JHI JOURNAL OF HEALTH INFORMATICS

CBIS'20 7 a 11 de Dezembro Foz do Iguaçú - Paraná Evento Virtual

Challenges of applying artificial intelligence responsibly at scale in healthcare

Aplicando inteligência artificial escalável com responsabilidade para saúde

Aplicando inteligencia artificial en la salud para escalar con responsabilidad

Amanda Furtado Brinhosa[1], Elizabeth Rocha Fernandes[2], Rafael de Castro Figueroa[3], André Vinícius Rocha Silva[4], Nicholas Roberto Drabowski[5]

Keywords: Artificial Intelligence; Medical Informatics; Social Responsibility

ABSTRACT

Objective: This article aims to present a case study on the main challenges of applying artificial intelligence (AI) in a scalable manner, following the precepts of the research field on responsible AI in the health area. **Method**: During the referred period, data were collected on operation, medical reports and examinations of the telemedicine system studied, which were analyzed and compared in different ways. **Results**: As a result, two different perspectives of the theme were defined, one regarding operational requirements, which are linked to the daily use of such technology, and another to technical requirements, which are related to the essential characteristics of these tools. Both were illustrated through reports and data from practical and real examples. **Conclusion**: From this set of information, some more critical points were raised and discussed, highlighting the caution necessary to work with AI in health and its benefits.

Descritores: Inteligência Artificial; Informática Médica; Responsabilidade Social

RESUMO

Objetivo: Este artigo tem por finalidade apresentar um estudo de caso sobre os principais desafios da aplicação de inteligência artificial (IA) de forma escalável e seguindo os preceitos do campo de pesquisa sobre IA responsável na área da saúde. **Método:** Ao longo do período referido foram colhidos dados de operação, de laudos médicos e de exames do sistema de telemedicina estudado, os quais foram analisados e comparados de diversas formas. **Resultados:** Como resultado foram definidas duas diferentes visões da temática, uma referente aos requisitos operacionais, que estão ligados ao uso diário de tal tecnologia, e outra aos requisitos técnicos, que estão relacionados às características essenciais dessas ferramentas. Ambos foram ilustrados através de relatos e dados de exemplos práticos e reais. **Conclusão:** A partir deste conjunto de informações, foram levantados e discutidos alguns pontos mais críticos, destacando a cautela necessária para se trabalhar com IA na saúde e seus benefícios.

Descriptores: Inteligencia Artificial; Informática Médica; Responsabilidad Social

RESUMEN

Objetivo: Este artículo tiene como objetivo presentar un caso de estudio sobre los principales desafíos de aplicar la inteligencia artificial (IA) de manera escalable, siguiendo los preceptos del campo de investigación sobre IA responsable en el área de la salud. **Método**: Se recolectaron datos sobre el funcionamiento, informes médicos y exámenes del sistema de telemedicina estudiado, los cuales fueron analizados y comparados. **Resultados**: Como resultado, se definieron dos perspectivas diferentes de la temática, una en cuanto a los requisitos operativos, que están vinculados al uso diario de dicha tecnología, y otra a los requisitos técnicos, que están relacionados con las características esenciales de estas herramientas. Ambos fueron ilustrados a través de informes y datos de ejemplos prácticos y reales. **Conclusión**: A partir de este conjunto de información, se plantearon y discutieron algunos puntos más críticos, destacando la precaución necesaria para trabajar con la IA en la salud y sus beneficios.

[1] Bacharel em Engenharia de Controle e Automação, Universidade Federal de Santa Catarina - (UFSC) e Engenheira de Aprendizado de Máquina, Portal Telemedicina, Florianópolis (SC), Brasil.

[2] Pós-graduada em Engenharia de Controle e Automação, Universidade Federal de Santa Catarina - (UFSC) e Diretora de Produto, Portal Telemedicina, Florianópolis (SC), Brasil.

