Requirements Modeling and Elicitation for Explainable Artificial Intelligence Based on i*

Álvaro Navarro[1], Ana Lavalle[1], Alejandro Maté[1] and Juan Trujillo[1]

[1]*Lucentia Research Group, Dept. of Software and Computing Systems, University of Alicante, Carretera San Vicente del Raspeig s/n, San Vicente del Raspeig, 03690, Spain*

Abstract

EXplainable Artificial Intelligence (XAI) has risen as one of the prominent and complex topics to be addressed to foster the adoption of Artificial Intelligence (AI) solutions. With independence of the domain, explainability is key in gaining user trust and understanding the rationale behind the AI systems' outputs. Multiple XAI techniques exist, each with its own level of confidence and technical level required to understand the results among other characteristics. While an adequate modeling of XAI requirements would aid in selecting the correct XAI techniques for the relevant stakeholders, it has been mostly overlooked so far in the XAI field. We claim that a more formal representation of user requirements would help developers with the design of their systems. Therefore, we provide: a requirements language that aids in XAI requirements elicitation and modeling, covering the key XAI concepts and elements. Our requirements language is based on the i* framework, a technique that can facilitate the communication across diverse stakeholders. The benefits of our proposal include ensuring the fulfillment of multiple stakeholders' needs as well as the generation of combined explanations. Furthermore, our proposal also makes it easier for non-expert users to select the XAI techniques applicable to the context.

Keywords

Conceptual Modeling, EXplainable Artificial Intelligence, iStar, Requirements Engineering

1. Introduction

Nowadays, Artificial Intelligence (AI) systems and Machine Learning (ML)/Deep Learning (DL) models, which are subsets of these systems, are quickly spreading across society. Hence, citizens are closely linked with AI systems (or ML models), playing the role of AI partners [1]. In this context, AI systems generate outputs that significantly affect citizens' lives such as medical diagnostics [2]. However, many AI systems present black-box nature, which limits their applicability. Unfortunately, this lack of transparency leads many AI systems to failure.

In order to address these problems, eXplainable Artificial Intelligence (XAI) emerged, aiming to help users to understand how the different AI models make their decisions. Given the rise of AI, XAI has become a crucial research topic. On the one hand, it allows AI/ML developers (also known as AI/ML experts) to better develop AI models that are free from inconsistencies and errors. On the other hand, it allows end-users to (i) understand the output of the models,

iStar'24: The 17th International i Workshop, October 28, 2024, Pittsburgh, Pennsylvania, USA*

alvaro.nl@dlsi.ua.es (Á. Navarro); alavalle@dlsi.ua.es (A. Lavalle); amate@dlsi.ua.es (A. Maté); jtrujillo@dlsi.ua.es (J. Trujillo)

0000-0002-9730-2106 (Á. Navarro); 0000-0002-8399-4666 (A. Lavalle); 0000-0001-7770-3693 (A. Maté); 0000-0003-0139-6724 (J. Trujillo)

CEUR Workshop Proceedings (CEUR-WS.org)

(ii) trust the learned rationale or rules of the models, and (iii) have confidence on the decisions made by the models.

Motivated by this, we propose a requirements language to model and elicit XAI requirements, which is based on one of the most widespread Goal-Oriented Requirements Engineering (GORE) modeling techniques: the i* framework [3]. It is a flexible language defined at a high abstraction level that can have been often extended in order to adapt its behavior and capabilities to different fields. In our proposal case, it includes a Meta-Object Facility (MOF)-based metamodel and it will be done through (i) specializing the actors that can appear in XAI systems and (ii) adding new elements including concepts and relationships from the XAI domain.

To the best of our knowledge, this is the first proposal to extend i* including a metamodel to formalize XAI requirements and considers different stakeholders and their requirements in a systematic way. We believe that our paper shows that traditional user's requirements approaches and conceptual, logical and physical models with their transformations (from the software engineering field) need to be adapted to XAI systems, in order to improve the way requirements and XAI models outputs are considered.

The rest of the paper is structured as follows. The background is presented in Section 2. Section 3 presents the related work. Section 4 presents our proposal, the extension of i* applied to XAI. Section 5 presents the case study where our proposal has been applied. Finally, the conclusions and the future work are detailed in Section 6.

