A process for evaluating Machine Learning models for Healthcare Applications

https://doi.org/10.56238/sevened2024.001-049

Cezar Miranda Paula de Souza[1], Cephas Alves da Silveira Barreto[2], Bruna Alice Oliveira de Brito[3], Victor Vieira Targino[4], Ramon Santos Malaquias[5], Itamir de Morais Barroca Filho[6] and Amanda Gomes de Oliveira Pereira[7]

ABSTRACT

Machine Learning (ML) has been advancing in the most diverse areas of knowledge. Among those, ML for Healthcare and Decision Support of Medical applications present specific challenges in evaluating, monitoring, and maintaining ML models. Once deployed, models are subject to performance degradation (drift). Therefore, continuous monitoring and evaluation are essential to establish minimum performance guarantees over unknown, real-world data. In medical applications, incorrect decisions can lead to life-threatening situations and irreversible outcomes. The present work proposes a process for ML model evaluation designed for Healthcare applications running on real-world data. To that end, a conducted Systematic Literature Review (SLR) aimed at determining the state-of-the-art techniques and methods for ML evaluation (detailed in another paper) and a case study applied the proposed process to ML models in an oncology ICU. The Case Study produced positive outcomes in establishing a feedback loop for models in use against real-world data.

[1] Teacher
Federal University of Rio Grande do Norte, Brazil
E-mail: cezarmiranda@gmail.com
ORCID: https://orcid.org/0009-0005-7189-8115
[2] Doctor
Federal University of Rio Grande do Norte, Brazil
E-mail: cephasax@gmail.com
ORCID: https://orcid.org/0000-0002-4756-8571
[3] Graduate
Federal University of Rio Grande do Norte, Brazil
E-mail: brna.oliveira03@gmail.com
ORCID: https://orcid.org/0009-0001-8116-495X
[4] Graduate
Federal University of Rio Grande do Norte, Brazil
E-mail: victorvieira.rn@gmail.com
ORCID: https://orcid.org/0000-0002-9036-6537
[5] Master
Federal University of Rio Grande do Norte, Brazil
E-mail: ramonstmalaquias@gmail.com
ORCID: https://orcid.org/0000-0002-8350-2836
[6] Doctor
Federal University of Rio Grande do Norte, Brazil
E-mail: itamir.filho@imd.ufrn.br
ORCID: https://orcid.org/0000-0003-1694-8237
[7] Institution: Federal University of Rio Grande do Norte (UFRN);
Institution affiliation: Undergraduate in Law and Student in the Postgraduate Program in Science, Technology and Innovation;
ORCID: https://orcid.org/0000-0003-4771-6754
E-mail: amandagomesop@gmail.com

Keywords: ML Evaluation, ML Monitoring, Continuous Monitoring, Feedback-loop, Drift, ML for Healthcare.

INTRODUCTION

Until a few years ago, academia perceived Machine Learning (ML) (and Artificial Intelligence (AI) as a whole) as a theoretical field, with applications only on curious little problems, challenging but of little practical value (Faceli et al., 2011). Nowadays, ML is an emerging topic in almost every knowledge field, with applications as diverse as financial forecasting, health monitoring, and human behavior recognition, among many others. Data science teams have shown that ML can offer many ways to improve efficiency, automate processes, reduce costs, and enhance the customer experience (Windheuser and Sato, 2021).

Bringing these experiments to practical use presents real challenges in itself, though. Deploying ML models in IT infrastructure and using them in real-world scenarios presents significantly different challenges, especially when real-world data can change rapidly (Mohandas, 2021). Among the current application areas, one in particular that stands out as a hotspot is Machine Learning for Healthcare and Software for Decision Support of Medical applications. Healthcare applications present specific challenges for ML model evaluation, monitoring, and maintenance, which are inherent to the clinical context, such as large volumes of complex, unstructured, or annotated data, concerns about the privacy of patient data, and critical requirements in terms of accuracy (Bin Rafiq et al., 2020).

Healthcare applications demand monitoring, evaluating, and continuously improving ML models as models must perform as expected. In most Healthcare applications, an incorrect prediction or classification can lead to life-threatening situations or undesired, irreversible outcomes.

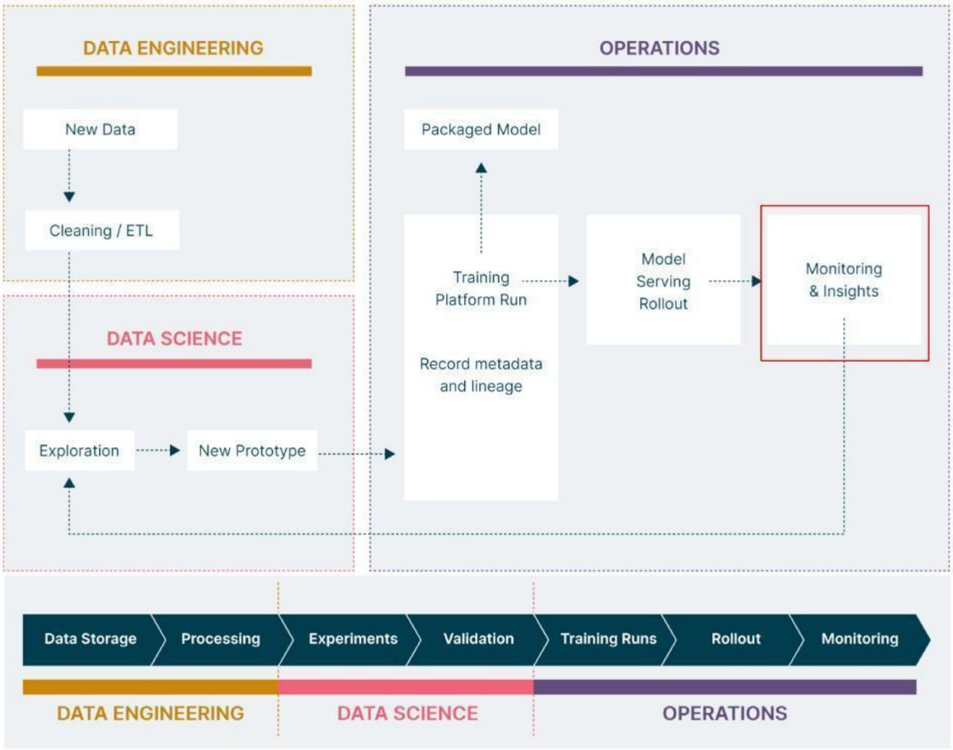
With the growing interest in ML, it is essential to understand the methodologies used to evaluate ML models to obtain reproducible solutions that can be successful in real-world environments (Maleki et al., 2020). In this context, MLOps, or Machine Learning Operations (which concerns itself with ML Models' lifecycle), is a discipline of Machine Learning that aims at managing the Intelligence Cycle for ML models (Windheuser & Sato, 2021). It encompasses concerns with monitoring and performance evaluation as well as change management.

In real-world ML systems, many sources of technical debt relate to data dependencies, model complexity, reproducibility, testing, monitoring, and dealing with changes in the external world (Sato et al., 2019). Changes to the ML models and the data used to train them must be managed and fed into the software delivery process.

Putting ML models into operation is just the beginning. Production data must be continuously evaluated, processed, and labeled into new training datasets, which can further improve ML models and close the feedback loop (Windheuser & Sato, 2021). That would allow models to adapt and create a continuous improvement process.

Unfortunately, industry and academia have yet to agree on the details (Dawson & Sato, 2021). Another challenge is the lack of consensus on terminology. For example, the most common understanding of “validation”, in the context of the lifecycle of an ML model, refers to the “validation of experiments in a controlled environment”, not “against real-world data, in a production environment”. Considering the diagram in Figure 1, most studies go to the “Validation of Experiments” phase (in the figure, within the Data Science block), and do not address issues of long-term operation, such as model evolution (in particular, the evaluation that occurs in the continuous monitoring stage). Unfortunately, having good statistical performance with experiments doesn’t necessarily translate to real-world performance (Harris et al., 2022). Data can change, and ML models should be able to adapt.

Figure 1: Machine Learning Lifecycle.



Source: - adapted from Dawson and Sato (2021).

Given this context, the main question is how to evaluate ML model performance continuously against real-world data in Healthcare applications. Medical data is subject to noise, bias, and missing information, which affects performance (Yu et al., 2021). At the same time, auditability concerns demand high interpretability of models’ decision processes due to strict security and correctness concerns.

The present work intends to establish an evaluation process for Machine Learning models against real-world data in Healthcare applications. Based on this high-level objective, this work intends to: (a) Investigate state-of-the-art techniques for ML evaluation and (b) describe guidelines for incorporating continuous monitoring and establish a feedback loop for model evolution.

A Case Study applied the proposed process in an oncology ICU (Intensive Care Unit) located in the city of Natal/RN on ML models used to predict the length of stay and classify the risk of death of patients. Contributions from the present work include (I) a catalog of best practices and strategies for ML model evaluation for Healthcare applications, (II) the proposal of an ML model evaluation process, and (III) a case study on applying the proposed process in a Healthcare context.

It is not part of the scope to investigate in-depth characteristics of specific methods or techniques but rather to identify their properties and features and to obtain a set of guidelines to propose an evaluation process for ML models. The following sections address this work's contributions, (2) related concepts, (3) methodology , (4) related work, (5) the proposed process, (6) a summary of the case study, (7) conclusions, and future work.

RELATED CONCEPTS

The subjects of the case study were two different models, one aimed at risk of death classification and another to predict the length of stay. Even though regression techniques would be valid for the ”length of stay” problem, both models used classification techniques. Also, the risk of death model used clustering techniques to determine the class labels, given that it was not previously defined. This section aims at providing minimal context for those concepts and the ones around model evaluation, maintenance, and change management.

UNBALANCED DATA

In practical classification applications, it is common for a class to be much more prevalent than others (Ian et al., 2005). Several ML algorithms have their performance impaired due to unbalanced data, as they tend to favor the majority class (Faceli et al., 2011).

