Do Explainable AI techniques effectively explain their rationale? A case study from the domain experts' perspective

Authors omitted for blind review

Abstract

AI systems are technologies impacting our lives. The systems learn from existing datasets that record human past decisions. Their performance is measured in terms of accuracy, precision, and recall for reproducing already known results. Understanding the system's rationale is crucial to check for bias and to accept such technology. Explainable AI (XAI) is the area devoted to open the AI black-box, designing guidelines to build explainable AI systems. Nevertheless, it is important to understand the users' needs for these explanations. This paper presents an investigation of the usefulness of XAI systems in the field of a cancer diagnosis from the domain experts' (oncologists)perspective. The main findings suggest domain experts (1) understood the outcomes of the XAI systems; (2) considered XAI outcomes as informative, rather than explanatory; (3) would like to go beyond the fixed presented perspective; and (4) missed the causal relation that would reveal the systems' rationale.

1. Introduction

Today, AI has become part of our daily lives. It is used to substitute human decision-making in several contexts, such as for credit limit checking [1], automatically interpreting x-ray to diagnose COVID-19 [2] or even to clean our houses with rumba [3]. The benefits are sensible, bringing productivity and assistance to our everyday experiences.

There are many real-world problems in which biased models can conceal inaccuracy, deficiency, and prejudice, contributing to ethical concerns and challenging AI methods' reliability. Some examples are prediction of crimes, credit fraud detection, clinical health assessment, loan appraisal, criminal justice risk score, self-driving autonomous cars, among others. The outcome from intelligent systems may be effective with high precision, but using the previous examples one can argue: *why did the systems deny John's payment on a web site?*, *why did the systems refuse Paul's credit request?*, *what is in the x-ray that tells Mary is sick?*. In these cases, understanding and interpreting the results of machine learning models is fundamental.

A large part of those methods are recognized as black-boxes capable of executing complex operations, but unable to explain or expose the causal model for the taken decision. In practical terms, a challenging situation is faced when stakeholders are put up with technological assistance and machine interference on critical tasks. As the learning algorithms influence their activities, there is a genuine need to trust and understand the mechanisms underlying the black-box models.

Hitherto, the benefits of using AI have surpassed our expectations. AI systems mostly learn from past human experiences registered in datasets. Human decisions are prone to errors, bias, and prejudices that individually may pass unnoticed, but when consolidated by machine learning, it becomes dangerous.

There are complex ML techniques, such as convolution neural networks, generative adversarial networks, and recurrent neural networks that produce precise results. Nevertheless, the results may uncover patterns of prejudice, errors, and unacceptable reasoning. Additionally, what is acceptable once may not be afterward. For instance, until the '60s being homosexual was a crime in most countries. Today, in most of the countries, it is just a matter of sexuality choice [4]. Even for not as sensitive issues as a sexual choice, we must stay vigilant against improper reasoning that leads to prejudice against minorities.

Systems' decision explanation is a desirable feature of any system [5]. In general, showing the business rule for the systems' features meets this need. However, a machine learning system does not have a clear and intelligible business rule to follow.

Computers learn from data that records human past decisions or from interacting with the world. Research efforts have been producing accurate and reliable systems with robust results. Machine learning models provide answers to assist in decision-making processes.

Still, most of the users involved have little understanding or knowledge about how these smart algorithms work internally. There is no proper comprehension of the rationale behind the learning models.

Dumping the systems' process in terms of nodes, weights, aggregation, and activation function may be useful for AI experts that need to follow the transformation from input to output. However, domain experts and end-users will not understand this data as an explanation. They need more, what kind of explanation do they actually need?

Explainable Artificial Intelligence (XAI) aims to reveal the hidden reasoning of intelligent agents. Intending to support the human comprehension, XAI techniques seek for more transparent, interpretative and explainable systems [6]. So the results of learning algorithms can disclose the causal explanation and the steps taken in machine outputs, predictions and recommendations. As a result, human experts, researches, and users can finally expect to understand how and why an algorithm made a certain decision.

This work evaluates the comprehensibility of XAI explanations from the domain experts' perspective, aiming to characterize the confidence and understanding of the results produced by explainable techniques. The central question is to analyze whether the unveiled rationale is acceptable by the professional specialist and, if not, provide the appropriate recommendations to improve the intelligent agents. To this end, it was developed a cancer diagnosis system, with high accuracy, using random forest from a public database. The experiment ran some cases and applied three XAI techniques to generate explanations. The results were presented to 12 oncologists. Semi-structured interviews were used to characterize the comprehensibility and transparency of the explanations generated by the three XAI techniques. A qualitative research method was applied and found interesting evidence that current XAI techniques are informative, but not explanatory. The final observations encourage a set of guidelines and suggestions for the development of XAI systems.