[3] Bacharel em Ciências Econômicas, Universidade Federal de Santa Catarina - (UFSC) e CEO e Cofundador, Portal Telemedicina, Florianópolis (SC), Brasil.

[4] Mestre em Engenharia Mecânica, Universidade Federal de Santa Catarina - (UFSC) e Engenheiro de Dados, Portal Telemedicina, Florianópolis (SC), Brasil.

[5] Mestre em Automação e Sistemas, Universidade Federal de Santa Catarina - (UFSC) e Gerente de Produto, Portal Telemedicina, Florianópolis (SC), Brasil.

Autor Coorespondente: **Amanda Furtado Brinhosa**
e-mail: **amanda@portaltelemedicina.com.br**

INTRODUCTION

One of the biggest challenges of incorporating new technologies in most fields today is reliability. Despite the rapid emergence of modern tools and advanced discoveries by academia, there is a large gap between them and the effective use in real-world applications. There are several reasons for that. When the use of artificial intelligence (AI) is present in such solutions, the discussions get even harder, mainly due to concerns with safety, fairness[1], scalability, and generalization, as previously reported in some studies[2-3].

The use of AI-based technologies in healthcare is even more challenging. It involves not only ethical issues and reliability but also the safety and transparency for the patient and health professionals must be considered. Also, the solutions have to be practical, without further hampering the work of health professionals. Contrasting contexts also need to be considered: in the same region, there may be a wide variety of equipment, technological levels, and types of patients. While solutions that use AI must generalize, they also need to be specialized, which may seem contradictory, but it is essential.

Although the field of pure artificial intelligence and data science grows exponentially, both in academia and in industry, its potential does not reach the same pace in healthcare[4]. The research and development (R&D) of AI-based applications aimed to help or save people's lives require a thoughtful design considering a huge number of details. One must also consider that very simple algorithms with reduced performance can also save lives if delivered fast enough. This article aims to help researchers and professionals to understand this delicate balance between being agile and careful when deploying AI-based solutions.

There are currently several healthcare solutions using AI[5-6] in some parts of their flow, some of which are more focused on the work of health professionals and others focusing on process automation. However, there is no "good practices" guide for implementing, maintaining, and even using these tools. Even though the majority of them are presented as innovations, sometimes with superhuman metrics, when tested in a real-world environment they do not perform well or adapt[7]. When it comes to Responsible AI this difficulty grows[8-9].

The main objective of this article is to present a detailed analysis of the incorporation of artificial intelligence in healthcare and its key challenges. Such information was built upon the experience of the last four years of the research and development team of Portal Telemedicina (2016-2020), which has AI-based solutions running in different scenarios in more than 300 hundred cities in Brazil and Africa. The variety of healthcare institutions being attended by these solutions (500+) made it possible to extract all kinds of data from this highly heterogeneous database and turn it into a benefit for patients and healthcare professionals. This work is expected to provide researchers and health informatics professionals with insights and considerations to create a culture of best practices in the innovation for the healthcare field using artificial intelligence as well as helping others to fast-forward their solutions to real-world applications.

METHODS

This article was divided into three main phases, based on the practical activities carried over the four years considered for this publication: research and development, process automation and validation. After all the experience acquired by the implementation of a telediagnosis system that collected thousands of data, which started in 2013, many demands were raised by the healthcare teams, resulting in the necessity of creation of new solutions, including some AI-based.

In the first phase, several researches and proof of concept were made in parallel to the development of new tools, as is the case of the Artificial Intelligence for Life Analytics (AILA), that is a customized AI platform that unifies all operations related, like training, preprocess and serving. In addition, a central data provider (CDP) was developed to help ingesting and creating data sets and an expert system based on regular expressions, called ARES (Automated REport Structurer), was produced to structure medical reports for creating labels for supervised training of AI models and for comparing predictions with medical reports made by specialist physicians.