2. Background

In the XAI scenario, the main element is the explanation, *i.e.*, the output that explains how the AI systems' (or applications) decisions were made. At the same of writing, the literature establishes that the explanations are categorized depending on its methodology, usage and scope [4, 5], as will be detailed in this work.

Furthermore, there are XAI task that represent explanation in different ways: text [4, 6], visual [4, 7], by emphasizing the most relevant features [4, 8], analysing a sub-group of the whole objects [4, 9], by providing examples of input-output relationship [10, 4], and by defining a more simple and interpretable model -keeping a similar performance than the original model- [4, 11].

Moreover, the goal orientation is essential to work with XAI systems, which has been established in works that faces XAI from a requirements analysis [12]. In order to achieve this goal in the specific context (actors, data, model, etc.), a XAI task should be executed. Each of these concepts will be explained in detail in this paper.

In this context, the i* framework [3] is aligned with the XAI conceptual modeling and requirements elicitation needs. Thanks to the extension of the i* framework [3] for the XAI context, which allows to model the XAI scenario in a high abstraction level, we can (i) specialize the actors/stakeholders that are involved in the AI systems where XAI is included, and (ii) add new elements including concepts and relationships from the XAI domain to take into consideration each XAI concept as above-presented explanations.

3. Related work

Thanks to its flexibility, the i* framework allows us to apply Requirements Engineering (RE) techniques in different domains, taking into consideration social and technological dimensions. This is possible even in safety-critical areas, such as human-centered processes. For instance, the educational context was studied in [13], where the authors made use of the i* framework to present how stakeholders contribute to student success. Therefore, it has been applied in many areas, as follows.

In the Big Data and Business intelligence (BI) contexts, different works have applied it. For example, in [14], the authors proposed a trace metamodel for Data Warehouses (DW). This work is focused on the relationships between requirements and multidimensional (MD) conceptual models, that is possible through taking advantage of the i* goal-oriented nature.

Moreover, the i* framework has been applied in the AI field. In [15], the authors aimed to provide a language for specifying ML requirements. This work provides a model to help ML experts to develop AI systems. In that work, the authors linked the ML systems with XAI. However, contrary to their argument, the field of application does not decide whether or not explainability should be applied, but should be taken into account as a point to guide explanations. Hence, the explainability is always essential and should be included in the AI systems.

Finally, in the context of XAI, different works have faced the lack of information of the AI systems through analyzing different points. In [16], the authors defined different XAI metrics and goals and emphasized the relevance of defining an explainable interface, but they did not cover the objective of how to extract the XAI requirements, that is a crucial step to face each specific case study. In other work [4], the authors explained different XAI information about the current XAI field such as different XAI techniques. However, the authors did not model and extract XAI requirements. Since our proposal goes beyond the scope of [12], we have aligned and integrated their concepts in our metamodel. Therefore, we present an enriched metamodel for XAI requirements, identifying actors (stakeholders) and their roles, which is not achieved yet in the XAI and RE fields.

Summarizing, there are works focused on applying i* or similar techniques to model and extract requirements in different contexts. Moreover, other works have emphasized the need of modeling and achieving the specific XAI goals. Despite these works, we are still missing a requirements language that bridges the gap between conceptual modeling, RE and XAI. The most closest proposal is presented din [12], but this proposal is not complete and does not include a metamodel. In this context, there are no formalized and complete proposals to capture XAI requirements, but we integrate these existent non-complete proposals in our proposal to provide a more complete one that captures the XAI requirements. We cover this gap by providing an enriched extension of i* -including a metamodel- with the aim of modeling and eliciting the XAI requirements.

4. Requirements Engineering for XAI

In this section, we present our proposal, which is based on the i* Goal-Oriented modeling framework [3]. The aim of our requirements' language is to model and elicit the stakeholders' needs and ensure that the most adequate XAI techniques are applied to the specific AI context. Then, we will present our defined XAI requirements metamodel that aims to drive the XAI processes.

In order to define these elements, which are considered in the metamodel, different works focused on XAI have been analyzed (*e.g.*, [4, 12]). Hence, we have detected and integrated the main XAI dimensions to define an enriched metamodel that formalizes the XAI requirements. Hence, we expand the current literature works and present a formal proposal that includes (i) new elements that current works do not cover, and (ii) the relationships between all these elements. In the following, this metamodel, which extends i* and proposes these new concepts and elements, will be described in more detail.