The remaining of this paper is organized as follows. Section 2 covers some required formal foundations regarding basic terminologies and highlights the distinct audiences for XAI applications. Section 3 outlines the methodology used for conducting this research. Subsequently, the study results are presented in Section 4. Section 5 presents some discussions and reinforces the relevant findings discovered in this paper. Finally, Section 6 puts forward some conclusive remarks and future work plans.

2. Theoretical Background

XAI attempts to unfold the reasons and internal logic of machine learning technologies, especially black-box algorithms such as neural networks, support vector machines, random forest, and others. Explainable models promote the dialogue between human-computer interaction (HCI) and artificial intelligence areas. Because to achieve a meaningful understanding of learning algorithms, it is precisely needed the apprehension and the interaction between human cognition and machine logic.

Key concepts concerning explainability and interpretability are explored in this section. The correlation and interrelationship of these terms prove not a monolithic idea, but a conflation of different theories [7]. The following discussion articulates practical issues of explanation, algorithmic fairness to prevent bias problems and ethical aspects of the methods.

2.1. XAI Concepts

The primary goal of *explainability* is the assignment of causal events allowing a practitioner to answer *why*-questions [8]. A semantic explanation is created to understand the behavior of AI systems. As humans prefer contrastive questions [9], it is relevant to note that *why*-question become more challenging. People do not ask why event Q happened, but rather they want to know why event P did not happen. This adds more sophisticated reasoning to XAI techniques because the process should consider counterfactual reasoning to focus not only on events that happened but also simulate occurrences that did not take place.

Some authors use the term *interpretability* as synonyms for *explainability* [6] and *intelligibility* [10]. The present article acknowledges the strong correlation between the mentioned concepts. However, there is a subtle distinction in regard to the goal addressed by each of the terms. Understanding these individual aspects is important to make XAI intention clear. While explanation refers to numerous ways of describing the rationale of a black-box model, *interpretability* refers to the quality a user can really apprehend the given explanation. Thus, the interpretation occurs after the explanation phase and depends on the human cognition to complete the understanding process.

The transparency of explainable models is analog to an X-ray, which enables examining the main stages of intelligent algorithms: design, processing, and learning. It should display latent interactions in the domain features (input layer) and describe the most relevant

ones for the obtained results. Learning from historical data may emphasize discrimination or favoritism acts hidden in the environment and assign discoveries to the algorithms' internal rules, violating social equity standards. Transparency intends to show whether the model performs as expected without bias issues or unfairness in the training data set.

Discrimination is related to improper or unequal conduct of people resulting in prejudice behavior against ethnic groups, gender and sexual diversity, religion, and other minorities. The research community must avoid and discourage learning models with unfair trends that harm a specific group of people [11, 12]. Social equity requires accountability (legal and moral) on the consequences of ethical problems. Interpretability and transparency can minimize such risks and become an asset for trust.

It is not a simple task. Some studies show different manners to deceive and add errors in machine learning algorithms. They use Evolutionary Algorithms to create "fooling examples" completely unrecognizable to humans [13] or produce small perturbations on the input layer [14] in such a way that Deep Neural Networks misclassifies the images. Another real-world case is the COMPAS risk tool, a computer program for assigning risk scores in the criminal justice system. Used in USA, it predicts the criminal defendant's likelihood of committing a future crime [15]. Data analysis indicated that the predictions were biased against African-American defendants [16] and not better than predictions made by people without criminal justice expertise [17].

In practice, those situations raise concerns about trusting black-boxes. As more people use machine learning algorithms, trust is a fundamental value for predictions and recommendations. XAI researchers claim the need for human understandable explanation to ensure technology acceptance and trust.

2.2. Different Players, Different Needs of Explanation

This paper focuses on the explanation needs from the AI system users' perspective. However, there are multiple stakeholders with different needs. As presented in Table 1, domain experts assess the AI systems' rationale to be able to trust the model and learn from it. Trust requires an understanding of and, eventually, an agreement with, the systems' process. The shift from explanation to understanding brings into account the explanation receiver psychological ability to explore the many possible interpretations of an explanation [18].

Most people are by the AI systems' outcome to some degree as in the case of an AI system checking a credit card purchase. The explanation must include evidence of judgement fairness. Similarly, regulatory agencies must also assess fairness on the AI system' judgment according to laws and norms, such as the European GDPR[19].

AI experts, data scientists, and programmers require a more mechanistic explanation showing all the steps involved in data transformation from input to output. We consider Hempel–Oppenheim mechanistic explanation model [20] either using deductive or inductive logic. Besides, they need an explanation that lets emerge problems and potential new functionalities. Managers and owners share the same needs as regulatory agencies and programmers, but in a more abstract level.

This paper investigates explanation from the perspective of domain experts. We are interested on the acceptability of existing XAI systems, suggestions and design guidelines based on the domain experts perspective.

2.3. Explainable AI Techniques

The XAI methods explored in this work are classified as post hoc and model-agnostic. Post-hoc applies explainable algorithms after the training phase and guarantees a separation of interpretability from the learning model. The model-agnostic approach is so called because it can be used either on black-box models (e.g., Deep Neural Networks, Random Forest) or intrinsically interpretable models (e.g., decision tree, linear regression). Also, agnostic models are not restricted to a single representation of the features, and they are free to present multiple types of explanations to the end-user [22].