OFFLINE LEARNING: CLASSIFICATION METHODS AND CLASSIFIER COMMITTEES (ENSEMBLES)

There are several families of ML classification techniques. The models in the Case Study touched on some of them. Model families have characteristics that will enable the choice according to the fit with the problem. Probabilistic methods expect independence between data attributes. Methods based on minimization are susceptible to running into local minima. Search-based models have great interpretability, describing the decision-making that led to the solution in the search space. Ensemble methods have the advantage of combining models from different families, favoring diversity in search space exploration.

Traditionally, model training uses classifier evaluation metrics for validating experiments. Such metrics are also natural candidates for tracking performance on real-world data, comparing against the baselines established during experimentation. Metrics commonly used include Confusion Matrix, Accuracy, Precision, Recall, F-measure (also called F1 Score), ROC curve, and ROC Area (or Area Under the Curve - AUC) (Faceli et al., 2011; Ian et al., 2005). Finally, Statistical Tests (also called Hypothesis Tests) can compare two or more models against each other. Those include the Wilcoxon tests, Friedman tests, and the Nemenyi test (Faceli et al., 2011).

ONLINE LEARNING

The techniques described above are part of the so-called offline learning techniques (also called batch learning), in the sense that they are techniques applied on known, previously obtained static datasets, consistently presenting the same result given they're provided with the same inputs (Carolan et al., 2022), presuming that data abides by some stationary distribution (Faceli et al., 2011). However, excellent offline results do not guarantee efficiency in the real world (Harris et al., 2022).

Online Learning intends to tackle challenges such as the ability to process large volumes of evolving data with non-stationary distributions. That is a common scenario in healthcare applications since Healthcare Information Systems (HIS) and Electronic Healthcare Records (EHRs) using IoMT (2.6) can produce increasing amounts of data in a constant flow and with an infinite time horizon, and distribution can change either seasonally or abruptly (as observed with pandemics such as COVID-19) de (de Fátima Marin, 2010; Malaquias, 2022).

Continuous data can change, and therefore ML models should change as well (Harris et al., 2022), to maintain a decision model that is accurate and consistent with the current state of the processes that generate data, which may change and evolve. In the presence of a non-stationary distribution, the learning system must also incorporate forgetting mechanisms in a way that allows the discarding of information and concepts that no longer reflect the current state of the problem.

CONTINUOUS LEARNING

Often the concepts of Online Learning and Continuous Learning intersect, and some consider them the same. Continuous Learning proposes to keep model fitness by constant reassessment and evolution (Parisi et al., 2019), continuously learning from a continuous data flow (Ettun, 2019). That could mean learning autonomously and adapting as the model acquires new production data.

Those principles seem ideal for healthcare scenarios but present challenges that are not yet fully resolved. One such issue is the catastrophic forgetting problem, where new data can negatively influence previous learning. Another one is convergence, where the model starts to predict itself and should not be updated (Harris et al., 2022). That can lead to an abrupt drop in performance when integrating new data or overwriting previous knowledge (Lee & Lee, 2020). That is called the plasticity-elasticity dilemma: finding a balance that is plastic for existing knowledge but elastic for incorporating what's new (Parisi et al., 2019).

DRIFT

Drift refers to changes that can affect input data, results, or the fitness of the model and is a concept closely associated with Online Learning (Bifet et al., 2023) but also observed in performance monitoring of models (of any model family) in production (Mohandas, 2021). Drift can occur quickly (abruptly), gradually, incrementally, or recurrently (Toor et al., 2020).

In healthcare applications, drifts tend to occur due to changes in the characteristics presented by patients. It can also occur due to changes in national health policies, for example, in the case of a drop in the number of smokers. Drifts can also happen due to emerging diseases. The COVID-19 pandemic brought severe, abrupt, and continuous drifts in data, leading to unprecedented scenarios in healthcare and causing the need to restructure operations and services (Duckworth et al., 2021).

HEALTH INFORMATION SYSTEMS, INTERNET OF MEDICAL THINGS (IOMT) AND THE HL7 STANDARD

In the healthcare sector, many suppliers manufacture a wide range of products. Most of these products claim to follow standard rules and protocols in the design process. However, validation formalism is missing to confirm this information (Pradhan et al., 2021). More often than not, Healthcare applications involve different data sources used for various purposes by various users. That leads to data heterogeneity, a predominant feature of clinical data (Malaquias, 2022). Due to the divergence of existing formats for communicating health data, the Health Level Seven (HL7) international standard proposes to provide interoperability standards for health data.

Recently, this field has seen exponential growth in products and research at different levels, like sensors, networks, and repositories. Medical data standards such as HL7 become fundamental tools for integrating data from diverse sources, with multiple manufacturers' standards and data models in place. Healthcare applications, previously centered on hospitals, evolved into patient-centered systems (Pradhan et al., 2021). There are increasing requirements for the volume of sensor-generated data, which influences the complexity of their analysis and the performance level of the provided solutions (El-Baz & Suri, 2023). These concerns are realities in healthcare, where HIS and EHRs can generate data in continuous streams with an infinite time horizon for each patient (Harris et al., 2022).

METHODOLOGY

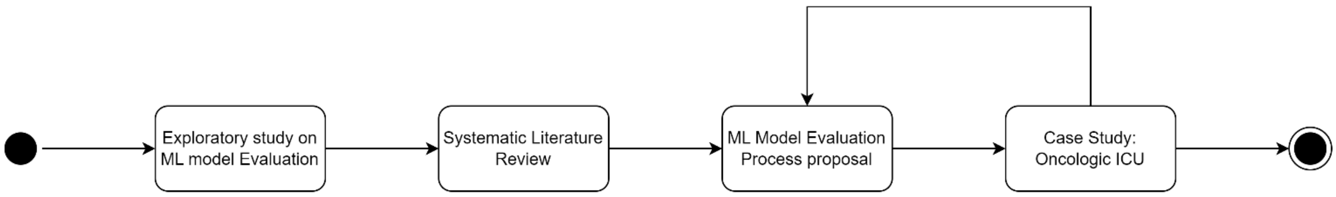
This work was carried out in Instituto Metrópole Digital (IMD), Universidade Federal do Rio Grande do Norte's (UFRN) Specialized Academic Unit, which has the mission of promoting the creation of a Technological Pole in Information Technology in the state of Rio Grande do Norte, Brazil, covering initiatives from the public, private and academic sectors. In this sense, IMD has, in addition to educational and academic initiatives, a series of actions aimed at the development of technological solutions, through projects developed through partnerships with both public and private initiatives.

The general problem emerged from the technological areas of healthcare, in healthcare applications developed by IMD itself. The work methodology was based on the Action research philosophy, seeking transformative change through the simultaneous process of taking action and doing research, which are linked together by critical reflection.

Initially, an exploratory study was conducted on Performance Assessment, Monitoring and Change Management for Machine Learning Models in a more high-level way, but including ML applications for healthcare. Then, a Systematic Literature Review (SLR) was carried out, with the objective of understanding the state of the art for Performance Evaluation, Monitoring and Change Management for Machine Learning Models for healthcare applications specifically.

Based on the SLR, challenges and opportunities were identified and, thus, the main requirements for evaluation, monitoring and change management for ML models in healthcare applications were specified. Once the solution requirements were defined, it was time to methodologically define how the solution should be built to satisfy the identified requirements. Then, a process for continuous evaluation, continuous monitoring and change management for ML models was developed, aimed at healthcare applications. That process was then applied to a Case Study in an Oncology ICU, to access the process usefulness, and obtain feedback to improve it. These activities are described in Figure 2.

Figure 2: Work methodology.


Exploratory study on ML model Evaluation
Systematic Literature Review
ML Model Evaluation Process proposal
Case Study: Oncologic ICU

Source: Authors (2023).

RELATED WORK

This section describes related work found in the literature around continuous monitoring and change management for Machine Learning, as part of the MLOps culture, considerations for evaluation of Machine Learning models, and specific concerns around the application of Machine Learning to Healthcare contexts.

CONTINUOUS MONITORING AND CHANGE MANAGEMENT IN MACHINE LEARNING

One of the key differences between traditional software engineering and ML development is that the real work doesn't start until after deployment into production. There are concerns with monitoring ML models in production, going through topics such as system health, performance, delayed outcomes, importance weighting, and identifying and measuring drift in online and offline learning (Mohandas, 2021).

Continuous learning (2.4) would be a possible approach to deal with model change (data distribution changes, outcome changes, model mappings, etc.). The model would undergo retraining using production data continuously (Ettun, 2019). The implementation of this principle, however, remains challenging due to the ”catastrophic forgetting problem” where new data can interfere with and even overwrite previous learning (Harris et al., 2022).

Parisi et al. (2019) discusses how continuous learning (2.4) in neural networks (2.2) can benefit from biological studies to outline guidelines for addressing the plasticity-elasticity dilemma. It also discusses the use of transfer learning for transferring previous knowledge between new stages of learning and reinforcement learning for autonomous exploration and development of new knowledge in self-organizing neural networks, Long Short-Term Memory (LSTM) networks, among other ANN techniques.

Ettun (2019) advocates using AutoML (Automated Machine Learning), a set of tools and services that abstract the details and knowledge needed to perform Machine Learning, automating tasks necessary for creating models, being related to no-code and low-code development trends. The idea would be to use different parameters and algorithms and would allow for the use of pre-built popular templates by simply specifying their parameters.

Bifet et al. (2023) addresses change management in online learning, citing concerns such as detecting changes in the data stream and adapting models as soon as possible while remaining robust against noise and outliers. Miranda et al. (2022) also echoes those concerns. They propose three strategies to deal with changes: the first focuses on monitoring predetermined statistics. In the second, revision and change-detection algorithms run parallel to building new models, asserting whether the drift is abrupt, gradual, local, or global. The third proposes using ensembles that generate models under different conditions and times and an ensemble manager that contains the rules for creating, deleting, combining, and revising models, which is primarily responsible for reacting to changes (Bifet et al., 2023).

