Interpretable Machine Learning in Clinical Practice

Wellington Cunha
New Jersey Institute of Technology
NJIT ID 31548454 – NJIT UC ID: wc44
wc44@njit.edu

Abstract

Machine Learning (ML) and Artificial Intelligence (AI) models have been permeating every aspect of our daily lives. They define the information we consume, how we consume goods, how we pay for them, how we move around, an even how we date. But there is still one area where it is still trying to show its worth and become a commodity: the health care space. Not that ML and AI are not being used in life-sciences in general, as it is widely used on genomics mapping, drug design and development, or how we pay for medical services. But when it comes to the decision-making process related to clinical practice, including disease diagnosis, there is still a lot of resistance, both from patients, as well from practitioners. And not even embedded solutions (like software embarked on medical devices and equipment) are free from this resistance. Among several reasons, one of the main is that most people still don't trust in machine-lead decisions for high-stake matters because they simply don't know how the machine is making those decisions. Opening those black boxes and making them intelligible for the audience (patients and practitioners) is probably the best way to remove those barriers and accelerate the adoption of Machine Learning and Artificial Intelligence based solutions for clinical practice.

1 Introduction

In recent years, with the huge increase in computational resources and the amount of data available, technology has been seen as a potential game-changer when dealing with health-related problems. It has the potential to augment the capabilities of practitioners (doctors, radiologists, physical therapists, etc.), not only to reduce the costs associated with providing health care services, but as well to improve the outcomes, like for instance by helping in early detecting diseases. Not to mention the possibilities in personalized treatments, drug development, and several other related fields.

The possibilities around the combination of sources of data, the amount of data being currently generated, and the knowledge accumulated by several researchers and clinical practitioners surpass, by far, the capacity of any human being alone. Already available data, such as biometrics information, can be combined with data produced for specific purposes (like blood tests, CT scan, X-Rays, etc.) and, along with diagnosis provided by thousands, eventually millions of practitioners, be used to train algorithms to detect and even predict the occurrence of diseases.

But, although the reliability and accuracy of those algorithms are being proved – sometimes even surpassing the accuracy of actual practitioners [1] – there are still a lot of concerns and resistance in using solutions based on Machine Learning and Artificial Intelligence.

From the patients' end, besides missing the human interaction, they are very concerned with privacy and how the patient's choice will be respected [2]. As have happened with any new technology introduced into medicine field (such as the very first vaccines), there is a natural resistance in using what is new and unknown.

Although smaller as the patients' concerns, there is also resistance from practitioners. Besides the fear of being "replaced by a machine" [3], practitioners have also been reluctant in using Machine Learning and Artificial Intelligence based solutions due to the lack of knowledge on how the models get to a diagnosis or prediction. Even when the results of a model are extraordinarily accurate, clinicians are likely to view any "black box" with suspicion [4].

Ethical concerns, like biases contained in the data or in the knowledge used to train those algorithms or in decisions that are not binary and would involve picking the "less-worse choice" (such as saving the baby or the mother. Or keep trying options to cure a patient or prescribe palliative treatments) are other factors for resistance.

The question of how algorithms and models work have recently become a legal matter: European Union's General Data Protection Regulation (GDPR) ensures "right to explanation" to data subjects (patients, customers, applicants, defendants, etc.) on how decisions made by algorithms in relation to them are made [5, 6]. The regulation states that data subjects have the right to ''meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing [automated decision making]" [7].

GDPR even goes further and ensures that "the data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her" [8].

Although not all the concerns can be addressed by developers and researchers in Machine Learning and Artificial Intelligence field, developers and researchers can certainly do their part in breaking some barriers to reduce this resistance. By shedding light into how the models are built and how they come to conclusions, patients and practitioners would feel more and more confident and comfortable to adopt solutions using those technologies.

By building more transparent, interpretable, and explainable systems, developers and researchers into Machine Learning and Artificial Intelligence space can help patients and practitioners in understanding and, therefore, trusting these solutions [9]. All of that while meeting regulation requirements. Interpretable Machine Learning principles and concepts are crucial for that purpose.

2 Possibilities of Machine Learning in health care

The use of computer-based solutions and applications, in their various forms, has been changing every aspect of our lives. In health care space and life sciences it is not different. Those solutions are helping in expediting the development, discovery, and design of new molecules; in simulations on how they will interact with other molecules; in analyzing how they will be interacting with our genes, proteins and/or pathogens, alone or along with other factors [10, 11].