Local Interpretable Model-agnostic Explanations (LIME) implements local surrogate models to explain individual instances [23]. This technique produces perturbations in the inputs and follows their effects on the model's prediction. The variation found in the output indicates a good estimation about the importance of some input features for a particular instance. Based on the new permuted samples, LIME applies an interpretable model to explain the predictions.

Formally, local surrogate models can be stated as follows :

$$\text{explanation}(x) = \arg\min_{g \in G} L(f, g, \Pi_x) + \Omega(g) \quad (1)$$

Let g be the explanation model, $\Omega(g)$ be a measure of complexity of g, G be the family of all possible explanations and Π_x the proximity measure

Table 1. Reasons for assessing AI systems' explanation (adapted from [21])

XAI audience	The need for explanation
Domain experts	Trust and agree with the AI model; obtain scientific knowledge
People affected by the systems' outcomes	Check for unfairness that may impact their lives
Regulatory agencies	Certify the system complies with laws and norms
Managers and system's owners	Verify legal compliance; look for new applications within their company
AI experts, data scientists and programmers	Verify performance and look for optimization and new functionalities opportunities

of the neighborhood samples around instance x. The explanation model g for the instance x is a minimization of loss L, which measures how close g is to the original prediction model f.

SHapley Additive exPlanations (SHAP) is a method that explains individual predictions and reveals the importance value of each feature [24, 25]. SHAP is based on the game theory. In this context, the *game* is the prediction task for an instance of interest; the *players* are the feature values; and the *gain* is the difference between the current prediction for this instance and the average prediction for all instances. The SHAP approach calculates the Shapley value [26] as the average contribution of a feature value to all feasible coalitions. It shows how to allocate the *gain* across the features in a fair manner.

Basically, the additive feature attribution method represents the Shapley value explanation as:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \tag{2}$$

where g s the explanation model, M is the number of input features, $z' \in \{0,1\}^M$ and $\phi_i \in \mathbb{R}$ is the feature attribution for a feature i.

The Permutation Feature Importance (PFI) method modifies the feature values and evaluates the effects in the prediction error of the model [27]. This technique measures the impact on the expected outcome, positive or negative, when the feature is permutated. If the model error increases, the feature importance is high, and the model relies on it. Otherwise, if the error does not change, the feature has low importance for the prediction model.

The PFI algorithm [28] considers a trained model f, a feature matrix $\boldsymbol{X}$ and a target vector $\boldsymbol{y}$ to estimate the original error $e^{orig} = L(\boldsymbol{y}, f(\boldsymbol{X}))$. For each feature $j = 1 \ldots p$, the algorithm generates $\boldsymbol{X}^{perm}$ by permutating j in $\boldsymbol{X}$ and measuring the new error $e^{perm} = L(\boldsymbol{y}, f(\boldsymbol{X}^{perm}))$. Then, the permutation feature important is calculated as $FI^j = e^{perm}/e^{orig}$.

3. Research Methodology

The objective of our research was to verify whether XAI systems accomplish the explanation needs of domain experts. There was no hypothesis to be tested, but a search for characterizing the domain experts' perspective, opinions, beliefs, and even emotions when interacting with such systems. We used in-depth interviews with qualitative analysis methods [29]. We interviewed 12 oncologists who interacted with three different XAI systems built upon an intelligent cancer diagnosis system. Participants were recruited from an open invitation basis. The flexibility of the method and the open-ended questions let emerge other issues related to AI technology acceptance. We used the Grounded Theory method for data analysis [30], with a particular emphasis on the coding procedure. This section describes the data collection instrument, the participants, the task, and the collected data.

3.1. Domain Area

Regarding the domain area, the objective was to use a critical field where the interpretability and explicability of the results are essential for the confidence in the learning models. We chose the medical area, more precisely the oncology scope that works with the prevention, diagnosis, and treatment of cancer.

Below, the criteria for choosing the medical field:

- Availability of public datasets;
- Importance of the area (oncology);
- Increasing use and relevance of AI in the medical domain;
- Availability of domain experts and people to observe and interview.

3.2. Learning Models and XAI Systems

Obtained from the open UCI Machine Learning repository, the dataset consists of indicators and risk factors to predict whether a woman will have cervical cancer. Features include demographics (e.g., age), lifestyle (e.g., smoker), and medical history (e.g., sexually transmitted diseases) taken from the University Hospital of Caracas in Venezuela.

The machine learning models used for classification were: Naive Bayes, Random Forest, Logistic

Regression, Support Vector Machine (SVM), Decision Tree, Multilayer Neural Network (MLNN), and KNN. Some interpretable models, such as decision tree, were included with the intention to compare their performance with the black-box models. The recordings were divided between 75% training and 25% testing data. Table 2 presents some performance metrics for the learning algorithms on cervical cancer dataset. The Random Forest ensemble method achieved the highest values of F1-score, accuracy, precision and recall. Therefore, random forest AI technique was the one selected to build the cancer diagnostic system to be presented to the domain experts.