On another spectrum, regulatory considerations can also impact change management for ML models, especially in Healthcare applications where criteria and prerequisites outside the technical scope can generate impediments to changes in a model. Carolan et al. (2022) discusses guidelines for regulating BF in healthcare.

The Foods & Drugs Administration (FDA), the American agency responsible for regulating health-related products, has published an action plan with intended changes for health systems (2.6) management regarding the use of ML (Food and Drug Administration [FDA], 2021), which have not yet been detailed or entered into effect as of the time of writing.

EVALUATION OF MACHINE LEARNING MODELS

In its classical interpretation, model validation is the process that occurs during model experiments where model performance gets measured against a set of previously curated data samples (Ian et al., 2005). However, that differs from what this work considers an evaluation. Evaluation occurs after the model gets deployed into production based on measurements obtained against data from the real world. There are challenges in this scenario, like the need for the periodic redeployment of models, due to performance decay over time; monitoring and alerts to identify biases, anomalies, and drift (2.5); regulatory concerns (compliance), among others.

In terms of evaluation criteria, Panesar (2019) makes an interesting differentiation between offline evaluation and online evaluation, describing the latter as continuously validating models using data from the real world, with a live assessment process, which is very close to change management processes in online learning. They propose the application of multi-variable tests and statistical tests (2.2).

Biswas (2021) proposes a framework for evaluating ML models based on a financial market methodology, the Model Risk Management (MRM), which is standard practice for any financial institution to assess the risk of models. It discusses two main aspects: risk level analysis, to quantify the damage of a solution, and bias and fairness, which is related to empiric asymmetries in the data and overfitting (Reagan, 2021). Validators should evaluate variables carefully, and to avoid bias, sensitive variables (such as age, gender, religion, and profession) should be excluded from data modeling unless they are critical to the problem (Biswas, 2021). Validators should also consider the level of risk in these analyses, especially in high-risk models, and define standardized techniques for attacking biases.

Luo et al. (2016) discusses guidelines for reporting results in ML models for healthcare with 12 reporting criteria that researchers should observe while preparing manuscripts. Stevens et al. (2020) and Vollmer et al. (2020) proposals are similar: the first discusses criteria divided into categories such as study design, data sources and processing, development, and validation of the model, while the second describes a framework composed of 20 criteria in 6 categories: inception, study, statistical methods, reproducibility, impact evaluation, and implementation. Any of those works describe specific metrics, though.

The suitability of an ML model to the problem it tries to tackle is subject to aspects other than statistical performance. Failures are part of the process, but ML projects could be doomed to fail from the outset due to misalignment between product metrics and model metrics (Kornhouser, 2021). Model evaluation should include business performance concerns, hence the need to define and quantify the success of a model considering the differences between business performance and statistical performance. Panesar (2019) describes the need for online evaluation of models by metrics and business objectives not directly related to the data, seeking metrics that guarantee representativeness for all project stakeholders.

Flach (2019) describes a measurement theory for ML based on the representational measurement theory (Krantz et al., 1971), which uses mathematical techniques: concatenation relation, scale, and transformations, aiming to create a broader coverage evaluation metric.

Walsh et al. (2020) describes requirements and recommendations for validating ML models, such as having baselines for comparison with publicly available methods and metrics validated by the community. It suggests comparing related methods and alternatives on the same datasets, ablation studies to measure the impact of components, continuously checking data distribution to ensure good representativeness, and defining confidence and error intervals to measure robustness.

Duckworth et al. (2021) discusses a model based on decision trees with boost by gradient and presents an algorithm to generate drift (2.5) monitoring metrics in a COVID-19 ICU context. Other tree-based methods, such as Random Forest and Decision Trees, can also use this method (2.2).

Similarly, Chiang and Dey (2019) describes a use case in healthcare using a model with multiple independent Random Forests with Feature Selection (RFFS). It also proposes Online Weighted Resampling (OWR) as a technique to alleviate the effects of drift in models based on Random Forest (RF) (2.2) and proposes a resampling mechanism that assigns weights to the samples to influence the bias towards those that occur more frequently.

MACHINE LEARNING FOR HEALTHCARE

The potential impact of ML in healthcare applications is genuinely exciting. Its limited adoption in clinical settings indicates that current strategies are far from ideal, though, and the promise that ML can increase the ability of humans to provide healthcare has yet to materialize (Wiens et al., 2019).

He et al. (2019) reviews the main practical problems ML faces in existing clinical workflows: data sharing; transparency (interpretability and biases); data correctness; patient safety and accountability (regulation, quality control); lack of data standardization and integration into existing clinical flows (interoperability); financial issues (monitoring and maintenance of models, hardware evolutions); educate the health workforce about ML (recognize benefits, limitations and better interact with data scientists).

Making clinical data usable for research is a formidable challenge, as data is often multimodal, presented in different patterns, and accumulated without context (Deasy & Stetson, 2021). Morin et al. (2021) reports on a data integration framework that enables continuous learning from multimodal health data while capturing and integrating longitudinal clinical data in a functional data mart with analytic workflows.

With the popularization of IoMT (2.6), volumes of medical data are getting larger and larger, which brings particular challenges regarding the storage and processing of this data. Toor et al. (2020) describes an approach to treat concept drift (2.5), starting from the Reactive Drift Detection Method (RDDM) method and proposing the Enhanced (E)RDDM.

Wiens et al. (2019) proposes a roadmap for the responsible implementation of ML in health, with recommendations for choosing the correct problems, developing meaningful solutions, addressing ethical implications, performing a rigorous evaluation of models, reporting results carefully, responsible implementation, and arrival on the market.

DISCUSSION

There are few works available in ML literature that focus on evaluating ML models against real-world data. Those call out the challenges but do not present structured solutions. There are attempts to propose frameworks and guidelines (Biswas, 2021; Luo et al., 2016; Stevens et al., 2020; Vollmer et al., 2020). However, they lack clear metrics or concrete steps to guide the evaluation process, keeping concerns at a high level of abstraction. They only provide general guidelines to keep in mind during the model lifecycle.

Some proposals (Ettun, 2019; Morin et al., 2021; Parisi et al., 2019) put concerns about the explainability and auditing of models under the spotlight. Continuous learning approaches seem very focused on neural networks, which by definition are black-box learning processes, and thus use unexplainable learning processes. With AutoML, on the other hand, as the process is automated, it can lead to unexpected results. Both monitoring and implementation of AutoML require additional care, with manual steps being mandatory to guarantee the performance of the new versions of the models before being used on real-world data.

On the other hand, Drift Detection Methods (Baena-Garcıa et al., 2006; Bifet et al., 2023; Duckworth et al., 2021; Nishida & Yamauchi, 2007; Settipalli & Gangadharan, 2021; Toor et al.; 2020;) provide metrics for monitoring the number of errors produced by a model against real-world data, which should decrease or, at least, remain constant if the model performance is stable. If the prediction errors increase, it would be indicative of drift (2.5).

Evaluation should also go beyond statistical methods for measuring model performance (Carolan et al., 2022; FDA, 2021; Kornhouser, 2021; Walsh et al., 2020) business performance and regulatory concerns should be considered throughout the model lifecycle and accounted for in the model evaluation. Figure 3 describes the proposed process for evaluating models in real-world scenarios.

A PROCESS FOR EVALUATING MACHINE LEARNING MODELS IN EFFECTIVE USE

This section describes the proposed process for evaluating Machine Learning models using a general adaptative approach that, while aimed at healthcare applications, could be extended to other application areas. The process is divided into phases and provides guidelines for producing metrics for model evaluation and monitoring, tailored to the type of model and application being observed.

PHASE 1: DOCUMENTATION

Before performing the evaluation, it's beneficial to gather detailed documentation of the experimental designs to understand the model, describing the target problem and intended predictions (Wojtusiak, 2021). To that end, reports of experimental results, descriptions of data sources, context, participants, outcomes, and predictors are desirable (Wiens et al., 2019).

Obtain Available Documentation

This activity seeks to acquire all available documentation on the models to contextualize the solution space and justify their reasoning. Identify whether it is an Offline (Batch) or Online Learning (Data Stream Learning) problem, which will already indicate the expected volume of new data.