Machine Learning models have been helping in the clinical trial process, starting by helping in the designing of trials, then in the collection of evidence, the analysis of the data and even to prepare the reporting of results to regulatory agencies [12]. The use of those solutions in procurement, sales, and distribution (supply chain); administrative tasks (like clinical documentation); and claims management has become a kind of commodity, in the meaning that no one can even imagine a health system existing without those solutions nowadays.

One of the areas where machine learning could have a huge impact on health care is in clinical practice [13]. But it is still struggling to gain traction there. Even in the automation of clinical tasks or in prioritizing the triage in emergency rooms, areas where the inputs and outputs are well known, there is a lack of confidence in using machine learning solutions for high-stake decisions [14].

If even in those processes that are well-known and most of the times oriented by a sort of "algorithm" (a procedure) developed and executed by humans, Machine Learning solutions receive pushbacks, in some advanced topics into clinical practice (such as precision medicine, personalized care, early detection, more accurate diagnosis and recommendations for the course of actions in the treatment of diseases) they are even farther from being accepted and popularized among clinical practitioners [15], especially due to the lack of transparency on how conclusions are made by those solutions.

3 Opacity vs transparency: the "black box" issue

Durán & Formanek, (2018) [16] set the grounds for the discussion about the opacity vs transparency by stating that there will always be some level of opacity in any procedure, including Machine Learning algorithms. That happens even for types of models that are, in theory, self-explanatory, like decision trees. The rationale behind this statement is that the level of opacity depends on the actor interacting with the solution or procedure. Keeping on the decision tree example, the statistician or data scientist developing the model will sure be able to understand the criteria used for splitting the branches (like, for instance, information gain), but that will be somehow unintelligible for a doctor.

On the other hand, the model developer (statistician or data scientist), although having deep understanding on how the algorithm built the tree for of a model, would have little (to none) knowledge on the causal relation that led to the diagnostic of a certain disease based on the input factors, as the doctor (the field expert) has.

Some model types, like decision trees and linear regressions, have transparency as something intrinsic to them [17], being considered transparent models. But even those model types have their limits, as they are interpretable only to some extent. A huge tree with several branches and layers will become unintelligible for humans, even for developers [5].

That is also one of the reasons why Deep Learning models, which are considered as an advanced Machine Learning method, are one of the model types that is most mentioned as "black box". As they are highly recursive, it is hard for humans to understand the causal relation between input and output [8].

Other factors that make some models less transparent are the use of hyper-parameters and feature engineering, which are even more difficult to translate to non-developers [5, 18]. Ren (2020) [19] proposes that we can use Machine Learning and Artificial Intelligence models to generate transparency to black box models. In opposite to that, Miller (2019) [8] argues that instead of trying to explain black box models, we should focus on building interpretable and explainable models since the beginning of the development, preferable during the problem statement phase (interpretability should be one of the goals of the projects), at least for high-stake decisions (as in health care and criminal justice), otherwise we will be only making the issue bigger.

We can certainly start the development of new solutions with that in mind and build algorithms that would produce knowledge while learning (especially for Deep Learning models during the feature extraction). But certainly, that will require more resources: humans (during the development) and computing (when training the model), especially when developing models with non-structured data (like images, audio, and video). But for some cases, the benefits may outweigh the costs [18].

Durán & Jongsma (2021) [15] propose a totally different approach, arguing that black box algorithms are less problematic than the academia believes. Their proposal is that, instead of focusing on transparency, we should focus on what they call "reliabilism" and trust. By giving transparency on how the algorithms are designed and implemented, it would show that they are reliable and, therefore, can be trusted. This is exactly how the drug development process works nowadays: there is a process to be followed, that will ensure that the drug is safe and reliable and, by knowing that the process was followed, practitioners and patients can trust in what the drug or therapy is supposed to provide.

Either by starting the development of solutions with transparency in mind or using other means for giving transparency. Or even in the case of making the development process transparent itself, one thing seems to be a consensus: developers should not expect that non-developers will blindly accept the results of their models. And it is in the hands of developers increasing that transparency (at least to the minimum acceptable level) by making their work and the results of their work interpretable.