Over the period contemplated in this paper, 22 models were developed and put in operation along 2 years on average. They count hundreds of versions and are used to detect 67 findings in 4 different modalities. Table 1 presents an example of findings detected by a version of an AI

Table 1 – Example of findings detected by chest X-ray and electrocardiogram AI models as abnormalities.

Model	Findings		
Abnormal Chest X-ray	Atelectasis	Fibrosis	Pneumothorax
	Cardiomegaly	Hernia	Infiltration
	Consolidation	Mass	COVID-19
	Edema	Nodule	Pleural Thickening
	Effusion	Emphysema	Pneumonia
Abnormal Electrocardiogram	Atrial fibrillation	Wide QRS complex	ST segment elevation
	Pathological Q wave	T wave abnormal	Sinus bradycardia
	Electrocardiographic right bundle branch block	Electrocardiographic left bundle branch block	Ischemia
	Early repolarization	Electrocardiographic left ventricle hypertrophy	Electrocardiographic complete atrioventricular block
	Wolff-Parkinson-White pattern	Electrocardiographic left ventricular strain	Electrocardiographic R wave abnormal

model for identifying abnormalities in chest X-rays and another version for electrocardiograms. The X-ray model has been in operation since May 2020 and it is used together with four other models to improve response time for prioritizing the exams reporting queue in a group of clinics.

The analytical work, using advanced statistical methods and data extraction techniques, were done mainly in the last two years, since the database is growing up constantly. Quantitative and qualitative analyses were performed on the data such as response times (human versus machine), system error rate, machine learning metrics (Area Under Curve, Confusion Matrix, etc), and others.

After the construction of all tools and models cited, the next phase was to automate a lot of processes, mostly according to the research and development team necessities and to accelerate the response time for improvements needed by the health professionals using these new solutions. With that, it was easier to move on to the last phase.

Although the different phases seem to have a specific order, they are not necessarily chronological. In fact, the validation phase is present in almost every step of all the processes and it is also a continuous activity over time.

Throughout the analysis carried out at all stages described above, several critical points were raised. Most of them are technical, but interdisciplinary, what makes the implementation of such innovative solutions in healthcare so challenging. For the purpose of this paper, they were separated into two main categories and presented as basic requirements for overcoming or avoiding these problems.

RESULTS AND DISCUSSION

The results are presented in two sections. First, there are the operational requirements that care about the aspects and worries of daily use of artificial intelligence in healthcare systems, in which all phases described in the Method are observable. Second, there are the technical requirements for using AI in this matter, which consists in more general concepts with the phases more implicit. Both are fundamental and this classification is only in order to better organize the ideas.

Operational requirements

The first fundamental operational requirement, that may sound obvious but is not always considered, is to include health professionals in the development of AI-based solutions. This not only helps them to gain more confidence and familiarity with the tools, but also to create good and useful solutions that can not exist without their expertise.

When taking into account AI models that are generated using a supervised training approach, an essential step is the creation of the annotated data set. Even when not using this approach, at least a test set labeled by experts to measure the model's performance will be required. Here is where the validation process will be first needed, although it is still the phase of research and development.

It is paramount to take care of the quality of the data used both for training and evaluation, including the labels chosen and the data after the annotation step. When the clinical information for creating labels is extracted automatically from medical reports, for example, a common issue can be how this information is written by distinct physicians. Another problem is determining the presence of it, meaning that if it is written that "there are no nodules in the X-ray" the label has to correspond to the absence of nodules and this observation can be put down in several different ways, making it difficult to extract precisely.

Still on that subject, although some studies show that a certain percentage of wrong labels in the dataset are tolerable[10], meaning that it will not affect too much the performance of AI models, in the healthcare field this is actually a really critical issue. When taking into account the difference between other types of pneumonia and COVID-19, for example, in which the findings can be quite subtle in imaging exams[11], if the data set used has a sufficient amount of wrong annotations, it will probably be almost impossible for the AI to distinguish the two conditions.