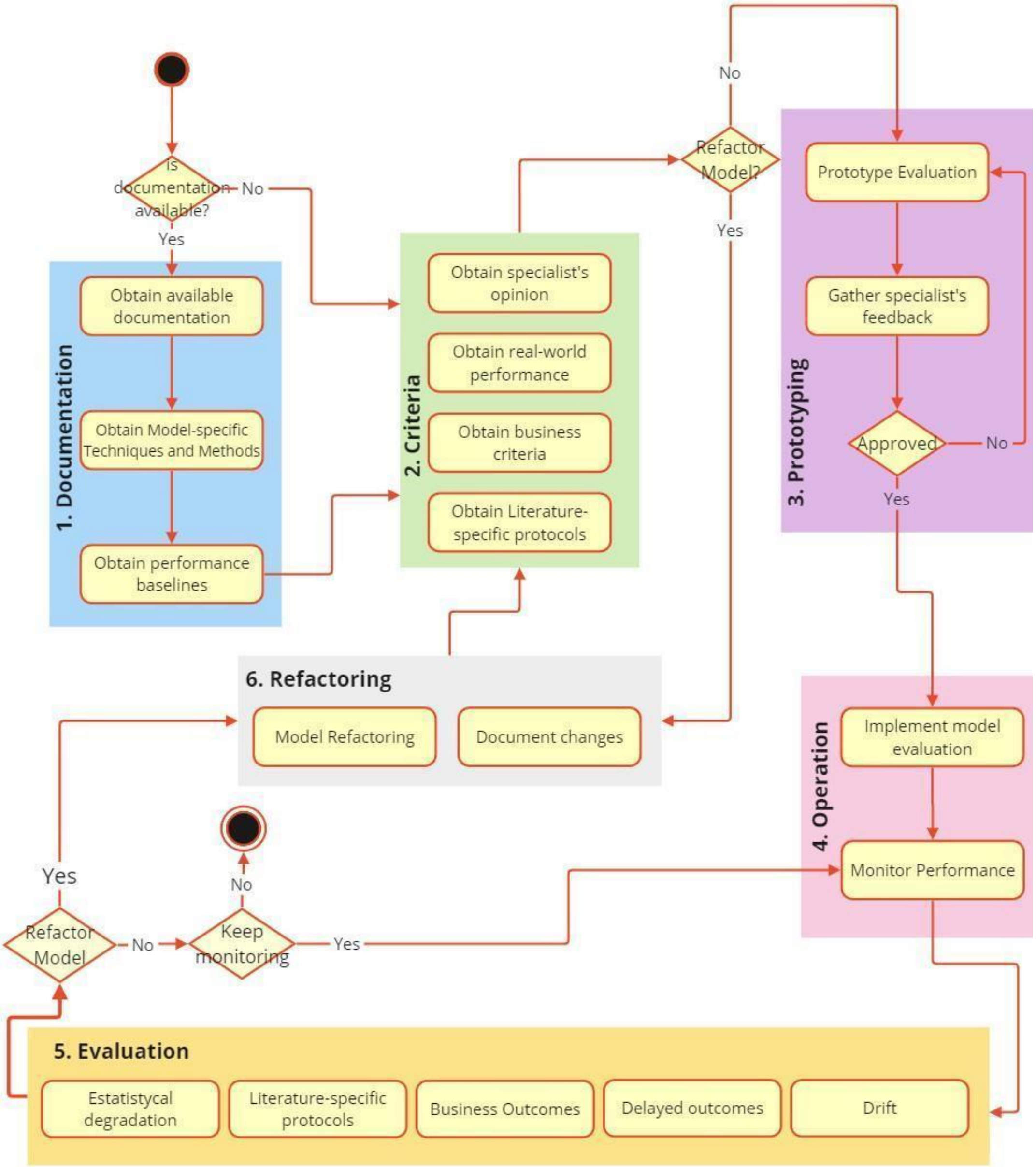
Obtain Model-specific Techniques and Methods

Based on the type of problem (classification, regression, clustering, etc.) and the specific characteristics of the family of models in use (tree-based methods, gradient-based methods, ANN-based methods, etc., identify performance evaluation methods and techniques for reference in future stages.

Obtain Model's Baseline Performance (Training Data)

Statistical analysis (accuracy, F-measure, Statistical Test, etc.) (2.2) is the gold standard for evaluating models during experimentation (Wojtusiak, 2021). Retrieve the model's baseline performance. That could be used to identify, during performance monitoring, whether there has been degradation compared to real-world data.

Figure 3: Proposal - Evaluation for ML Models in (real-world) effective use.



Source: Authors (2023).

PHASE 2: CRITERIA

Once available documentation has been gathered and context has been established for the model and it’s intended application, phase 2 aims at iteratively developing and defining the evaluation criteria to be considered and guide the evaluation process of the models. Specialists should be consulted, real-world performance should be gathered, and business-related criteria that can help assess model validity should be determined, as well as literature specific protocols whenever possible and applicable.

Obtain Specialist's Opinion

In addition to quantitative performance measures, qualitative approaches can expose problems associated with biases that quantitative measures may have missed (Wiens et al., 2019). Obtain opinions from specialists on model business performance and insights that might help prototyping model evaluation. Factors should include product and business fitness, desired key results, and critical success factors. "Specialist" refers to people with technical knowledge of the application area and anyone knowledgeable on business objectives the model can influence.

Obtain Real World Performance

Assess model performance on real-world data, and compare it against baseline performance data collected from the training and validation phases, to identify gaps between expected and real-world performance.

Obtain Business Criteria

Based on specialists' feedback, determine business metrics (which are not necessarily related to the statistical performance) that are influenced by (or influence) the model's results.

Obtain literature-specific metrics and protocols

Obtain area-of-application-specific protocols and metrics to establish performance benchmarks. For healthcare applications, performance should consider clinically relevant evaluation metrics (Wiens et al., 2019). Should use publicly available methods and metrics, which are endorsed by the healthcare community (Walsh et al., 2020).

PHASE 3: PROTOTYPING

Responsible for prototyping model evaluation, to be validated with specialists. Before starting the prototyping phase, verify that the current model performance is within an acceptable threshold. Otherwise, the process moves to phase 6, Refactoring (5.6).

Prototype Model Evaluation

Qualitative and quantitative performance metrics should represent all the stakeholders of the project (Panesar, 2019). Those should include statistical, business, and area-of-application-specific metrics. Visualizations can facilitate a specialist's assessment. The proposal can also include manual steps to collect stakeholder feedback during monitoring.

Present Prototypes / Collect Expert Opinion

Specialists should validate the artifacts produced in the previous activity. Present and discuss those to reach a consensus if it meets the application's needs for performance monitoring. Otherwise, the process iterates to create a new proposal and submit it to specialists' assessment until it's approved.

PHASE 4: OPERATION

Once prototyping has been completed and approved, it's time to get the model evaluation implemented and operational so that metrics can be derived and monitoring can begin, generating data that will be later used to determine if the model is performing as intended or if refactoring will be necessary.

Implement Model evaluation

This activity involves the coding and configuration to deliver the tailored model's evaluation. Monitoring can use both in-house developed solutions and existing platforms. Either way, the chosen solution should monitor the previously defined evaluation criteria. That information should also be available to stakeholders. Might involve integrating with other data sources for cross-validation of Business criteria, Literature Specific Criteria, etc.

Monitor Model Performance

It is crucial to monitor model performance continuously after deployment (Panesar, 2019). Variance, drift, and other changes require confronting current results against baselines frequently. Perform monitoring of statistical performance, area-of-application-specific protocols and metrics, and business metrics. That process should occur continuously, include feedback mechanisms for specialists, and be automated as much as possible (Mohandas, 2021; Sato et al., 2019). Make use of visual monitoring through management dashboards and automated monitoring.

PHASE 5: EVALUATION

Once model evaluation is operational, and monitoring is in place, the evaluation can be performed in a continuous fashion, in several ways, such as statistical and quantitative performance evaluation, comparison against literature specific metrics, evaluation of business impact through correlation with business metrics, among others.

Evaluate Statistical / Quantitative Performance Degradation (Model Decay)

As proposed in Wojtusiak (2021), test whether the models performance is as expected. Confront current performance against baselines obtained from the experimentation phase, and, eventually, against real-world historical data (sliding window analysis).

Evaluate Literature-specific Protocols

Evaluation of literature-specific protocols intends to acquire scientific validity by using community-recognized metrics from the application area (Walsh et al., 2020). Especially for healthcare, where there is resistance and distrust in adopting ML models, using clinically relevant metrics for performance analysis adds validity to the evaluation process (Wiens et al., 2019).

Evaluate Business Results (Product/Business fitness)

Misalignment between model and business metrics can cause undesirable effects on performance. Business performance is often lost in the over-promotion of model performance (Kornhouser, 2021). Having statistically accurate performance metrics while failing to meet business expectations is a recipe for failure. Evaluation Should include business specialists' qualitative feedback over real-world data or against available business metrics. Also, consider the cost and the impact of implementation (Wiens et al., 2019).

Evaluate Delayed Outcomes (Delayed Outcomes)

Sometimes the actual results are not immediately available to measure model performance on real-world inputs. That is especially true if a significant delay happens in acquiring results or if real-world data needs to be annotated by specialists (Mohandas, 2021). Analyze whether there are delayed outcomes and whether any adjustments to the model response time are necessary.

Evaluate Drift

Drift can lead to model decay over time, affecting input data, model results, or even the actual concept mapped by the model. Identify drift, and propose metrics and methods for monitoring it. There are several drift detection methods and metrics available in the literature (Baena-Garcıa et al., 2006; Bifet et al., 2023; Chiang & Dey, 2019; Duckworth et al., 2021; Nishida & Yamauchi, 2007; Toor et al., 2020).

PHASE 6: REFACTORING

If a model’s performance is found to be lacking in some or all of the monitored metrics, a decision can be made to refactor and evolve the model and reach better performance results. This phase discusses concerns related to how refactoring should be reflected into the evaluation process.

Refactor / Evolve Model

Eventually, during previous phases, a decision can be reached (automatically or manually) that the model’s performance is insufficient. If new evaluation criteria or changes in currently mapped criteria occur, the process may involve getting specialists’ feedback again.

Document Change

Once refactoring is complete, the new model(s) might need updated documentation. It might also be necessary to determine whether to modify the evaluation process if criteria have changed, which leads the process either to cycle back to the Criteria Phase (5.2) or to resume monitoring (Operation Phase – 5.4).

CASE STUDY: EVALUATING ML MODELS IN AN ONCOLOGY ICU

The case study is a suitable research methodology for research in software engineering, as it studies contemporary phenomena in their natural context. However, understanding of what constitutes a case study varies (Runeson & Höst, 2008). As with the systematic review, it is necessary to adopt a protocol and guidelines that favor conducting the case study in a reproducible way and that facilitates its interpretation for the reader. Therefore, the guidelines proposed in Kitchenham et al. (1995), Runeson and Höst (2008) and Yin (2009).

The planning of the case study includes the description of the research questions, the subjects who participated in it, the objective used, the units of analysis, the evaluated artifacts and the evaluation criteria, as well as the procedures used for data collection (Barroca Filho, 2015).

RESEARCH QUESTIONS

Research questions need to maintain a chain of evidence consistent with the study's conclusions (Yin, 2009). However, although research questions state what needs to be known in order to achieve the study's objectives, these may evolve over the course of the study, over its iterations (Runeson & Höst, 2008).

- **RQ1**: Is the proposed process useful for evaluating the performance of ML models in Healthcare applications?

- **RQ2**: Which are the benefits, problems and challenges that using the proposed evaluation process?