4 What is interpretability in Machine Learning?

Interpretability is a problem as old as Machine Learning itself [6]. But what is interpretability in the context of Machine Learning? According to Murdoch *et al* (2019), Interpretable Machine Learning is the ability (or action) of extracting knowledge contained in the data with the use of Machine Learning models; or the ability (or action) of extracting knowledge learned by the model itself during training [20].

They argue that Machine Learning models should be used not only to produce prediction, but as well (and sometimes mainly) to produce knowledge. That includes the use of Machine Learning models for the solely purpose of generating knowledge about a problem [20]. Therefore, interpretability is a way of giving transparency to what the model has learned. And that leads to explainability.

Miller (2019) uses interpretability and explainability interchangeably as "the degree to what extent a human can understand the cause of a decision from a Machine Learning model". Citing Lewis, Miller defines that "To explain an event is to provide some information about its causal history. In an act of explaining, someone who is in possession of some information about the causal history of some event – explanatory information, I shall call it – tries to convey it to someone else" [9].

Holzinger *et al* (2019) [6] also define interpreting and explaining as "means to provide the causal relationship of 'why', in a logical manner ('this led to that')". And Vellido (2020) [7] defines interpretability and explainability with a subtle difference between them: interpretability alone puts the practitioner as the active actor/component, by having him/her "interpreting" what the model is providing (and that interpretation may pass thru some biases filters), while explainability makes the model itself as the active actor, leaving not much room for interpretation. At the end, in the context of Machine Learning, Holzinger *et al* (2019) [6] defines interpretability as the "capacity of a model to explain itself to the intended audience (because 'transparency is not enough')".

This definition takes interpretation in Machine Learning to another level, as the model must be transparent and explain itself according to the audience and the context, therefore generating explanations for different personas [5]. The interpretation needs to be, somehow, translated to the terms and jargons of the practitioner interacting with it. And in health care this adds another layer of complexity, as the same type of information (a prediction or recommendation) needs to be translated to several different personas, such as doctors, nurses, staff planners, health insurance companies, and so on.

Summarizing those definitions in a general statement to be considered in this paper from now on, we can say that "interpretability is the ability of the model to explain itself to different kinds of personas that interact with it, without giving margin for different interpretations and considering all the factors taken in account when providing its outputs (e.g., causal relationship, feature engineering, hyper-parameters, etc.)".

5 Generating interpretability

Almost all authors split the process of producing interpretability into two classes: model based (also called ante-hoc by some of them) and post-hoc [20]. Model based approach happens when the knowledge is obtained, and interpretability is produced while the model itself is being developed. Although this approach tends to produce better results in terms of interpretability/explainability, as previously mentioned, this requires more resources. Another downside is that the interpretability is exclusive for that specific model.

Post-hoc happens when the knowledge is obtained after the model is developed (by analyzing the results and how the module came to them). Usually, a second model is created to interpret/explain the results of the first one (the black box). One advantage of post-hoc is that the interpreting model is model agnostic in relation to what is interpreting [6] and can be reused. But on the other hand, as also briefly mentioned, they can pile up problems and the second model can be as hard to explain (and be trusted) as the first one [8].

In both cases, models should be able to give two types of interpretation. Global interpretation refers to how the data used for training generated the model at the population level (which factors were considered when building the model and the respective weight for each one of those factors). Personalized or local interpretation refers to how the trained model gets to a result for a unique new case, explaining why, for that specific case (which was not used during the development), the model generated the output [17, 18]. While Global predictions are easy to provide explanations for, Local predictions are hard to interpret/explain.

It is almost an axiom in Machine Learning that simpler models (like Linear Regression or Decision Trees) produce results that are not so good as complex models (like SVM – Support Vector Machine or Deep Learning) in terms of outcomes (accuracy, precision, recall, etc.). However, the simpler the model, the easier it is to interpret or generate interpretability for, while the complex the model, the harder it is to interpret or generate interpretability for. Murdoch *et al* (2019) define two metrics to evaluate the efficacy of the model in predicting the results and the ability of them to generate interpretation: descriptive accuracy and predictive accuracy [20].

So, based on what we have seen in this section so far, we can say that ante-hoc interpretable models tend to have high descriptive accuracy and low predictive accuracy. For the post-hoc, it is exactly the opposite: they tend to have high predictive accuracy and low descriptive accuracy. But those accuracies are also affected by the complexity of the model: the simpler the model, the higher the descriptive accuracy and the lower the predictive accuracy; the more complex the model, the lower the descriptive accuracy and the higher the predictive accuracy. And the level of predictions (Global or Local) also interferes on those accuracies: Global predictions have high predictive and descriptive accuracies while Local predictions has low predictive and descriptive accuracies.