In this context, the i* extended metamodel that includes these concepts and elements is presented in Fig. 1. In Fig. 1, core i* elements are represented in a blue color, whereas the new elements proposed tailored for XAI requirements are represented in a yellow color for the new concepts. Moreover, Fig. 1 also shows how we have included and extended the context (red square), problem (blue square), and solution (purple square) concepts presented in [12]. Summarizing, we have added different elements presented below.

First, we have defined a sub-goal for XAI. This Goal is called "XAIGoal" and contains one attribute: "typeGoal". This attribute can be assigned as "ModelPerformance" [17], "Coherence" [18, 19], "Domain" [6] or "Social" [20]. These different values have been selected to scope each dimension of the ML and XAI fields. All of them will be also included in the Qualities' and Indicators' blocks that will be exposed below. Moreover, the "XAIGoal" is subdivided in "ScopeGoal" (local or global) and "GranularityGoal" (shallow, medium or deep) [12].

Second, we have defined "XAITask" as a sub-task. This proposed sub-task contains an attribute that can be represented in different types of XAI tasks. More specifically, these types are "text" [4, 6], "visual" [4, 7], "features relevance" [4, 8], "local" [4, 9], "example" [10, 4] and "simplification" [4, 11]. Finally, the "XAITask" receives the "Domain", which is also incorporated in the Actors' block, as follows.

Third, the Actor's block has been extended. This extension is essential to present how the different types of stakeholders involved in the XAI process have been taken into consideration. On the one hand, there are two new kinds of actors called "EndUser", which represents the users that interact with the explainable interface, and "Recipient", which represents the users affected by theses interactions. On the other hand, a new role, called "XAIRole", has been added. Both types of stakeholders that inherit from the "Stakeholders" class that inherits from the "Actors" concept. Each one can play a role -or more than one- and can "collaborate" with other actors. In more detail: (i) the "MLExpert" has the attribute "MLKnowledge", which can be "Classification", "Regression" or "Clustering", as will be also presented in one of the indicators' elements added. This "MLExpert" also has knowledge about ML models, which interacts has the attribute "typeModel" that represents the type of model (opaque, black box; transparent, white box) and interacts with data (structured, semi-structured and unstructured) as input and output. Moreover, data inherits from the "resource" class and interacts with the "source" class, which also