Table 2. Performance metrics for learning algorithms

Model	F1-score	Accuracy	Precision	Recall
Naive Bayes	0.83	0.77	0.94	0.77
Logistic Regression	0.96	**0.95**	0.95	0.96
Decision Tree	0.95	0.94	0.95	0.95
Random Forest	**0.97**	**0.95**	**0.97**	**0.97**
SVM	0.94	0.94	0.94	0.94
MLNN	0.91	0.94	0.89	0.94
KNN	0.91	0.93	0.89	0.93

Afterward, model-agnostic XAI methods were chosen to be integrated with the cancer diagnostic system. SHAP, LIME, and PFI were selected for being largely used for delivering AI explanations in the technical literature. Figure 1 shows the output for the cancer diagnostic system example (feature title in Portuguese).

3.3. Interviewees Profile and Questionnaire

Once the XAI methods were applied on the learning model, this work is interested on the acceptability of the explanation by the domain experts. The goal is to review if the specialists received a proper clarification and insights about the systems' rationale.

The recruitment of participants aimed at health professionals with experience in the oncology field. As the cancer diagnosis requires a variety of other healthcare professionals to create a patient's overall treatment plan, this research interviewed a multidisciplinary team of physicians and pharmacists from different institutions.

A total of 12 domain experts (DE) responded to the open-ended questions until the saturation point is achieved, and no more interviews were necessary. To ensure anonymity, the volunteers' names were coded as DE01, DE02, DE03, DE04, DE05, DE06, DE07, DE08, DE09, DE10, DE11, and DE12. Table 3 details the interviewees' profile.

Table 3. Interviewees profile

Code	Gender	Institution	Area
DE01	M	INCA[a]	physician
DE02	F	Fiocruz[b]	pharmacist
DE03	F	Fiocruz	pharmacist
DE04	F	Fiocruz	pharmacist
DE05	F	SLucas[c]	physician
DE06	F	private clinic	physician
DE07	F	Fiocruz	pharmacist
DE08	M	Fiocruz	physician
DE09	F	Fiocruz	physician
DE10	M	Fiocruz	pharmacist
DE11	F	D'Or[d]	physician
DE12	M	D'Or	physician

[a] Brazilian National Cancer Institute.
[b] Osvaldo Cruz Foundation.
[c] São Lucas Hospital.
[d] D'Or Oncology Hospital.

The interviews followed a semi-structured format with open-ended questions. The script sought extensive and in-depth responses from the domain specialists. During the interviews, we tried to create an environment of spontaneity and informal conversations. Two groups of questions were created for the questionnaire. The first group is related to the understanding and interpretation of XAI results for the cervical cancer diagnostic system. In this context, the domain experts evaluated the interpretability and transparency of SHAP, LIME, and PFI explainable techniques. The second group of questions analyses the level of confidence on black-boxes predictions in the medical field.

Domain specialists were encouraged to talk about their experiences through the open-ended questions. Before starting the interview, the participants signed an informed consent form containing information about the research objectives and the use that can be made of the collected material. The interviews were individual and lasted an average of 50 minutes. The questionnaire was carried out showing the XAI results in real-time. All the answers were recorded for further and detailed analysis.

3.4. The coding procedure

Grounded Theory is the methodology underlying the interviews analysis phase. From the observation of relevant phenomena in the conversation (discourse, interaction, expression, etc.), researchers create the codes, concepts, and categories used to produce new theories [29]. In our case, the observed phenomena were captured in the interviews' transcriptions, which were input for the coding process. We started the interpretation and analysis of the interviews performing Open Coding, which is the process of fracturing the data