		Descriptive Accuracy	
		Low	High
Predictive Accuracy	Low	Local	Ante-hoc approach Simple models
	High	Post-hoc approach Complex models	Global

Figure 1: Comparison between predictive accuracy vs descriptive according to the interpretability producing approach, prediction level and complexity of models.

Therefore, we can say that the model chosen (simple or complex), how interpretability was generated (ante-hoc or post-hoc) and the level of use for predictions (local or global) are the dimensions that determine the level of both description and prediction accuracy.

6 Trade-offs between description and prediction accuracies

So far, we have seen that more interpretable models often have less predictive accuracy than less interpretable models [5]. That implies that there are always trade-offs between description accuracy and prediction accuracy. But also, there is always a limit line: if the prediction accuracy is too low, it doesn't matter how high the descriptive accuracy is [20]. And for some cases, it does not matter how high the predictive accuracy is, if the model cannot be interpreted, it would serve no purpose. These assumptions are shared by almost all authors.

It is possible to obtain high prediction accuracy and high description accuracy, but those models are, in general, too unstable and can rarely be used for predictions purpose. Usually, they serve only the purpose of explaining causal relation [20].

Besides the trade-offs between description and prediction accuracies, another factor to be taken in consideration when evaluating the trade-offs are the extra-costs (computing, time, etc.). That is valid when developing the model, training it and, especially, when the model is being used to generate predictions or knowledge [21]. For instance, a solution that would take hours to generate its outputs would be of little help in some applications, like Emergency Rooms.

Field experts should be involved not only in designing how those interpretations would be given, but they should ultimately be the ones making decisions about the trade-offs [5]. For some cases high accuracy may be the priority and, for others, high interpretability can be the goal. When developing Machine Learning models, developers usually have the objective of reducing the error, but the real-world purpose of a model is usually different [18]. Another point to take in consideration is that, when referring to predictive accuracy, it does not mean exactly accuracy, as traditionally used in Data Science field, as sensitivity or positiveness might be more important [17].

Vellido (2020) [7] reiterates that interpretability is very important in the field of health care for the adoption in practice. However, there is a need for integrating the health care experts in the designing of data analysis and interpretation and that they will be the ones making decisions about those trade-offs.

Among all the authors consulted for this paper, Rudin (2019) [8] is the one that goes against the flow and states that the trade-offs between accuracy and interpretability is a myth. According to Rudin's work, the trade-off is not true, as some more simple and explainable model can be as much as accurate as complex black box models. Or at least the level of accuracy is not that large, and one can give up some accuracy in favor of interpretability without much difference on the results at the Local level. And for some cases, where the models are used as an aid to provide information and evidence to practitioners, more interpretability can lead to more predictive accuracy in the final outcome.

In any case, one assumption that seems to be a consensus among all the authors is: transparency is not enough.

7 How to communicate the results of a model

Miller (2019) [9] suggests that, when tailoring the interpretation/explanation of models to each specific audience, we should use some lessons learned from social sciences. The major findings obtained from research on how to communicate results in those sciences are:

- Humans want to contrast explanations, which means, they want to compare the results and know why this was chosen and not that. Interpretations may also provide explanation by showing similar cases (or features) that had the same outcome as the prediction [18].
- Humans do not want all the complete explanations. They want one or two causes (among several presented), which are usually selected by them, and this selection is heavily influenced by their cognitive biases.
- Probability matters little: although the likelihood is important, referring to probabilities is not effective when communicating with the general public. And even for specialists, probabilities will be taken into consideration after other factors, as a way of backing them up.
- Explanations are social: they are a form of knowledge transfer. And this transfer will happen better in the form of a conversation, an interaction or a good tale being told, rather than charts, tables, numbers, formulas, etc. As in probability, those assets (charts, tables, etc.) will be used only to support a decision already made considering other factors.

Miller (2019) [9] acknowledges that these assumptions are not feasible for all applications and audiences, but in several cases, they have the potential to improve the explanation. We should, as much as possible, start with them and then craft the message to the intended audience. And observing how humans would explain decisions to each other also provide a good starting point on how to communicate the results.