Related to the data labelling and its validation stage, another problem can be raised. If the algorithms are trained only with knowledge from a specific group of health professionals or type of data (e.g. images acquired from different equipment, with distinct conditions and quality), the AI will almost certainly perform badly in another context, clinic or hospital since there are many scenarios possible. This is known as bias, that is defined as a disproportionate weight in favor of or against something[12] and can cause a lot of trouble[13]. The bias problem is the main reason why the team decided to rely on proprietary data instead of public, because all transformations, processes will be well known and controllable, making possible the mitigation of biases.

One example of this problem is shown in Table 2. A model for detecting effusion in chest X-rays was trained on a variety of data, very mixed, with the purpose of being applicable in a generalized way, reaching 80% of precision. When the model was tested with data labeled by another group of physicians, from a specific brazilian region, the model performed worse, reaching only 39% of precision. Because of that, the best option was to calibrate the model for that particular knowledge, since it would be used only for that case.

Going deeper into the metrics, bringing not only the R&D phase, but also automation one, the classical way to evaluate the performance of AI algorithms is calculating some of them, like specificity, sensitivity and accuracy[I]. They are very important, and illustrate how well the model will operate, but it has to be constantly monitored and the metrics have to be recalculated periodically. It is also

[I] 20 Popular Machine Learning Metrics. Part 1: Classification & Regression Evaluation Metrics (2019). Available at: https://towardsdatascience.com/20-popular-machine-learning-metrics-part-1-classification-regression-evaluation-metrics-1ca3e282a2ce (Accessed: 9 October 2020).

[II] Data leakage is when the training data contains information about the target, but similar data will not be available when the model is used for prediction.

Table 2 – Metrics for a model that detect effusion in chest X-rays tested with different data sets. The specific test set had 5255 patients and 10261 X-ray images. The threshold to maximize precision was 0.91 for the original model and 0.57 for the calibrated model.

Metrics	Original test set	Specific test set	Specific test set after model calibration
Accuracy	80%	97%	98%
Recall	54%	48%	33%
Precision	80%	39%	47%

Table 3 – Metrics for a model that detects abnormalities in electrocardiogram created by Portal Telemedicina's team in July of 2019 and tested again with another test set in September 2020. Both test sets had 10000 images. The threshold to maximize recall was 0.44.

Metrics	July 2019	September 2020
Accuracy	85%	82%
Recall	69%	36%
Precision	41%	34%

important to check if the test set used to evaluate the performance represents the reality well and if it did not suffered any data leakage[II] when creating it.

Although the quality of the data preparation and training process is decisive in the performance of AI models, they are not the unique factor. When AI-based solutions go to operation, companies have to develop supporting systems to have these models making predictions such as Application Programming Interfaces (API's) and preprocessing algorithms. These must also be considered when calculating the metrics. For that reason, the R&D team reapplies new test sets from and in a real-world production environment, which is why it is so important to have automated processes. It has been found that there is always a non-negligible drop in performance due to parsing, programming libraries versioning and communication or integration problems.

Finally, even with all the care in creating and serving the artificial intelligence models, it should not be forgotten about their maintenance, that is the last operational requirement considered for this publication. Technical debts have to be on the team's radar, as well as the bugs have to be fixed as soon as possible, everything with automated monitoring, alerts and visualizations. Beyond infrastructure issues, AI algorithms may need to be retrained including new data or may be specialized for a new scenario to work better.

Artificial intelligence models can also become outdated through time. In Table 3 is shown an example of that. A model for detecting abnormalities in electrocardiogram with only one year and three months old already presented a drop in performance, which reinforces the importance of revisiting all phases described in the method.

As said before, the technical requirements consist in a more general way of conceptualization. Although the operational ones include several technical issues, this section presents aspects more connected to the essential characteristics of AI-based systems, divided into seven categories, and that is why the phases are more implicit.

The first requirement here is **consistency**, which covers the main issues for serving AI models in real time. The main point of attention is that the development flow has to be perfectly consistent with the operation one, also meaning that it has to be checked if the same models behave the same and in the expected way for different hospitals, equipment and even countries. This idea is also relevant for checking if the algorithm is applying the correct models for the right purpose.

