Gender Perspective . Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 1(1), 18-24. https://doi.org/10.60087/hckggn20

[10] Islam, M., & Rana, M. S. (2023). CONTAMINANT IDENTIFICATION IN WATER BY MICROBIAL BIOSENSORS: A REVIEW. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 1(1), 25-33. https://doi.org/10.60087/jgrkv103

[11] slam, M. (2023). BRIEF REVIEW ON ALGAE BASED BIOFUEL. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 1(1), 46-54. https://doi.org/10.60087/7xz85292