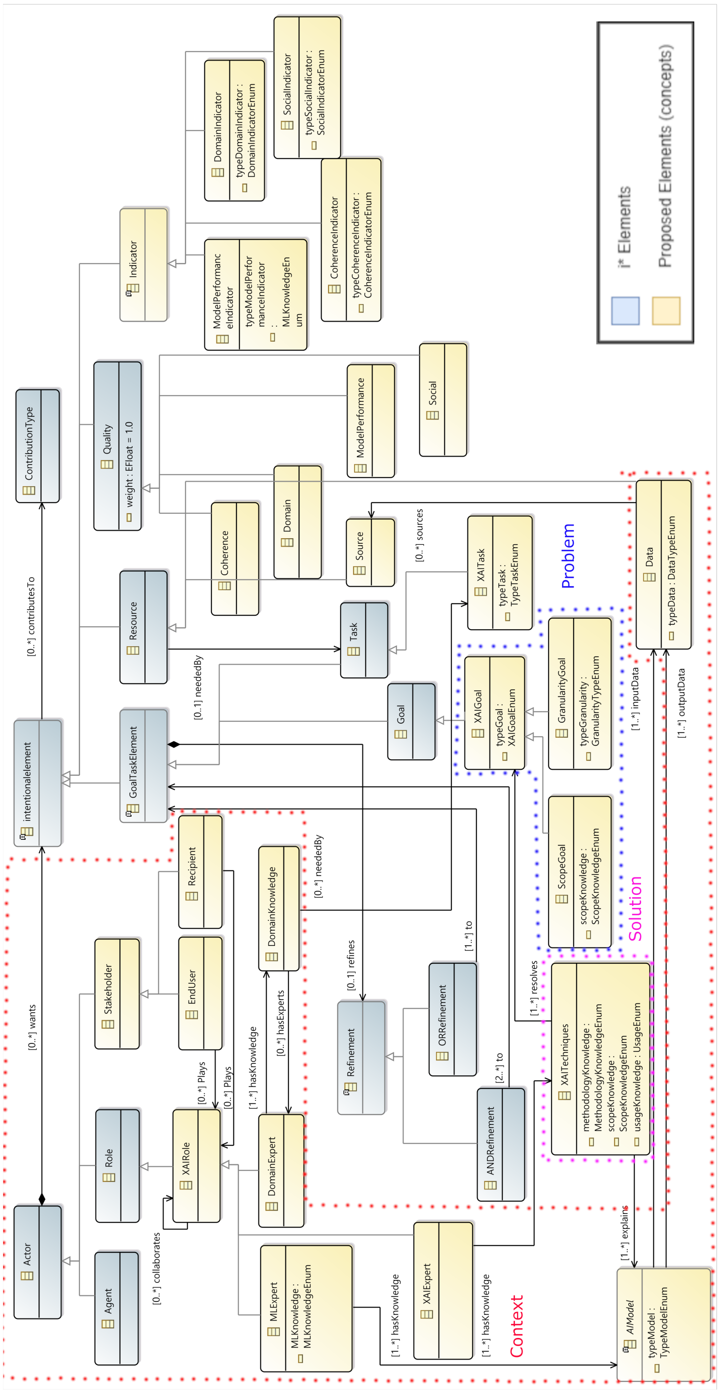


Figure 1: Proposed metamodel: i* for XAI.

inherits from “resource”; (ii) the “XAIExpert” role has knowledge about the “XAITechniques” (that explains the ML models to resolve the XAI goals) and their classification [5]. Specifically, there are three attributes represented in “methodologyKnowledge” (data: depending on the features in the input data instance; model layers: depending on the propagation input-output layers), “scopeKnowledge” (Local; Global), and “usageKnowledge” (Intrinsic, Post-Hoc); and (iii) the other role added is the “DomainExpert”, who is linked to the “DomainKnowledge” that is also included in the XAI task as a “Resource”.

Fourth, the Indicators’ block is added, that inherits from “intentionalelement”. The new concept “Indicator” is a super-class of four different concepts. First, “ModelPerformanceIndicator” that contains the attribute “typeModelPerformanceIndicator” (“classification” [21], “regression” [22] or “clustering” [23]). Second, “CoherenceIndicator” that contains the attribute “typeCoherenceIndicator” and can be “identity”, “separability” or “stability”. According to [18], these three axioms defines that (i) identical objects must have identical explanations, (ii) non-identical objects can not have identical explanations, and (iii) similar objects must have similar explanations, respectively. Third, “DomainIndicator” that contains the attribute “typeModelDomainIndicator”. This attribute can be “completeness”, “correctness” or “compactness”. According to [24], there are the three Cs of the interpretability. These three Cs represent (i) the accuracy, (ii) the reliability, and (iii) the feature dimensionality of the explanation, respectively. Fourth, “SocialIndicator” that contains the attribute “typeSocialIndicator” and can be “Specific metrics” or “Goals”. Both types should verify the social impact of the explanation in the “EndUser”.

Fifth, the Qualities’ block has been extended. There are four sub-classes that inherit from “Quality” and should be weighted. More specifically, “ModelPerformance” [17], “Coherence” [18, 19], “Domain” [6] and “Social” [20].

Consequently, by including the above-presented concepts and elements, we allow users to define their needs instead of focusing on technical details, which is crucial since these technical details are not relevant at this abstraction level.

5. Case study

To exemplify the applicability of our proposal, we have applied it proposal in areal project focused on AI-driven ADHD diagnosis and treatment: the BALLADEER project[1]. Therefore, we present Fig. 2. As it shows, our proposal helps to capture user requirements, in this case, for the neurologist. Thus, thanks to our proposal it is possible to easily detect which elements are important for specific users, which is essential to correctly design explanations that should meet the needs of these users.

Moreover, our proposal also allows to achieve the explainable interface, in this case, for the neurologist. In this way, this explainable interface shows how is possible to take advantage of the defined user requirements model (achieved thanks to the modeled and elicited XAI requirements). However, due to the different constraints presented in our case study (as the coherence indicators in Big Data scenarios [18]), there are limitations to implement the explainable interface. Therefore, the explainable interface, which will be presented below, is defined in an ideal context, without taking into account the different constraints.

[1] https://balladeer.lucentia.es/en/home-2/.