This brings the discussion to the next requirement, strongly related to the previous one, that is the **reproducibility** of the results. There has to exist a mechanism that guarantees that everything has the same behavior when applying the same configurations. For this purpose, it is highly recommended the automation of processes that can easily input human errors when testing and evaluating the solutions.

Another crucial thing is the **versioning and traceability** of the models. These can be also covered by processes' automations. It must be known where the data is stored and where it came from, how it was transformed and with what parameters and which code was used at each stage.

Some other challenge in the technical area is the **preservation of data quality**. Although data has to be processed to be used, all transformations applied to it must minimize the loss of information. This is a very delicate point, because not every available data is going to have a good quality and processing it can be disastrous. Portal Telemedicina, for example, receives ECG images that are digital and acquired by modern machines, but also has to deal with images that are cellphone pictures of an exam printed by an analog equipment. Both have to be accepted by the existing AI models, but sometimes they need to be treated differently to work correctly, that is, maintain consistency.

Although all requirements lead to the application of AI in a **responsible** manner, there are some techniques that specifically help the acceptance for the use of these solutions in healthcare. In literature, visualizations that demonstrate what information the model found relevant to make a certain decision is referred to as explainability[14]. This resource can be presented as heatmaps when analyzing images, for example.

It can be a powerful tool for understanding what the models are using as information to its predictions, but also gives more confidence for people that are using it. There are a lot of curious cases that the performance metrics were very high, but when looking at these

visualizations, it was discovered that the AI used worthless information[15].

One of the main and recent concerns about AI, which is another part of the Responsible AI study field, is fairness. In the context of health, this is even more worrisome and is also related to the bias problem mentioned in the previous section. A model has to treat the inputs that it receives equally, meaning that it has to work well in all populations from that specific application.

For example, if an algorithm prioritizes a patient care queue by evaluating X-ray images, it can not use the patient sex or age as a fact for prediction decision, just the image itself. In this case, it is very important to check the epidemiological profile of the available data, studying the sex, age and another features distribution and try to balance possible differences, turning the model more fair. Not many solutions care about this issue yet[16], but in healthcare its importance is growing fast and validation phases can be a great ally in this analysis.

Another characteristic related to the main topic of this publication is **scalability**, that should cover every achievable dimension of the problem wanted to be solved. The solution has to scale in training, accepting various amounts of data and processing it as fastest as possible. It also has to expand in different categories, like modalities and specialties, and in distinct data formats, like PDF and PNG.

It should be considered that the solutions have to scale in types of AI models and their versions as well, being easy to compare and update them. Additionally they have to support simultaneous development of new features by the R&D team, accept different types of architectures and client users.

The last fundamental requirement raised is related to the **antifragile** philosophy[17]. It is impossible to predict all problems that can happen when in operation time. There has to be a constant monitoring of errors, evaluation of metrics (including time for responding to a request to the AI model) and maintenance of the models. This brings again the importance of keeping everything consistent, automated as much as possible and well validated.

For the sake of closure, it is important to highlight that this paper is not based on hypothesis tests, as well as other formal statistical techniques, due to the practical context encouraged by the agile methodology used by the R&D team. The fast changes and constant version releases did not allow the researchers to present a more rigorous scientific analysis.

Additionally, it would have been very enriching to present comparisons between the brazilian and african scenarios. Considering the variance in exams, equipment and also because of the natural ethnic and genetic differences, that can probably lead to very distinct fairness and metrics when comparing these populations, a deeper study may be more appropriate.

CONCLUSION

The growth of artificial intelligence applied to healthcare is undeniable. The continuous improvement of AI architectures and new techniques show the great potential for this technology to contribute to facilitating the physicians workflow and, in consequence, saving more lives and more accurately. It is the time for healthcare professionals to interact daily with such systems, so it is of extreme importance that this contact be positive. On the other side, the technology adoption can potentially be delayed in several years if the systems being delivered nowadays do not follow the minimum requirements described in this paper.