To answer these questions, in addition to observing and monitoring the evolution of the ML models in the PAR platform, two sets of questionnaires were applied: the first one aimed at the team of data scientists and developers of PAR (available at https://forms.gle/eVRg8x3FRbmp1MpYA), and a second one for healthcare professionals involved in the day-to-day use, evaluation and evolution of the models on the oncology ICU (available at https://forms.gle/EjXsmP37nv5FUvubA).

SUBJECTS OF STUDY AND DATA COLLECTION

The proposed process was used by the team of developers and data scientists responsible for the PAR platform (henceforth called the IT team), which included: a director of IT Innovation, a Data Scientist and two Data Science fellows (both specialized in ML), a Software Development Coordinator and a Software Development Fellow (both specialized in Java and Spring Boot).

The process was also used by the team of healthcare professionals involved in the implementation, adequacy and evolution of the PAR platform in the oncology ICU (henceforth called the Healthcare team), including: A Healthcare Innovation Manager, an ICU managing physician, a statistician, 3 fellows working in the ICU.

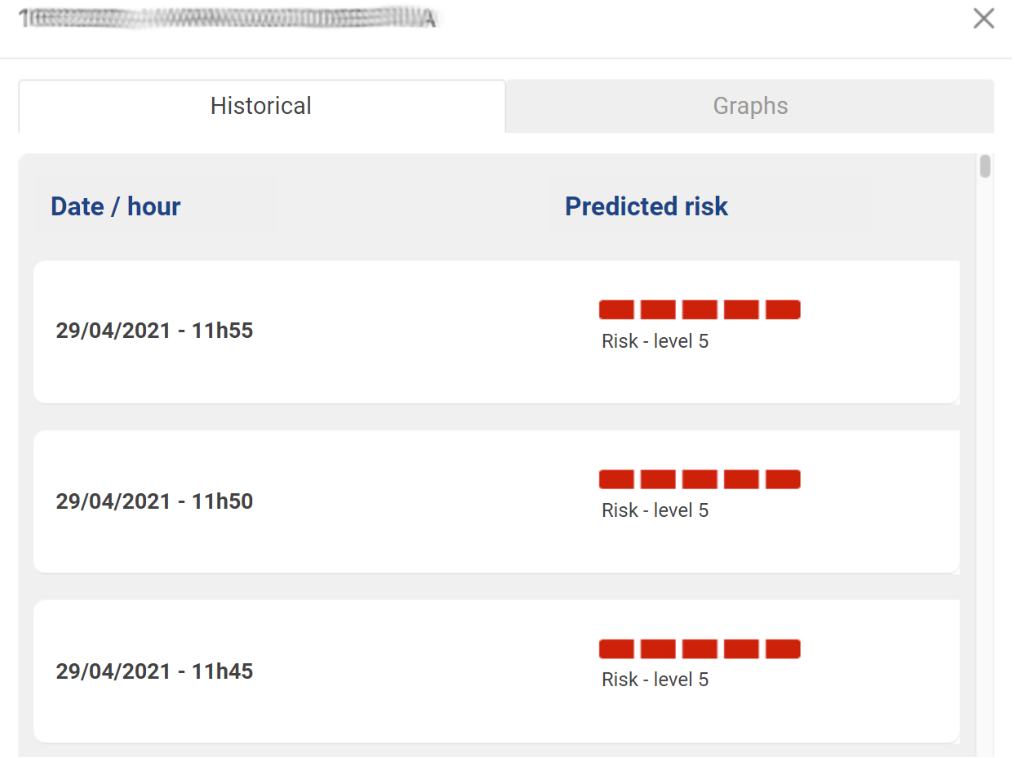
In this study, the following data was collected:

- Responses to the questionnaires applied to the IT team,
- Responses to the questionnaires applied to the Healthcare team,
- Notes on follow-up meetings with both the IT team and the Healthcare team while running the process.

OBJECT OF STUDY

The Case Study was conducted on features available on the Remote Assistance Platform (PAR). PAR is a computational solution based on the Internet of Things (IoT) infrastructure, which aims to promote, through the use of web and mobile applications, intelligent, personalized, remote (distance) and real-time monitoring of vital signs and the environment of patients with chronic conditions, in critical situations and/or in need of constant monitoring, whether they are hospitalized in Intensive Care Units (ICUs), wards, urgency/emergency or homecare beds. The PAR Platform has ML models available to classify Risk of Death (RoD) for patients and predict Length of Stay in ICU beds (LoS). Those ML models, shown in Figures 4 and 5, are the Objects of Study.

Figure 4: PAR platform – Risk of Death classification.



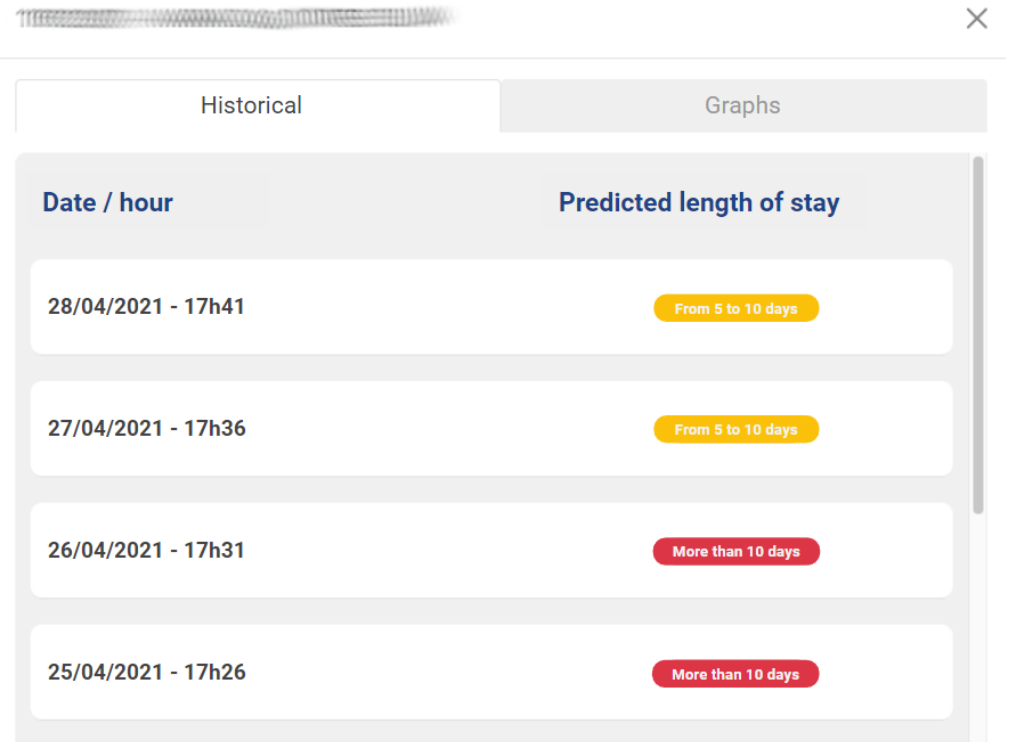
[redacted] ×

Historical | Graphs

Date / hour	Predicted risk
29/04/2021 - 11h55	🟥🟥🟥🟥🟥 Risk - level 5
29/04/2021 - 11h50	🟥🟥🟥🟥🟥 Risk - level 5
29/04/2021 - 11h45	🟥🟥🟥🟥🟥 Risk - level 5

Source: Authors (2023).

Figure 5: PAR platform – Length of Stay prediction.



[redacted] ×

Historical | Graphs

Date / hour	Predicted length of stay
28/04/2021 - 17h41	From 5 to 10 days
27/04/2021 - 17h36	From 5 to 10 days
26/04/2021 - 17h31	More than 10 days
25/04/2021 - 17h26	More than 10 days

Source: Authors (2023).

EXECUTION

During execution of Phase 1, Documentation, all information regarding the ML models was gathered. After experimentation with several ML models, the following decisions were taken: both problems were addressed as classification problems (even though the LoS problem is inherently a regression problem, ranges of time were used to adapt it as a classification problem). Since RoD classes were not labelled, a first round of clustering was performed to create candidate risk classification labels. Those were then decided upon using feedback from the Healthcare team. Given demands of high explainability, and performance outcomes during experimentation, Random Forest (RF) ensembles were chosen as the models for both LoS and RoD. Performance baselines were recorded based on the available documentation.

During the execution of Phase 2, Criteria, the Healthcare team focused heavily on the RoD model, leaving the LoS model outside of most discussions. It was also decided, due to the Healthcare team's feedback, that the RoD model was not performing as expected, with a heavy bias detected towards the highest risk classification denoting class imbalance. Business criteria was lightly touched upon, and a heavy focus was given to comparison between the RoD prediction and ICU literature-specific risk scores, such as MPM0, SAPS, SOFA, NEWS and APACHE. It was decided that comparison with those scores should be used as one of the criteria for the model performance Evaluation.

Execution proceeded with a detour to the Refactoring Phase (Phase 6) for the RoD model, and discussions around the LoS model halted completely. Meanwhile, execution of Phase 3, Prototyping, started with discussions around integration with Electronic Healthcare Records Systems and Laboratory Exams Systems to retrieve data needed for calculating the literature-specific risk scores. Due to restrictions around data privacy and the unstructured nature of most data, it was decided that for a MVP all missing information would be fed into PAR by the Healthcare team manually. It was decided that for that same MVP, the MPM0 score was to be used, given it was the one that demanded less manual input. Evaluation was to be performed against the MPM0-III score and historical performance data over previously produced risk classifications for patients, that were also going to be manually validated by the Healthcare team. Prototypes were produced based on those agreements. Figure 6 describes the proposed process activities as visited during the execution.