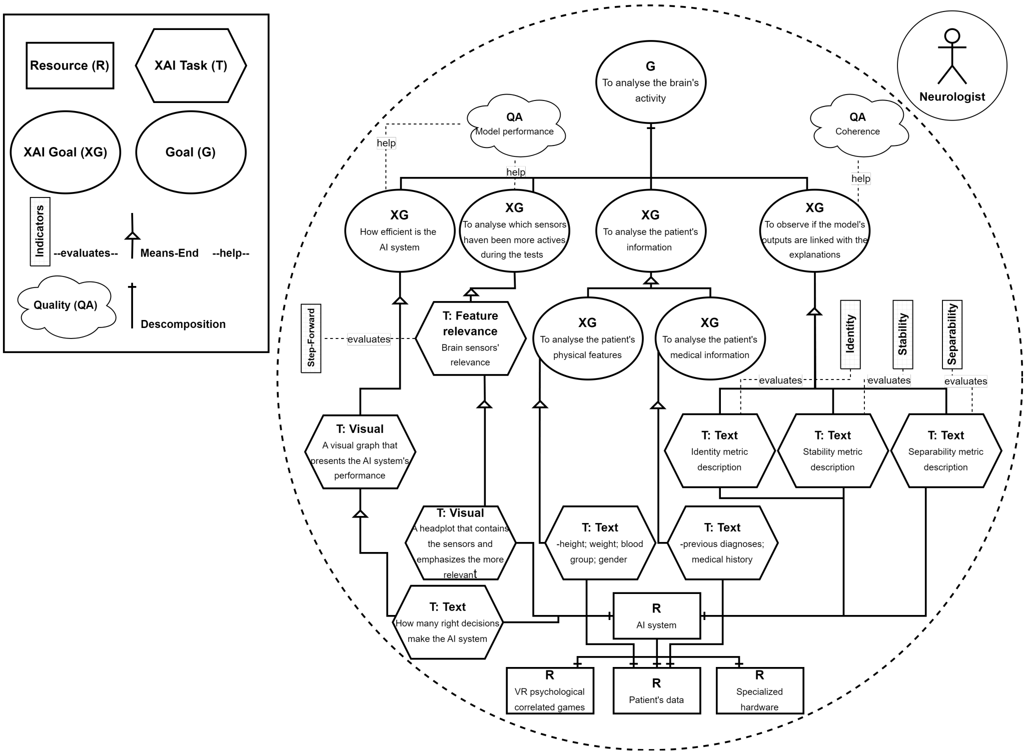

Resource (R)
XAI Task (T)
XAI Goal (XG)
Goal (G)
Indicators
--evaluates--
Means-End
--help--
Quality (QA)
Descomposition
Neurologist
G
To analyse the brain's activity
QA
Model performance
QA
Coherence
help
help
help
XG
How efficient is the AI system
XG
To analyse which sensors haven been more actives during the tests
XG
To analyse the patient's information
XG
To observe if the model's outputs are linked with the explanations
Step-Forward
evaluates
T: Feature relevance
Brain sensors' relevance
XG
To analyse the patient's physical features
XG
To analyse the patient's medical information
Identity
Stability
Separability
evaluates
evaluates
evaluates
T: Visual
A visual graph that presents the AI system's performance
T: Text
Identity metric description
T: Text
Stability metric description
T: Text
Separability metric description
T: Visual
A headplot that contains the sensors and emphasizes the more relevant
T: Text
-height; weight; blood group; gender
T: Text
-previous diagnoses; medical history
T: Text
How many right decisions make the AI system
R
AI system
R
VR psychological correlated games
R
Patient's data
R
Specialized hardware

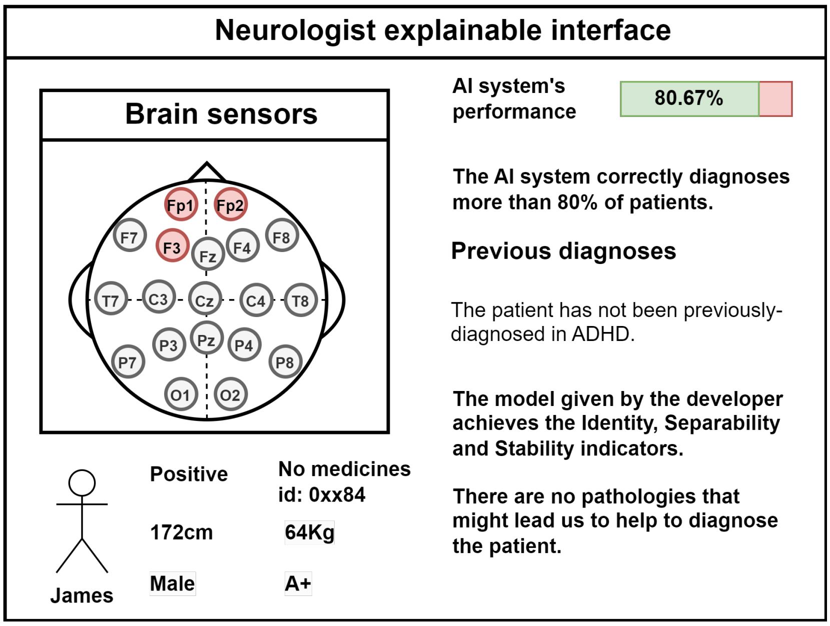
Figure 2: Captured requirements for the neurologist.