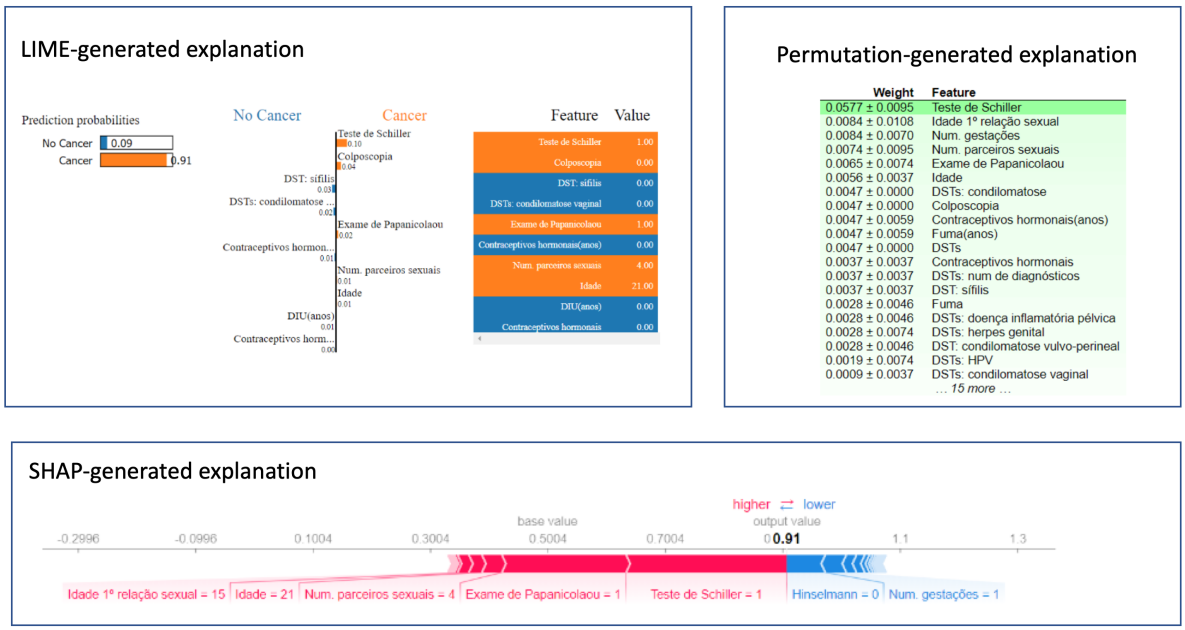

Figure 1. LIME, SHAP and PFI output for the cervical cancer diagnostic system example.

to identify categories, properties, and local dimensions [30]. Through open coding, the set of codes found represented potential understandings or impressions of explainable AI techniques. Work progressed in timeboxed intervals of four weeks, where the research team discussed the uncovered codes. This process was supported using QDA Miner Lite [1], and it led us to the identification of 32 significant codes.

Subsequently, the next step was the development of Axial Coding, which is the process of reassembling/regrouping codes by relating categories to subcategories to explain the phenomena of interest. Plenary meetings handled the discussion regarding the pertinence of open codes and their association with the axial codes. The research team repeated these activities until the analysis produced no new codes or re-organization of the axial codes. This process identified three main categories of explanation needs.

Finally, Selective Coding refines the process and search for the core category. We identified the core phenomena in the plenary meetings, and then researchers investigated for links explaining the overall theory grounded on data.

4. Findings

In this section, we report the outcomes of the qualitative analysis. Each subsection represents a category (an axial code) identified in the analysis. The three categories are presented in order of relative importance regarding the goal of this investigation. It should be noticed that the first category is the core category resulting from the selective coding process. For this reason, and due to space restrictions, only the first category will be comprehensively detailed in the following sections. Furthermore, for the same reason, we selected for presenting in this section the most important open codes defined during the analysis. The number of presented/identified open codes for each of the three categories in the order of presentation is as follows: 10/17 (**explainability of XAI techniques**), 6/11 (**trusting in AI**), and 2/4 (**recommendations for the medical domain**).

4.1. Explainability of XAI methods

The analysis core category is related to how information and explanation are perceived and understood by domain experts when using XAI methods. Surprisingly, contrary to what the term *explainable* would influence one's understanding of XAI techniques, some interviewees were concerned with the techniques' lack of explainability. In Table 4, the open codes associated with the explainability of XAI techniques are enumerated. It also provides a small description of each code and lists the interviewees' IDs in whose answers the codes originated. Although the interviewees' list provides a quantitative perspective

[1] https://provalisresearch.com/products/qualitative-data-analysis software

regarding the data, which can be convenient for some justifications given in the analysis, the reader should be aware that it is not the focus of qualitative analysis. The latter is concentrated on examining individual cases selecting those that are nonconforming or important cases for the analysis (i.e., "positive on the dependent variable") [31].

The set of open codes contains both perceptions regarding how the XAI techniques present the AI model results (e.g., **visual elements are easy to interpret** and **identification of minor influencing features**) and improvement suggestions (e.g., **improve visual elements**). These different perceptions were directly influenced by the first group of questions as described in Section 3.3, which inquired the participants not only about their understanding of the results but also the information missing in the results.

As mentioned at the beginning of this section, the most salient aspect of the analysis is related to the explainability of the XAI methods, which is primarily evidenced in the **XAI is not explanatory** code. One of the interviewees (DE09) indicated that "it (PFI) shown what parameters were used. However, the explanation regarding why is not here. It informs me of the parameters, but the relationship between the parameter and cancer I do not know. The (XAI) technique did not explain this." Another domain expert (DE10) complements the view from DE09 stating that "actually, listing the Schiller's Test, Papanicolaou exam etc, is not an explanation for me. This does not explain why. Hence, it does not explain; it justifies." These excerpts reveal that, in the participants' view, the information provided by the XAI techniques only indicate what factors (i.e., features) influenced the results but not why they influenced the outcomes.