Another point to consider, for the specific case of health care, is that health care workers are usually overwhelmed with the number of patients, with a lot of information to read and to write (on medical charts and EMR systems). If not delivered correctly, in a manner that improves the outcomes and crafted considering the characteristics above, both for the patients as well for the practitioners, providing interpretations may be seen as only another burden for them [5].

Considering these factors, we can say that the use of Natural Language Processing (NLP) models could be used to translate the predictive models and their decisions on a kind of story to be told to practitioners. And in order to reduce their burden, those models could even "speak" to them (a good number of doctors already use speech-to-text solutions when filling up charts – text-to-speech would not be a totally unknown technology for them). We also should consider the fact that every human being has a limited cognitive capacity, and we need to find ways to provide interpretation for high-complex models in a way that does not overwhelm the practitioners.

Holzinger *et al* (2019) [6] highlights that one important factor when generating interpretability/explainability is to inform the sequence of events that lead to a certain (local) prediction/prescription. This is very important in the case of models that use non-structured data, such as in image analysis, as creating interpretation for how a computer vision models get to the conclusions is a task that is almost impossible.

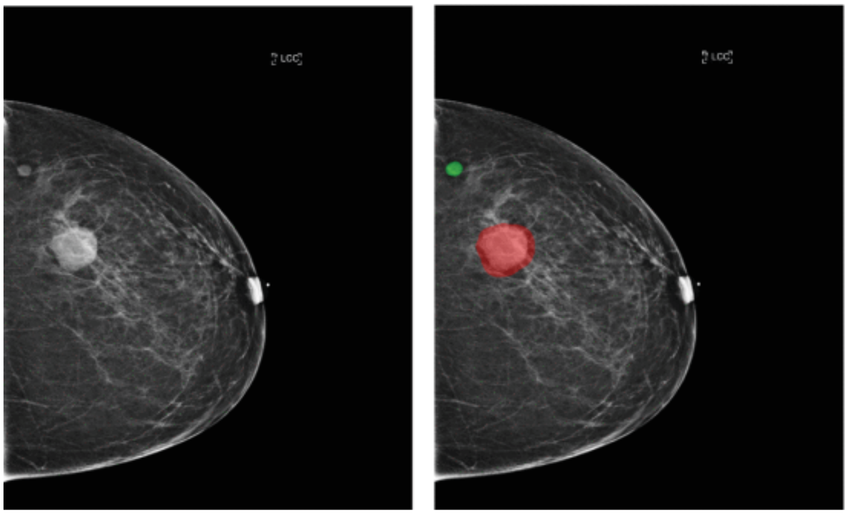


Figure 2: Mammogram “tagging” potential benign (green) and malignant (red) tumors that can lead to the diagnostic of breast cancer.

In the specific case of image analysis, where a small piece of information (for instance, a local edema or nodule), can lead to a global diagnosis (as per the previous example, pneumonia or breast cancer, respectively), using the same image, annotated, is a good way of providing interpretability. Fig 2 provides one example of “tagging” potential benign (green) and malignant (red) tumors on mammograms [22] and Figure 3 one example of using masks to show potential local (lung opacity) and global (pneumonia) dignostics on chest X-rays [23].

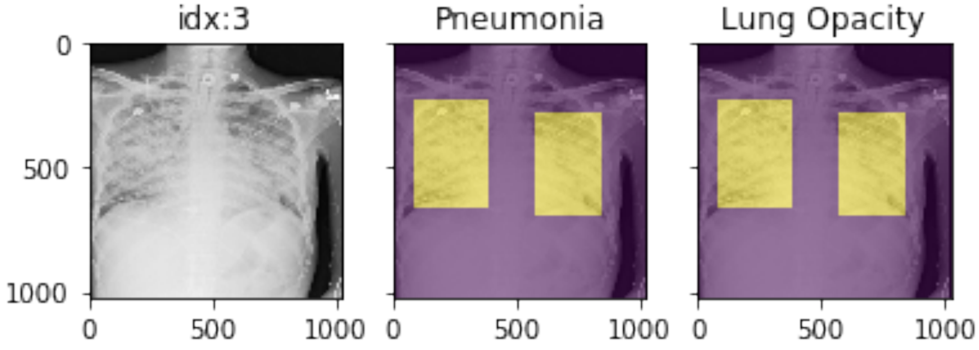
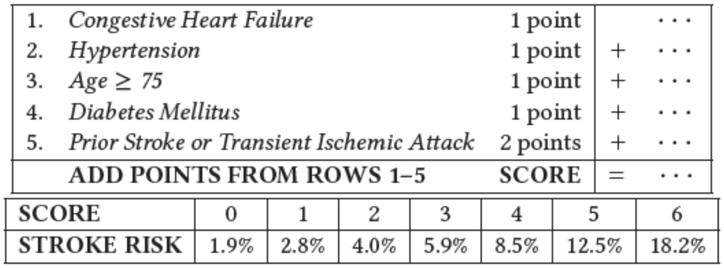