Even though AI can be a powerful tool for health professionals, it should be noted that it is just that: a tool. Well defined processes are much more important than standalone algorithms.

Some benefits of artificial intelligence models are their capability of analyzing in real time large batches of data with low cost, resource optimization, process automation and even more. However, there are some human skills, like complex decisions based on years of practice, that AI probably will never have and this is the reason why health professionals always have to participate in the development of these tools, using them in their favor.

Finally, when artificial intelligence is used and incorporated into healthcare systems with all considerations presented in this article, focusing on research and development, process automation and validation phases, it can be really useful. It is a great challenge to rethink under this new perspective, but it is a necessary challenge for AI to be used responsibly and to scale to impact and save more lives.

ACKNOWLEDGMENTS

The AI system described in this article was developed with support of the *Fundação de Amparo à pesquisa do Estado de São Paulo (FAPESP)* in the project number 2016/10374-3, and the warehouse that enabled the analytic work was developed under the project number 2017/25238-0. We also would like to thank all health professionals that gave us feedback and great insights along this journey.

REFERÊNCIAS

1. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635. 2019 Aug 23.
2. Beede E, Baylor E, Hersch F, Iurchenko A, Wilcox L, Ruamviboonsuk P, Vardoulakis LM. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. InProceedings of the 2020 CHI Conference on Human Factors in Computing Systems 2020 Apr 21 (pp. 1-12).
3. Stoica I, Song D, Popa RA, Patterson D, Mahoney MW, Katz R, Joseph AD, Jordan M, Hellerstein JM, Gonzalez JE, Goldberg K. A berkeley view of systems challenges for ai. arXiv preprint arXiv:1712.05855. 2017 Dec 15.
4. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nature biomedical engineering. 2018 Oct;2(10):719-31
5. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y. Artificial intelligence in healthcare:

past, present and future. Stroke and vascular neurology. 2017 Dec 1;2(4):230-43.
6. Gonçalves, Carol. Inteligência Artificial na saúde: aplicações, benefícios e ameaças. Revista Hospitais Brasil. 2019 May;97:10-20.
7. Henriksen A, Bechmann A. Building truths in AI: Making predictive algorithms doable in healthcare. Information, Communication & Society. 2020 May 11;23(6):802-16.
8. Clarke R. Principles and business processes for responsible AI. Computer Law & Security Review. 2019 Aug 1;35(4):410-22.
9. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, Jung K, Heller K, Kale D, Saeed M, Ossorio PN. Do no harm: a roadmap for responsible machine learning for health care. Nature medicine. 2019 Sep;25(9):1337-40
10. Frénay B, Verleysen M. Classification in the presence of label noise: a survey. IEEE transactions on neural networks and learning systems. 2013 Dec 17;25(5):845-69.
11. Li X, Fang X, Bian Y, Lu J. Comparison of chest CT findings between COVID-19 pneumonia and other types of viral pneumonia: a two-center retrospective study. European radiology. 2020 May 12:1-9.
12. Lexico. Oxford English and Spanish Dictionary, Thesaurus, and Spanish to English Translator, URL: https://www.lexico.com/, Archived on September 30th, 2020.
13. Kaushal A, Altman R, Langlotz C. Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. JAMA. 2020;324(12):1212–1213. doi:10.1001/jama.2020.12067
14. Belle V, Papantonis I. Principles and Practice of Explainable Machine Learning. arXiv preprint arXiv:2009.11698. 2020 Sep 18.
15. Wieland Brendel and Matthias Bethge. 2019. Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. arXiv:1904.00760 [cs.CV].
16. Piano SL. Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. Humanities and Social Sciences Communications. 2020 Jun 17;7(1):1-7.
17. Taleb NN. Antifragile: Things that gain from disorder. Random House Incorporated; 2012 Nov 27.