Despite the issue regarding explainability, most domain experts acknowledged that the **visual elements are easy to interpret** and were able to perform the **identification of major/minor influencing features**. The excerpt from the DE12 interview depicts his impression regarding the interpretation of the visual elements with Lime: "the size... the size of the bar, right? The strength is greater for this factor. If the patient has done colposcopy, the chances for having cancer or another disease is much greater." Regarding the influencing features, when analyzing results generated with Lime, DE04 mentioned that "it is clear what descriptors (features) are related to having or not having cancer in this case along with this table showing data with colors." On the other hand, several participants argued that PFI **visual elements are difficult to interpret**. In the words of DE08: "I'm trying to understand the output... the weight and the variable... is there an ordering? You see, I can't understand if it has an ordering. Visually it is really unintelligible..."

Still concerned with the understanding of the results, several **experts did not agree with the results**. In this case, we analyzed whether this was related to the data used to build the model or the XAI technique. We found the responses complaining about this issue when using Shap or Lime (DE06 and DE09) were associated with problems in the model itself. For instance, DE09 answered, "I can understand the chart and the information, I just can't agree with the parameters (features)." On the other hand, users who did not agree with the PFI results were unaware of what a global technique represents, particularly concerning the fact that the results do not represent a single case (i.e., a diagnostic for a specific patient). For instance, DE11 asked, "I do not know if this was only with colposcopy, which is an important exam, but why it got a lower rank?"

Turning to the open codes related to the participants' improvement suggestions, we can separate them into two pairs of open codes. The first pair represents the participants' demand for more information to support and increase their confidence when making decisions. For some interviewees, the XAI methods lack detail regarding how the results were achieved. In other words, it is necessary to **add traceability information** to 'see' what is happening inside the models. As can be seen in the DE09 observations, this concern is directly related to the lack of explainability discussed earlier as it is explicit the need to understand the rationale behind the results: "It (Lime) should be able to provide access to a parameter (feature) so that it can be 'opened' and find out how it works. For instance, if I want to know how Shiller's test was used, then it (Lime) says how it was considered (in computing the results)... I would like to see the rationale behind it when I click on the name (of a feature). Why it was used in that way." Moreover, supplementing this aspect, several interviewees understood that the XAI techniques should **enumerate all features** so that "it is possible to see what is and what is not influencing the outcomes (DE08)."

The second pair contains improvements associated with the visualization of information in the XAI method. The domain experts suggested to **improve visual elements** and **add new types of data visualization**. As an example of these suggestions, DE03 indicated that charts could be used for data: "I think that a chart can help. When you put the characteristics (features) along with the values and add a bar chart with the values in it, it makes all the difference."

Table 4. Interviewees' frequency for the codes regarding explainability of XAI techniques

Code	Description	Interviewees with relation to the code			
		Shap[a]	Lime[a]	PFI[a]	# Total[b]
XAI is not explanatory	The domain expert indicated the XAI techniques' information does not represent an explanation, i.e., why and how the outcomes were obtained	09, 10	09, 10	05, 09, 10	3
Visual elements are difficult to interpret	The domain expert could not understand values, numbers, quantities or concepts contained in the results presented to them	-	-	04, 05, 06, 07, 08, 09, 12	7
Visual elements are easy to interpret	The domain expert could understand the values, numbers, quantities or concepts in the results presented to them	01, 03, 04, 05, 06, 07, 09, 11, 12	03, 04, 05, 07, 08, 09, 11, 12	02	10
Identification of major influencing features	The domain expert was able to identify the most important features influencing the results presented to them	01, 02, 03, 04, 05, 06, 07, 08, 09, 10, 11, 12	01, 02, 03, 04, 05, 06, 07, 08, 09, 10, 11, 12	01, 02, 03, 04, 06, 07, 09, 10, 11, 12	12
Identification of minor influencing features	The domain expert was able to identify the least important features influencing the results presented to them	01, 02, 05, 07, 09	02, 04, 06, 07, 09, 12	02	8
Expert does not agree with the result	The domain expert indicated the result is not consistent with the physician practice	06	06, 09	01, 05, 06, 07, 11	6
Enumerate all features	The domain expert indicated the XAI techniques should present all features used in the model	04, 08	05, 11, 12	-	5
Improve visual elements	The domain expert expressed his/her opinion regarding the visual aesthetics used in the information shown in XAI techniques (e.g., position, fonts, and colors)	-	04, 06	08	3
Add new types of data visualization	The domain expert shown interest in viewing information using different visualization formats (e.g., charts, diagrams, and mind maps)	-	03	03, 06, 08, 11, 12	5
Add traceability information	The domain expert requested information about the data used to generate the AI model along with the link to the model features	09, 10, 12	09, 10, 12	09, 10, 12	3

[a] Domain experts' IDs. We omitted 'DE' from the ID for abbreviation purposes. For instance, '07' should be read as 'DE07.'
[b] The total number of interviewees whose responses are related to the code.