Figure 3: Chest X-ray with masks for local (lung opacity) and global (pneumonia) mask applied.

The images (Figure 2 and Figure 3) above also show another factor to be considered when providing explanation. According to Lipton (2018) [18], an interpretable model should have as one of its characteristics, what is called “decomposability”, which is the ability to show each phase or step of the prediction process. Or any criteria used to come up to a conclusion.

Stiglic *et al* (2020) [17] suggest that a set of rules (pretty much as in association rules) that describes why a prediction or assessment was made is very effective in explaining the decisions made by models at the local level. In the same line, Ustin & Rudin (2017) [14] suggests using the most important factors (features), converting them to a risk score (as in Figure 4), which is something already disseminated in clinical practice. As stated before, one just needs to be wary about adding too many levels and complexity, which will make the risk score itself unintelligible.



1.	*Congestive Heart Failure*	1 point		· · ·
2.	*Hypertension*	1 point	+	· · ·
3.	*Age ≥ 75*	1 point	+	· · ·
4.	*Diabetes Mellitus*	1 point	+	· · ·
5.	*Prior Stroke or Transient Ischemic Attack*	2 points	+	· · ·
	ADD POINTS FROM ROWS 1–5	SCORE	=	· · ·

SCORE	0	1	2	3	4	5	6
STROKE RISK	1.9%	2.8%	4.0%	5.9%	8.5%	12.5%	18.2%

Figure 4: Risk score to assess the chances of a cardiac arrest.

For models with too many important factors or in the case where decisions are not binary, Stiglic *et al* (2020) [17] suggest the use of SHAP models (Figure 5) to show what factors were considered for local level predictions. Feature importance charts (Figure 6) can also be used to show how each factor (feature) was considered at global level.

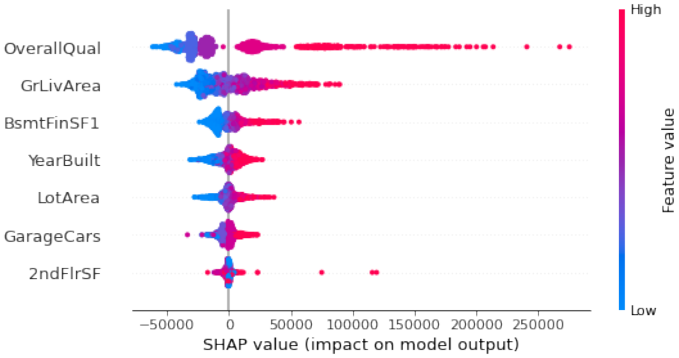


Figure 5: SHAP model template showing the impact of factors / features.

One point still under discussion is how to deal with uncertainty in the model [6]. The main open question is how to communicate the errors contained in every and each model? And Ghassemi *et al* (2021) [13] propose that, for some cases, we should move from just the explanation to justification, in the meaning that, after providing the outputs, the model also provides a justification on why that output was generated. It would be mainly a change of tone in order to put the solution as an aid, and not as the main actor.