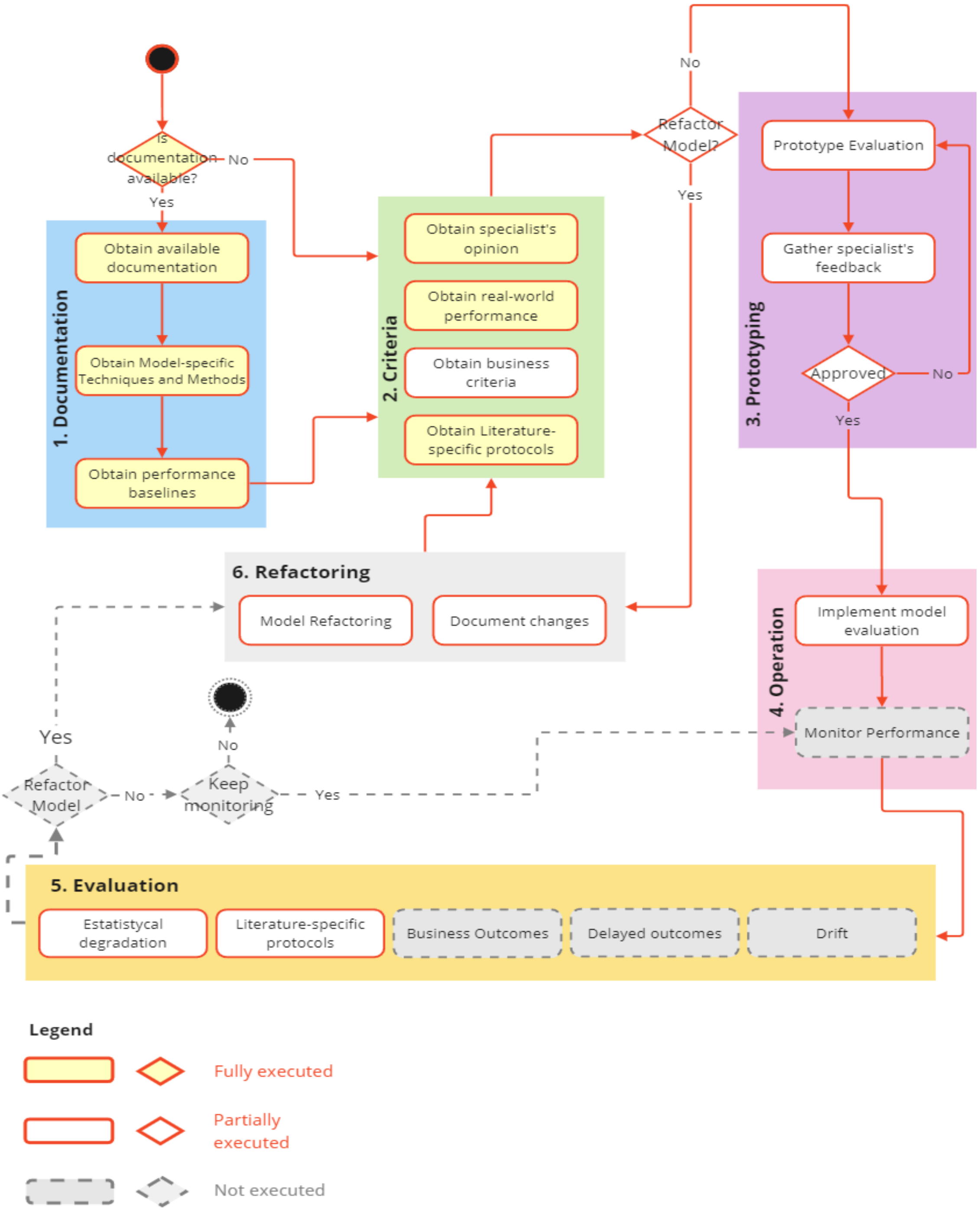
Observations made during execution:

- Low availability of sample data for model training;
- Low representativeness of samples, indicative of class imbalance;
- Few instances and attributes were used (250 instances, 30 attributes);
- Short experimentation time;
- Datasets containing sparse instances, inherent to the nature of the data and scale variance between attributes (which could be mitigated using normalization techniques);
- Difficulties in gathering feedback from the Healthcare team, given the critical work environment and low availability of professionals, which had to coordinate evaluation

activities with the day-to-day ICU routine with critical care patients (mission-critical environment);

- Decision to use offline learning despite both models having the characteristics of typical online learning scenarios;
- APACHE 4 would be the best match in terms of literature-specific scores, since it's capable of predicting risk of death and also Length of Stay;
- Operation and monitoring were hampered by the focus on refactoring the risk model;

Figure 6: Process according to Case Study execution.



Source: Authors (2023).

THREATS TO VALIDITY

Due to low availability from the Healthcare team, refactoring of the RoD model and work on the LoS model were affected, which impacted on activities like Evaluation Prototyping (5.3.1), Evaluation Implementation (5.4.1), Monitoring (5.4.2), the Evaluation Phase itself (5.5) and Model Refactoring (5.6.1). The evolution of the process itself benefited from this unfolding of events, though, and this exception route was mapped back into it. Some activities like Obtaining (5.2.3) and Evaluation of Business Criteria (5.5.3) were only ever discussed superficially, and challenges like obtaining other data sources to further enhance models and the evaluation process presented themselves, forcing execution to steer towards manual inputs by specialists.

Given this context, four types of validity were assessed:

- **Construction Validity:** The capture of data from the execution of this case study was carried out through the application of impersonal questionnaires (without subject identification), to the development teams and domain specialists. Data interpretation was conducted by an unbiased researcher to avoid interpretation errors.
- **Internal Validity:** The case study was carried out with developers experienced in developing Healthcare applications and Machine Learning models. At the start of the study, the proposed evaluation process was discussed in detail, feedback was collected, and adjustments were made, increasing confidence in the process.
- **External Validity:** This case study was carried out within a platform which solves real Healthcare problems, with advanced monitoring solutions implemented on the PAR platform and in active use by a reference hospital, making it possible to evaluate ML model performance against real-world Healthcare data, which allows for generalization of this case study's results and research responses to the healthcare practice.
- **Validity of Conclusion:** In this case study there was no control group, as the objective was not to compare the proposed process with other ML model evaluation processes (which, if exists, were not identified by the Systematic Literature Review). Therefore, there was no way to establish statistical relationships. Quantitative and qualitative data that contributed to the evaluation of the proposed process were used, considering its context, that is, the evaluation of ML models in effective use for Healthcare applications.

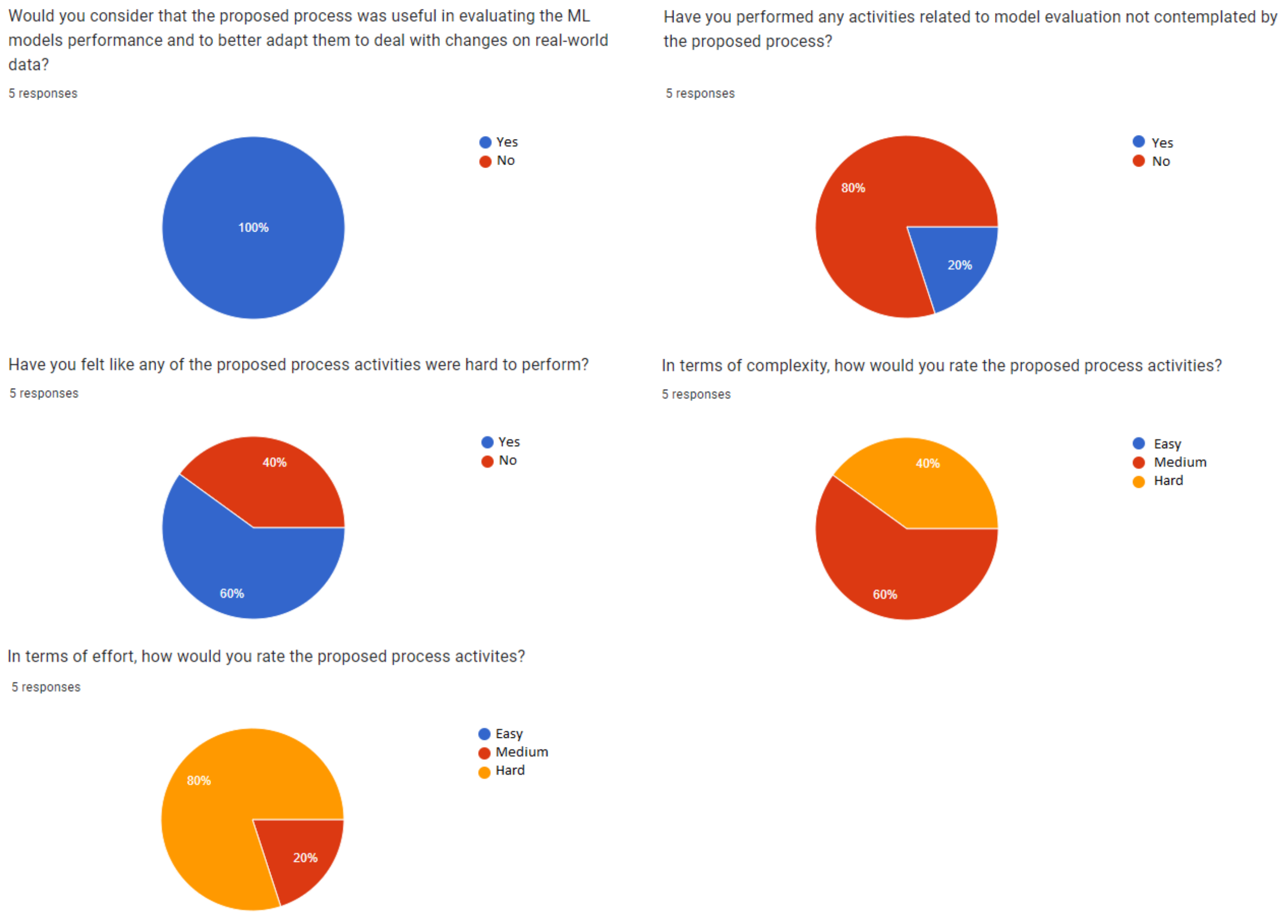
RESEARCH QUESTIONS - RESPONSES

As previously stated, two sets of subjects were taken into account, and for each, a set o questionaries was applied. Questionary responses can be found on the following URLs https://bit.ly/45NFJE5 and https://bit.ly/43Osdht. Figure 7 shows aggregated responses for the IT team questionaries, and Figure 8 shows the same for the Healthcare team.