4.2. Trusting in AI

This category was mainly a result of a specific question included in the questionnaire regarding trust in AI. The question was included precisely because of the relationship that can be established between trust and explainability, as discussed in Section 2.1. However, despite the authors' intention in exploring this matter, the participants also raised other concerns that were not necessarily related to explainability.

Three main themes emerged in the discussion regarding this theme during the interviews: the quality of the data and the AI models construction, the role of humans in the process of adopting AI and XAI techniques, and ethical concerns. Several participants mentioned the models should be **tested/validated (5)**[2] before applying in critical domains such as in medicine. Also, **data curation (4)** can be employed to assess and manage data quality.

As is the case of a considerable number of professions that can be affected by the introduction of AI, the domain experts manifested their concern with their professional position. They stated that AI **do not replace the health professional (4)** and called attention to the **necessity of the human factor (2)** in their area. **Ethical considerations (1)** was also regarded as important as wrong diagnostics can life-threatening. And **biased algorithms (3)** can also lead to erroneous decisions which can be harmful to patients as well.

[2]In this and the following section, the number of respondents is shown beside the open code. Both in bold.

4.3. Recommendations for medical domain

This last category is peripheral with respect to the more broad goal of this paper in understanding XAI techniques from the user perspective. Only a few open codes were classified under this category. An interesting aspect that is particularly important in the medical domain is the scientific rigor expected from the interventions used in the area. As a consequence, this worldview is promptly taken to other aspects of the medical area including the adopted technologies. Thus, several respondents mentioned the importance of **selecting, using, and testing diverse sets of data (3)** and **evidencing the model applicability and utility (5)**.

5. Discussion

Based on the interviews, we reached some interesting findings that provide evidence for building XAI systems' design guidelines. The literature in psychology, education, and philosophy make it clear that explanation, as a cognitive process, is closely related to, and perhaps even the same as causal reasoning,

since explanations often refer to causation or causal mechanisms, and causal analyses are believed to have explanatory value [32] [33] [34]. Explanations relate the event being explained to principles and invoke causal relations and mechanisms [35].

5.1. Informative, not explanatory

As clearly stated by the participants, the three XAI systems strive to show the elements from the input that actually took part in producing the final results. The presented information helped them to feel more "comfortable" with the AI system's result but within limitations.

Although the explanations seemed superficial, not actually opening the AI black-box, they could fill in some gaps with their own knowledge. Participants emphasized the informative nature of the XAI methods output, but an explanation would require at least a display of a clearer causal relation among the elements of the explanation. This corroborates the idea that an explanation must go beyond showing the data, but actually showing the causal relationships among the parameters [36, 37].

5.2. Lost in translation

The explanations provided by the XAI methods privilege some factors over others. This selection process may be misleading. For instance, a factor alone may be not as important as others, however within a conjunction of factors it might be decisive for providing a certain diagnosis. Consequently, the translation of what actually matter for a given scenario may be concealed.

This finding corroborates the idea that an explanation is a more complex artifact. It should allow users further investigation on the agent's reasoning letting clear the causal relationships among the elements of the explanation [38].

Choosing what to emphasize in the explanation may mislead the users, hiding existing or displaying nonexistent bias in that system that can be decisive for accepting or rejecting an AI system [39].

5.3. Explanation as an active artifact.

The XAI methods current outputs would comply neither with the mechanistic view that requires reveling the causal relationships within the entire transformations from the input to the output nor to the functional needs that would abstract and let the user to explore further perspectives so to adjust to their particular cognitive needs [35]. As observed by most participants, in different ways, they wished to explore the explanation to verify the points of accordance of disagreement between their reasoning and with the system's rationale and their own.

These findings corroborate the idea that an explanation should be an active artifact to explore the reasoning process rather than a statistic view. Explanations call for exploring the causal relations among the parameters from different perspectives to accommodate the users' cognitive skills and beliefs [18].

6. Conclusion

This paper sheds light on the acceptability of XAI methods from the domain experts' perspective. There are other studies comparing XAI methods [40, 41, 42], but focusing on general instead of the specific cognitive needs from the different explanation consumers or focusing on model visualization issues [22] [43], performance [37], and auditing [44].

Based on interviews with oncologists (domain experts) interacting with different XAI methods, we identified some key aspects that still must be addressed. We concluded XAI methods are informative rather than explanatory, limited to the features choice display and constrained to a specific view (rather than allowing explanation exploration from different perspectives).

We believe that research efforts concentrated most focus to address AI developers as XAI users. Our research focused on the domain expert's perspective within the context of understanding and accepting an XAI explanation. We will further explore the perspective of the domain expert in the context of using the AI system embedded in their activities. We will also broaden the research to encompass other XAI users, such as patients and regulatory agencies. We understand that our research is an important first step to actually build XAI design guidelines to accomplish the explanatory needs of domain experts.

References

[1] J. Bryson and A. Winfield, "Standardizing ethical design for artificial intelligence and autonomous systems," *Computer*, vol. 50, no. 5, pp. 116–119, 2017.