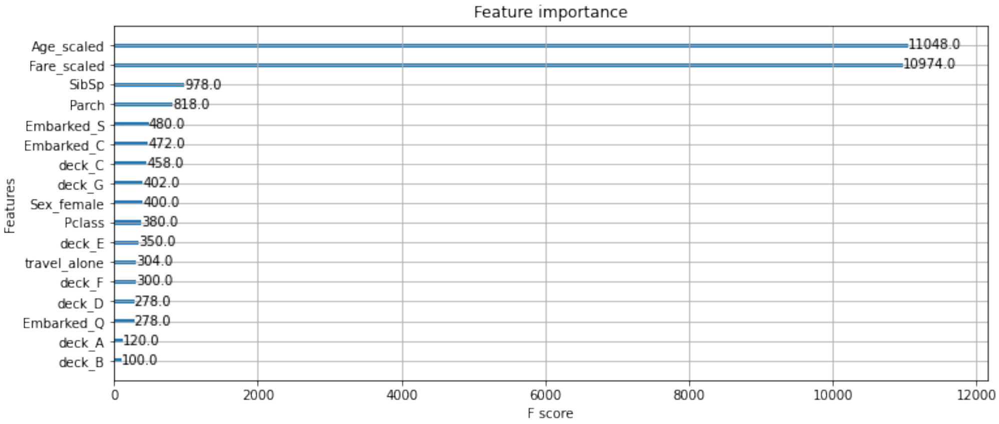


Figure 6: Feature importance map template showing the importance of each feature when building the model.

Miller (2019) argues that studying (and thereafter modeling) the ways humans communicate decisions and behavior to each other provides good insights on how to create explainable models. And that includes biases and social expectations, that can be valuable in learning how to communicate the results [9]. This would lead to somehow a kind of personalization in the interpretability: models generating interpretations would tailor the way of interpreting to each one of the practitioners interacting with them.

A question that we pose for future discussion is: why do not engage User Interface/User Experience (UI/UX) designers since the very beginning of the process of developing a Machine Learning solution?

8 Other challenges

Besides the issues and challenges with interpretability already mentioned, there are others to take in consideration. Probably the first one is how to evaluate the interpretation methods [20]. For sure we could involve the field experts in the process of assessing the descriptive accuracy [7], but given the fact that the ability of explain measured from the target audience is very subjective, how to define a method that could be used to assess the goodness of the explanation? And taking one step ahead, how to measure models' ethics and legality compliance [18]?

Another challenge is in relation to biases. First question in this matter is how to identify the existence of prejudice biases contained in the training dataset? And after identifying the existence, if and how to correct for those biases? What if a model predicts that, for young black males, the root cause of a cardiac arrest is an overdose, because most practitioners treat cardiac arrest on young black males as an overdose in emergency rooms? Should developers try to adjust (correct) the models for that?

And finally, there is the question of responsibility [15]: who should be responsible for a bad decision made by or with the help of an algorithm? Who should be held accountable?

9 Conclusion

The popularization of Machine Learning solutions in clinical practice still has a long way to go. But one thing is known for sure: while practitioners and patients cannot understand how those solutions work, there will always be pushbacks in accepting and adopting them. There are some considerations that are almost consensuses among researchers in this field and that can help in making models more transparent and, therefore, help in reducing the resistance.

The first one is that developers should start thinking on how their models can be interpreted (or can explain themselves) since the very beginning of development – and even during the hypothesis formulation – and that should be one of the goals of Machine Learning projects in health care space. Rudin (2019) [8] proposes that "no black box should be deployed when there exists an interpretable model with the same level of performance" in a way to change the business model in favor of explainable models. Rudin (2019) goes even further, asking for this to become part of regulations because that would be a way to hold accountable companies developing and deploying models for high-stake decisions without accounting for interpretability.

The second consensus is that we should always rely on the experts (not just in health care, but in any field) to validate the models, not only for explainability (descriptive accuracy), but as well for the level of performance (predictive accuracy). Those experts should be the ones making the final decision about the acceptable levels of performance and explainability.

The third consensus is the one that is most stressed by researchers: Machine Learning based solutions must assist, and not replace practitioners [5]. The practice of medicine is highly expert-dependent. The expert or experts should be the ultimate deciding authority on a given subject. For highly sensitive medical decisions the decision is usually made in the context of plurality of options of experts, and technology should only offer support [16].

Durán & Jongsma (2021) [15] make a strong case for not allowing Machine Learning and Artificial Intelligence models to make decisions. Although at the same time they defend that the model does not need to "explain" but be reliable. And as the model would only provide evidence for clinical decision making (and not decide how it should be acted), that evidence would be consequently self-interpretable.

Another rational for the consideration that models should not make decisions is that clinical decision-making process should have the participation of the patient (or someone on his/her behalf). By the end of the day, the patient should have autonomy to even decline a treatment and that is another strong reason why Machine Learning and Artificial Intelligence based solutions should work to produce information and evidence, and not to make decisions.