RQ1: Is the proposed process useful for evaluating the performance of ML models in Healthcare applications?

The IT Team unanimously recognized that the proposed process was useful for evaluating the performance of the models. When questioned about the activities performed outside of those proposed by the process, one of the developers expressed that they felt the need to search for alternatives, which, if found, would be discussed with the larger group. The Healthcare team, on the other hand, was unanimous in declaring that they find the ML features useful in the PAR platform, and that none of them is unnecessary. They also unanimously recognized that the process was useful in evaluating performance of the models and in increasing trust in the estimates produced by those models, which was one welcomed additional benefit of the process. All in all, the case study successfully answered RQ1, and responses were widely positive, meaning the proposed process was considered useful for ML model evaluation.

Figure 7: IT team questionnaire responses.



Would you consider that the proposed process was useful in evaluating the ML models performance and to better adapt them to deal with changes on real-world data?

5 responses

Have you performed any activities related to model evaluation not contemplated by the proposed process?

5 responses

Have you felt like any of the proposed process activities were hard to perform?

5 responses

In terms of complexity, how would you rate the proposed process activities?

5 responses

In terms of effort, how would you rate the proposed process activites?

5 responses

Source: Authors (2023).

RQ2: Which are the benefits, problems and challenges that using the proposed evaluation process?

When asked about it, the IT team declared recurring challenges in interacting and articulating with the Healthcare team, specially in getting their feedback on several activities and moments throughout both the ML models creation process and the evaluation process. The Healthcare team on the other hand didn't explicitly declare any issues with the process. In terms of complexity of the process's activities, the IT team considered them to be of moderate-to-high complexity, while the Healthcare team considered them of moderate complexity. In terms of effort, most of the IT team considered effort high, while the Healthcare team considered them of moderate effort. When asked about the time effort involved, responses varied, and no consensus was reached. Some responses reported a couple dozen hours, others up to 6 months, and others declared they wouldn't know how to estimate it, both on IT and Healthcare teams.

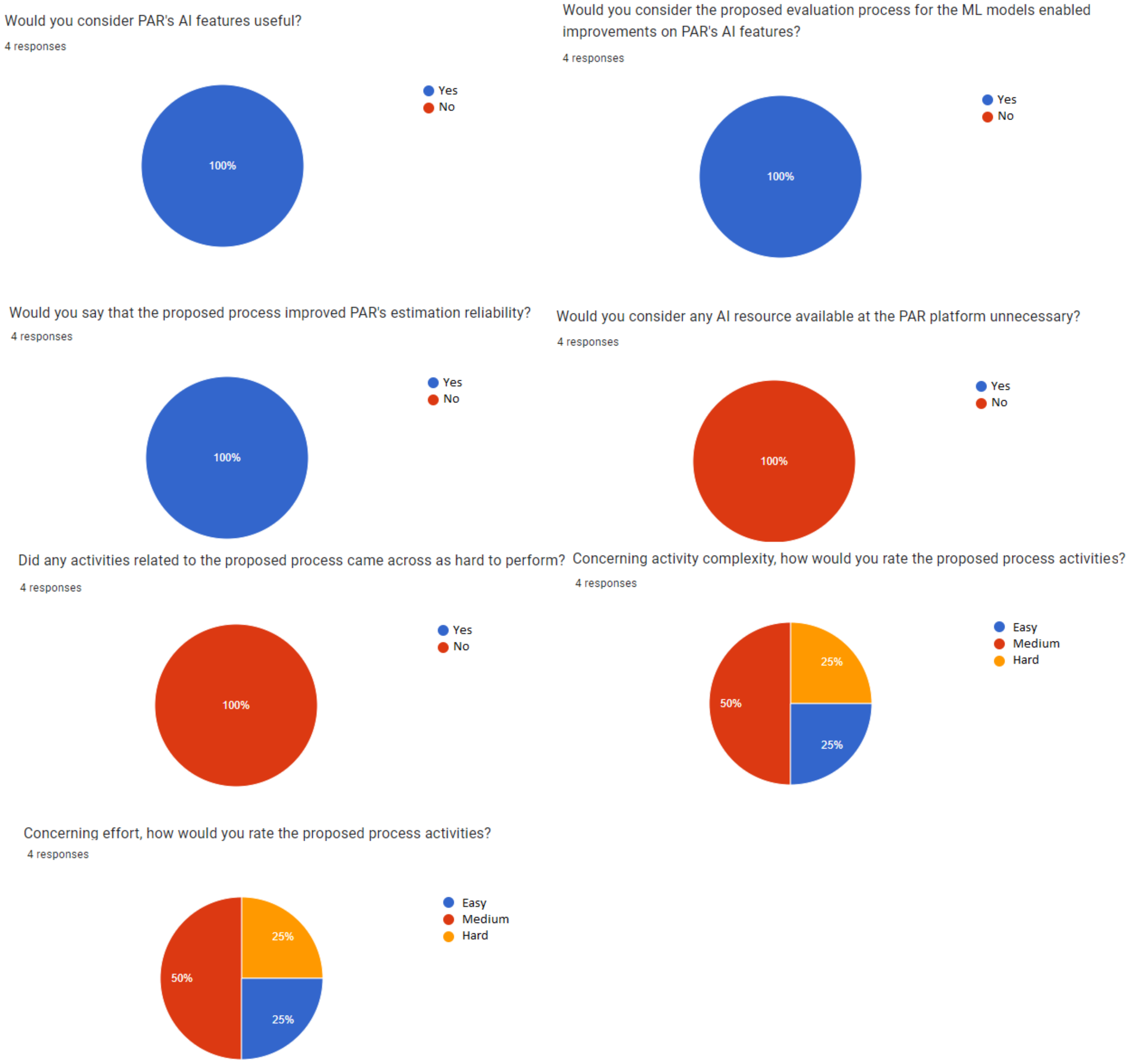
On the subject of benefits of using the proposed process, the IT team stated better model fitness against real-world data, bringing clarity on the path forward for model evaluation, identification of gaps and failure points, better exchanges with the Healthcare team, efficiency in decision making, model evaluation systematization. The Healthcare team stated that easiness of understanding, efficiency, management feedback and multidisciplinary efforts were the greatest benefits brought by the proposed process.

CONCLUSIONS AND FUTURE WORK

Throughout this work, the state-of-the-art for ML model evaluation was analyzed, specifically for Healthcare applications. Based on the best practices identified, an Evaluation Process was proposed, and to validate that proposal, a Case Study was conducted on an Oncology ICU, over ML models in effective use, exposed to real-world, unknown data.

Given specialists' limited availability, much due to the critical-mission scenario of an ICU and its time-consuming day-to-day routine, it wasn't possible to fully apply all proposed activities. Nonetheless, conclusions could be reached, as well as positive outcomes which testify for the usefulness and need for such an Evaluation Process, even if the technical and methodological complexities exposed left plenty room for improvements. Based on the obtained outcomes, the following conclusions could be achieved.

Figure 8: Healthcare team questionnaire responses.



Source: Authors (2023).

RELATED TO ML APPLICATIONS FOR HEALTHCARE

Ensuring interpretable and explainable models builds confidence in a model's decisions. Especially in healthcare applications, it favors interpretation, exchanges with specialists, and favors auditing efforts. Artificial Neural Network techniques don't benefit from an explainable decision process though, given its black-box nature.

Validation costs should be kept in mind when working on model evaluation. For instance, the cost of labeling data would be high, due to the low availability of healthcare specialists, so the process should be automated whenever possible, including integration with other data sources for cross-validation, whenever possible. To that end, investments should be made in integration with other systems and automation in the generation of labels, for instance. The ability to handle missing or corrupted data should also be prioritized. If that's not the case, efforts should be made in user experience whenever manual specialist input is needed. The cost of mistakes made, on the other hand, can be extreme, given that could mean loss of lives and/or permanent consequences to patient's health.

In a Healthcare application context, metrics and scenarios which optimizes efforts should always be prioritized. And, in such contexts, ethical concerns and data confidentiality issues should also be kept in mind, such as biases, discrimination, accountability, transparency, fairness, and privacy. To that end, the adoption of online learning techniques could be beneficial, since it does not store data, which is discarded after the learning is acquired. On the subject of regulatory concerns, it's also important to keep in mind that the FDA is in the process of establishing guidelines targeted at ML applications for Healthcare.

RELATED ML MODEL EVALUATION

Regarding ML model evaluation, there are some important factors to note, such as performance (responsiveness) and scalability, accuracy and reliability of predictions, which is especially critical in healthcare applications, resilience and robustness (capacity to adapt to changes in data or environment), also critical to healthcare applications. The ability to detect errors should also be prioritized and would benefit in both resilience and robustness. To that end, some techniques such as active learning and transfer learning could also be used.

Other factors to keep in mind include defining clear metrics, simplicity and clarity of the models and their learning process, response time and latency, which also includes concerns around delayed outcomes. On an infrastructure level, integration with other healthcare applications and data sources should be considered whenever possible, and mechanisms should be in place to monitor performance over time and refit the model when necessary. That should be an ongoing process, since data and environment can change over time, especially in healthcare applications.

RELATED TO MLOPS AND ML INFRASTRUCTURE

There are also opportunities related to MLOps such as for optimizing the decision process through the usage of visualization tools, model refactoring and version control, observability and alerting opportunities, among others. MLOps can be used to guide and address concerns on infrastructure, CI/CD and change management. Change management practices include code review, test, monitoring, continuous improvement, documentation and regulatory compliance.