[2] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. R. Acharya, "Automated detection of covid-19 cases using deep neural networks with x-ray images," *Computers in Biology and Medicine*, 2020.

[3] Z. A. Neemeh, "Cultural affordances in ai perception," in *CogSci*, pp. 2441–2446, 2019.

[4] S. Chang, "The sex ratio and global sodomy law reform in the post-wwii era," *Journal of Population Economics*, pp. 1–30, 2020.

[5] V. Fortineau, T. Paviot, and S. Lamouri, "Automated business rules and requirements to enrich product-centric

information," *Computers in Industry*, vol. 104, pp. 22–33, 2019.

[6] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[7] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[8] L. H. Gilpin *et al.*, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th DSAA*, pp. 80–89, IEEE, 2018.

[9] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining explanations in ai," in *conference on fairness, accountability, and transparency*, pp. 279–288, 2019.

[10] Y. Lou *et al.*, "Accurate intelligible models with pairwise interactions," in *19th ACM SIGKDD*, pp. 623–631, 2013.

[11] D. Pedreshi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *14th ACM SIGKDD*, pp. 560–568, 2008.

[12] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2016.

[13] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *IEEE CVPR*, pp. 427–436, 2015.

[14] C. Szegedy *et al.*, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[15] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[16] J. Angwin *et al.*, "Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. 2016," *URL https://www. propublica. org/article/machine-bias-risk-assessments in criminal sentencing*, 2019.

[17] J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science advances*, vol. 4, no. 1, p. eaao5580, 2018.

[18] T. Chakraborti, S. Sreedharan, S. Grover, and S. Kambhampati, "Plan explanations as model reconciliation," in *2019 14th ACM/IEEE HRI*, pp. 258–266, IEEE, 2019.

[19] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.

[20] S. T. Mueller, R. R. Hoffman, W. Clancey, A. Emrey, and G. Klein, "Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable ai," *arXiv preprint arXiv:1902.01876*, 2019.

[21] A. Arrieta *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Info Fusion*, pp. 82–115, 2020.

[22] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *22nd ACM SIGKDD*, pp. 1135–1144, 2016.

[23] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," *arXiv preprint arXiv:1606.05386*, 2016.

[24] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, pp. 4765–4774, 2017.

[25] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," *arXiv preprint arXiv:1802.03888*, 2018.

[26] L. S. Shapley, "A value for n-person games," *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 307–317, 1953.

[27] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously.," *JMLR*, no. 177, pp. 1–81, 2019.

[28] C. Molnar, "A guide for making black box models explainable," *URL: https://christophm. github. io/interpretable-ml-book*, 2018.

[29] J. Corbin and A. Strauss, *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage publications, 2014.

[30] A. Strauss and J. Corbin, "Strauss, anselm, and juliet corbin. basics of qualitative research: Grounded theory procedures and techniques. newbury park, ca: Sage, 1990.," 1990.

[31] J. Mahoney, "Qualitative methodology and comparative politics," *Comparative political studies*, vol. 40, no. 2, pp. 122–144, 2007.

[32] J. Y. Halpern and J. Pearl, "Causes and explanations: A structural-model approach. part ii: Explanations," *BJPS*, vol. 56, no. 4, pp. 889–911, 2005.

[33] D. S. Krull and C. A. Anderson, "The process of explanation," *Current Directions in Psychological Science*, vol. 6, no. 1, pp. 1–5, 1997.

[34] J. Trout, "Scientific explanation and the sense of understanding," *Philosophy of Science*, vol. 69, no. 2, pp. 212–233, 2002.

[35] T. Lombrozo and D. Wilkenfeld, "Mechanistic versus functional understanding," *Varieties of understanding: New perspectives from philosophy, psychology, and theology*, p. 209, 2019.

[36] P. Langley, "Scientific discovery, causal explanation, and process model induction," *Mind & Society*, vol. 18, no. 1, pp. 43–56, 2019.

[37] D. Gunning *et al.*, "Xai—explainable artificial intelligence," *Science Robotics*, vol. 4, no. 37, 2019.

[38] J. Pearl and D. Mackenzie, *The book of why: the new science of cause and effect*. Basic Books, 2018.

[39] M. Nourani, S. Kabir, S. Mohseni, and E. D. Ragan, "The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems," in *AAAI HCOMP*, vol. 7, pp. 97–105, 2019.

[40] K. Hickey, L. Zhou, and J. Tao, "Dissecting moneyball: Improving classification model interpretability in baseball pitch prediction," in *53rd HICSS*, 2020.

[41] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[42] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, pp. 818–833, Springer, 2014.

[43] A. Chattopadhay *et al.*, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE WACV*, pp. 839–847, IEEE, 2018.

[44] M. Hall *et al.*, "A systematic method to understand requirements for explainable ai (xai) systems," in *IJCAI Workshop on XAI, Macau, China*, 2019.