Providing information and evidence to patients is another strong case for interpretability in the context of Machine Learning in clinical practice.

References

[1] Richens, J. G., Lee, C. M., & Johri, S. (2020). Improving the accuracy of medical diagnosis with causal machine learning. Nature communications, 11(1), 3923. https://www.nature.com/articles/s41467-020-17419-7

[2] Richardson, J. P., Smith, C., Curtis, S., Watson, S., Zhu, X., Barry, B., & Sharp, R. R. (2021). Patient apprehensions about the use of artificial intelligence in healthcare. NPJ digital medicine, 4(1), 140. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8455556/

[3] Panch, T., Szolovits, P., & Atun, R. (2018). Artificial intelligence, machine learning and health systems. Journal of global health, 8(2). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6199467/

[4] Verghese, A., Shah, N. H., & Harrington, R. A. (2018). What this computer needs is a physician: humanism and artificial intelligence. Jama, 319(1), 19-20. https://jamanetwork.com/journals/jama/article-abstract/2666717

[5] Ahmad, M. A., Eckert, C., & Teredesai, A. (2018, August). Interpretable machine learning in healthcare. In Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics (pp. 559-560). http://www.ieee-iib.org/2018/Aug/article1/iib_vol19no1_article1.pdf

[6] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(4), e1312. https://doi.org/10.1002/widm.1312

[7] Vellido, A. (2020). The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural computing and applications, 32(24), 18069-18083. https://link.springer.com/article/10.1007/s00521-019-04051-w

[8] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence, 1(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

[9] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence, 267, 1-38. https://doi.org/10.1016/j.artint.2018.07.007

[10] Lavecchia, A. (2019). Deep learning in drug discovery: Opportunities, challenges and future prospects. *Drug Discovery Today*, 24(10), 2017–2032. https://doi.org/10.1016/j.drudis.2019.07.006

[11] Seyhan, A. A. (2019). Lost in translation: The Valley of death across preclinical and clinical divide – identification of problems and overcoming obstacles. *Translational Medicine Communications*, 4(1). https://doi.org/10.1186/s41231-019-0050-7

[12] Weissler, E. H., Naumann, T., Andersson, *et a*l (2021). The role of machine learning in clinical research: transforming the future of evidence generation. *Trials*, 22(1), 1-15. https://doi.org/10.1186/s13063-021-05489-

x

[13] Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., & Ranganath, R. (2020). A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, 2020, 191. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7233077/

[14] Ustun, B., & Rudin, C. (2017, August). Optimized risk scores. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1125-1134). https://doi.org/10.1145/3097983.3098161

[15] Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, 47(5), 329-335. https://jme.bmj.com/content/47/5/329

[16] Durán, J. M., & Formanek, N. (2018). Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds and Machines*, 28, 645-666. https://link.springer.com/article/10.1007/s11023-018-9481-6

[17] Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., & Cilar, L. (2020). Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5), e1379. https://doi.org/10.1002/widm.1379

[18] Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3), 31-57. https://doi.org/10.1145/3236386.3241340

[19] Jia, X., Ren, L., & Cai, J. (2020). Clinical implementation of AI technologies will require interpretable AI models. Medical physics, 47(1), 1-4. https://doi.org/10.1002/mp.13891

[20] Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. In *Proceedings of the National Academy of Sciences*, 116(44), 22071–22080. https://doi.org/10.1073/pnas.1900654116

[21] Okay, F. Y., Yıldırım, M., & Özdemir, S. (2021, October). Interpretable machine learning: a case study of healthcare. In *2021 International Symposium on Networks, Computers and Communications (ISNCC)* (pp. 1-6). IEEE. https://ieeexplore.ieee.org/abstract/document/9615727

[22] Wu, N., et al (2019). Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE transactions on medical imaging*, 39(4), 1184-1194. https://ieeexplore.ieee.org/abstract/document/8861376

[23] Cohen, J.P., Viviano, J.D., Bertin, P., Morrison, P., Torabian, P., Guarrera, M., Lungren, M.P., Chaudhari, A., Brooks, R., Hashir, M. &amp; Bertrand, H.. (2022). TorchXRayVision: A library of chest X-ray datasets and models. *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning*, in Proceedings of Machine Learning Research, 172, 231-249. https://proceedings.mlr.press/v172/cohen22a.html