In this XAI and ADHD context, the neurologist aims to analyse the patient's brain behaviour. As Fig. 3 shows, this analysis is focused on the sensors that have captured the EEG signals. In this context, the neurologist analyses which sensors have been more relevant and the patient's information to understand the AI system diagnosis process and communicate it to the patient. Finally, this interface presents a headplot that includes 19 sensors. If the model had been trained with another database, this interface would be redesigned to represent the specific sensors and their brain locations.

Consequently, by taking into consideration the actor and their associated elements, it is possible to achieve a final AI system that correctly supports the different stakeholders involved to (i) predict ADHD cases and (ii) understand why each case has been classified as positive or negative case.

6. Conclusions and future work

In this paper, we have presented a requirements language to model and elicit XAI requirements based on i*. By aligning and including current works, our proposal expands them and presents a more complete and formal proposal that includes a metamodel. Hence, our proposal helps to capture the main elements and concepts in XAI, including stakeholders, relationships, and constraints involved in the XAI process. In this way, our approach lets identify and specify



Neurologist explainable interface

Brain sensors

James	Positive	No medicines id: 0xx84
	172cm	64Kg
	Male	A+

AI system's performance 80.67%

The AI system correctly diagnoses more than 80% of patients.

Previous diagnoses

The patient has not been previously-diagnosed in ADHD.

The model given by the developer achieves the Identity, Separability and Stability indicators.

There are no pathologies that might lead us to help to diagnose the patient.

Figure 3: Neurologist explainable interface.

stakeholders' needs, takes into account the ML models being used and the kind of information that end-user requires to interpret the results. Furthermore, our proposal also facilitates the selection of the XAI techniques appropriately applicable to the specific AI context.

Compared to current practice, our approach ensures that all relevant stakeholders and XAI dimensions are taken into account and facilitates the generation of combined explanations that fulfill end-users' information needs, thereby enabling confident use of AI systems.

In order to show the applicability of our proposal, we have applied it to an existing project on AI-driven Attention-Deficit/Hyperactivity Disorder (ADHD) diagnosis and treatment: the BALLADEER project. As a result of the application of our proposal, we have been able to define an explainable interface that provides the explainable interface for the neurologist.

Finally, in future work we will explore how to improve or make more systematic the derivation of requirements to interfaces by facilitating the traceability of requirements and the updating of interfaces if requirements are updated. In this way, we will be able to connect XAI requirements with their corresponding implementations, ultimately enabling an MDA approach that can facilitate quicker and less costly XAI implementations.

Acknowledgments

This work has been co-funded by (i) the AETHER-UA project (PID2020-112540 RB-C43) funded by the Spanish Ministry of Science and Innovation, and (ii) the BALLADEER (PROMETEO/ 2021/ 088) project funded by the Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital (Generalitat Valenciana).

References

[1] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.-Z. Yang, Xaiexplainable artificial intelligence, Science Robotics 4 (2019) eaay7120. doi:10.1126/scirobotics.aay7120.

[2] N. Amoroso, D. Pomarico, A. Fanizzi, V. Didonna, F. Giotta, D. La Forgia, A. Latorre, A. Monaco, E. Pantaleo, N. Petruzzellis, et al., A roadmap towards breast cancer therapies supported by explainable artificial intelligence, Applied Sciences 11 (2021) 4881. doi:https://doi.org/10.3390/app11114881.

[3] F. Dalpiaz, X. Franch, J. Horkoff, istar 2.0 language guide, arXiv preprint arXiv:1605.07767 (2016). doi:https://doi.org/10.48550/arXiv.1605.07767.