CASE STUDY RESULTS

Regarding the Case Study, it is important to state that an option was made early on the study to do "the minimum viable work" given the context and limitations found. There are many opportunities to evolve and expand upon this Case Study, for instance, to other Healthcare ML applications or even within the existing PAR platform models. Nevertheless, the proposed process was useful in filling in the gaps in model evaluation in an effective, methodologic and reproducible way.

It is important to recognize the technical and methodological complexity involved, given the high abstraction levels, multidisciplinary teams, and healthcare specific complexities such as incomplete and noisy data, presenting low representativeness (class imbalance), difficulties in obtaining specialists feedback and participation, and the highly demanding environment in which the work has been performed (Oncology ICU).

Opportunities to expand and iterate over this work could be pursued in data-acquisition and data-source integration with other healthcare applications such as Electronic Healthcare Record systems, and Laboratorial Exams systems, among others, which could raise data and model prediction confidence. In this regard, it's important to recognize ethical and technological challenges, such as proprietary applications and data sources, and unstructured data repositories.

CONTRIBUTIONS AND LIMITATIONS

The current work has provided a catalog of strategies and best practices for ML model Evaluation, specifically targeted at healthcare applications. It has also presented a process aimed at providing evaluation routines and guidelines for creating model-specific evaluation and change management routines. A Case Study was presented containing a report of the experiences in applying the proposed process, as well as testimonies of it’s usefulness.

Some of the proposed activities were not contemplated in the Case Study, though. The Case Study also included 2 models at first, but a greater focus was given to one in detriment of the other. Another limitation was the fact that the Case Study was conducted on a single healthcare unit, an oncology ICU, with a limited set of health conditions and in the same geographical area. Further studies are needed to evolve the process and validate the remaining activities.

FUTURE WORK

New case studies can be conducted over different models and ML for healthcare applications. This could lead to improvements and better detailing for the process and its phases and activities. There’s also a wide range of opportunities to explore in terms of visualization, automated alerting, as well as incorporating MLOps tools and techniques to assist the models' change management, like on monitoring, maintenance and version control.

Another possible route to take would be to expand the proposed process to provide criteria for a maturity level evaluation for ML models, ML applications and institutions that are adopting ML practices and tools. This maturity assessment could include aspects such as continuous monitoring and maintenance, change management, performance feedback, and automation level, among others.

ACKNOWLEDGEMENTS

Thanks to the Ministry of Science, Technology and Innovation (MCTI) and the National Council for Scientific and Technological Development (CNPq) for funding the research through the Call RHAE - Human Resources in Strategic Areas.

REFERENCES

1. Baena-Garcıa, M., del Campo-Ávila, J., Fidalgo, R., Bifet, A., Gavalda, R., & Morales-Bueno, R. (2006, September). Early drift detection method. In Fourth international workshop on knowledge discovery from data streams (Vol. 6, pp. 77-86).

2. Bifet, A., Gavalda, R., Holmes, G., & Pfahringer, B. (2023). Machine learning for data streams: with practical examples in MOA. MIT press.

3. Bin Rafiq, R., Modave, F., Guha, S., & Albert, M. V. (2020). Validation methods to promote real-world applicability of machine learning in medicine. In 2020 3rd International Conference on Digital Medicine and Image Processing (pp. 13-19).

4. Biswas, P. (2021). AL/ML model validation framework. Retrieved from https://towardsdatascience.com/ai-ml-model-validation-framework-13dd3f10e824

5. 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 Medical Informatics*, 10(1), e34038.

6. Chiang, P. H., & Dey, S. (2019). Offline and online learning techniques for personalized blood pressure prediction and health behavior recommendations. *IEEE Access*, 7, 130854-130864.

7. Dawson, R. & Sato, D. (2021). Guide to evaluating mlops platforms [White paper]. Thoughtworks. Retrieved from https://www.thoughtworks.com/en-cn/what-we-do/data-and-ai/cd4ml/guide-to-evaluating-mlops-platforms

8. Deasy, J. O., & Stetson, P. D. (2021). A platform for continuous learning in oncology. *Nature Cancer*, 2(7), 675-676.

9. Marin, H. F. (2010). Health information systems: general considerations. *Journal of Health Informatics*, 2(1).

10. Duckworth, C., Chmiel, F. P., Burns, D. K., Zlatev, Z. D., White, N. M., Daniels, T. W., ... & Boniface, M. J. (2021). Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19. *Scientific reports*, 11(1), 23017.

11. El-Baz, A., & Suri, J. S. (Eds.). (2023). *Cloud Computing in Medical Imaging*. CRC Press.

12. Ettun, Y. (2019). How to apply continual learning to your machine learning models. Retrieved from https://towardsdatascience.com/how-to-apply-continual-learning-to-your-machine-learning-models-4754adcd7f7f

13. Carvalho, A., Faceli, K., Lorena, A., & Gama, J. (2011). *Artificial Intelligence–a machine learning approach*. Rio de Janeiro: LTC, 2, 45.

14. Barroca Filho, I. D. M. (2015). Development of mobile applications based on existing web information systems (Master's thesis, Federal University of Rio Grande do Norte).

15. Food and Drug Administration. (2021). Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device Action Plan. Retrieved from https://www.fda.gov/medical-

devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

16. Flach, P. (2019). Performance evaluation in machine learning: the good, the bad, the ugly, and the way forward. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 9808-9814).

17. 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.

18. He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. *Nature medicine*, 25(1), 30-36.

19. Ian, H. W., & Eibe, F. (2005). *Data Mining: Practical machine learning tools and techniques*.

20. Kitchenham, B., Pickard, L., & Pfleeger, S. L. (1995). Case studies for method and tool evaluation. *IEEE software*, 12(4), 52-62.

21. Kornhouser, W. (2021). Measuring Success of Machine Learning Products: Business Performance vs Model Performance. Retrieved from https://towardsdatascience.com/measuring-success-ef3aff9c28e4

22. Krantz, D. H., Suppes, P., Luce, R. D., & Tversky, A. (1971). *Foundations of measurement* (Vol. 2). academic press.

23. Lee, C. S., & Lee, A. Y. (2020). Clinical applications of continual learning machine learning. *The Lancet Digital Health*, 2(6), e279-e281.

24. Luo, W., Phung, D., Tran, T., Gupta, S., Rana, S., Karmakar, C., ... & Berk, M. (2016). Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *Journal of medical Internet research*, 18(12), e323.

25. Malachi, R. S. (2022). A middleware based on the HL7 FHIR standard for Health Information Systems (Master's thesis, Universidade Federal do Rio Grande do Norte).

26. 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*, 30(4), 433-445.

27. Miranda, L., Viterbo, J., & Bernardini, F. (2022). A survey on the use of machine learning methods in context-aware middlewares for human activity recognition. *Artificial Intelligence Review*, 1-32.

28. Mohandas, G. (2021). Monitoring machine learning systems. Retrieved from https://madewithml.com/courses/mlops/monitoring/

29. Morin, O., Vallières, M., Braunstein, S., Ginart, J. B., Upadhaya, T., Woodruff, H. C., ... & Lambin, P. (2021). An artificial intelligence framework integrating longitudinal electronic health records with real-world data enables continuous pan-cancer prognostication. *Nature Cancer*, 2(7), 709-722.

30. Nishida, K., & Yamauchi, K. (2007). Detecting concept drift using statistical testing. In *Discovery science* (Vol. 4755, pp. 264-269).

31. Panesar, A. (2019). *Machine Learning and AI for Healthcare: Big Data for Improved Health Outcomes*. Apress.

32. Parisi, G. I., Kemker, R., Part, J. L., & Kanan, C. Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural Networks*, 113, 54-71.

33. Pradhan, B., Bhattacharyya, S., & Pal, K. (2021). IoT-based applications in healthcare devices. *Journal of healthcare engineering*, 2021, 1-18.

34. Reagan, M. (2021). Understanding bias and fairness in AI systems. *Towards Data Science*, Mar, 24.

35. Runeson, P., & Höst, M. (2009). Guidelines for conducting and reporting case study research in software engineering. *Empirical software engineering*, 14, 131-164.

36. Sato, D., Wider, A., & Windheuser, C. (2019). Continuous Delivery for Machine Learning-Automating the end-to-end lifecycle of Machine Learning applications.

37. Settipalli, L., & Gangadharan, G. R. (2021). Healthcare fraud detection using primitive sub peer group analysis. *Concurrency and Computation: Practice and Experience*, 33(23), e6275.

38. Stevens, L. M., Mortazavi, B. J., Deo, R. C., Curtis, L., & Kao, D. P. (2020). Recommendations for reporting machine learning analyses in clinical research. *Circulation: Cardiovascular Quality and Outcomes*, 13(10), e006556.

39. Toor, A. A., Usman, M., Younas, F., M. Fong, A. C., Khan, S. A., & Fong, S. (2020). Mining massive E-health data streams for IoMT enabled healthcare systems. *Sensors*, 20(7), 2131.

40. Vollmer, S., Mateen, B. A., Bohner, G., Király, F. J., Ghani, R., Jonsson, P., ... & Hemingway, H. (2020). Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *bmj*, 368.

41. Walsh, I., Fishman, D., Garcia-Gasulla, D., Titma, T., Harrow, J., Psomopoulos, F. E., & Tosatto, S. C. (2020). Recommendations for machine learning validation in biology. *arXiv*.

42. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., ... & Goldenberg, A. (2019). Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9), 1337-1340.

43. Windheuser, C., & Sato, D. (2021). How to get mlops right - tackling the complexity of building and deploying machine learning in your organization. Thoughtworks. Retrieved from https://www.thoughtworks.com/content/dam/thoughtworks/documents/report/about-us/tw_report_how-to-get-mlops-right-aws-and-thoughtworks-report.pdf

44. Wojtusiak, J. (2021). Reproducibility, Transparency and Evaluation of Machine Learning in Health Applications. In *HEALTHINF* (pp. 685-692).

45. Yin, R. K. (2009). *Case study research: Design and methods* (Vol. 5). sage.

46. Yu, C., Liu, J., Nemati, S., & Yin, G. (2021). Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1), 1-36.