[4] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information fusion 58 (2020) 82–115. doi:https://doi.org/10.1016/j.inffus.2019.12.012.

[5] A. Das, P. Rad, Opportunities and challenges in explainable artificial intelligence (xai): A survey, arXiv preprint arXiv:2006.11371 (2020).

[6] D. Das, S. Chernova, Leveraging rationales to improve human task performance, in: Proceedings of the 25th International Conference on Intelligent User Interfaces, 2020, pp. 510–518. doi:https://doi.org/10.1145/3377325.3377512.

[7] P. Gupta, N. Puri, S. Verma, D. Kayastha, S. Deshmukh, B. Krishnamurthy, S. Singh, Explain your move: Understanding agent actions using specific and relevant feature attribution, International Conference on Learning Representations (ICLR) (2020). URL: https://par.nsf.gov/biblio/10166401.

[8] S. Prasanth Kadiyala, W. L. Woo, Flood prediction and analysis on the relevance of features using explainable artificial intelligence, AICSconf '21, Association for Computing Machinery, New York, NY, USA, 2022, p. 1–6. URL: https://doi.org/10.1145/3516529.3516530. doi:10.1145/3516529.3516530.

[9] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144. doi:https://doi.org/10.1145/2939672.2939778.

[10] E. M. Kenny, M. T. Keane, Explaining deep learning using examples: Optimal feature weighting methods for twin systems using post-hoc, explanation-by-example in xai, Knowledge-Based Systems 233 (2021) 107530. URL: https://www.sciencedirect.com/science/article/pii/S0950705121007929. doi:https://doi.org/10.1016/j.knosys.2021.107530.

[11] H. Khosravi, S. B. Shum, G. Chen, C. Conati, Y.-S. Tsai, J. Kay, S. Knight, R. Martinez-Maldonado, S. Sadiq, D. Gašević, Explainable artificial intelligence in education, Computers and Education: Artificial Intelligence 3 (2022) 100074. URL: https://www.sciencedirect.com/science/article/pii/S2666920X22000297. doi:https://doi.org/10.1016/j.caeai.2022.100074.

[12] T. Li, L. Han, Dealing with explainability requirements for machine learning systems, in: 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC), 2023, pp. 1203–1208. doi:10.1109/COMPSAC57700.2023.00182.

[13] N. Yang, T. Li, How stakeholders' data literacy contributes to student success in higher

education: a goal-oriented analysis, International Journal of Educational Technology in Higher Education (2020). doi:https://doi.org/10.1186/s41239-020-00220-3.

[14] A. Maté, J. Trujillo, A trace metamodel proposal based on the model driven architecture framework for the traceability of user requirements in data warehouses, Information Systems 37 (2012) 753–766. URL: https://www.sciencedirect.com/science/article/pii/S0306437912000701. doi:https://doi.org/10.1016/j.is.2012.05.003, special Issue: Advanced Information Systems Engineering (CAiSE'11).

[15] J. M. Barrera, A. Reina-Reina, A. Lavalle, A. Maté, J. Trujillo, An extension of istar for machine learning requirements by following the prise methodology, Computer Standards & Interfaces 88 (2024) 103806. URL: https://www.sciencedirect.com/science/article/pii/S0920548923000879. doi:https://doi.org/10.1016/j.csi.2023.103806.

[16] S. Mohseni, N. Zarei, E. D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable ai systems, ACM Trans. Interact. Intell. Syst. 11 (2021). URL: https://doi.org/10.1145/3387166. doi:10.1145/3387166.

[17] Y. Zhang, Y. Weng, J. Lund, Applications of explainable artificial intelligence in diagnosis and surgery, Diagnostics 12 (2022) 237. URL: https://www.mdpi.com/2075-4418/12/2/237. doi:10.3390/diagnostics12020237.

[18] M. Honegger, Shedding light on black box machine learning algorithms: Development of an axiomatic framework to assess the quality of methods that explain individual predictions, arXiv preprint arXiv:1808.05054 (2018). doi:https://doi.org/10.48550/arXiv.1808.05054.

[19] D. V. Carvalho, E. M. Pereira, J. S. Cardoso, Machine learning interpretability: A survey on methods and metrics, Electronics 8 (2019) 832. doi:https://doi.org/10.3390/electronics8080832.

[20] K. Höök, Steps to take before intelligent user interfaces become real, Interacting with computers 12 (2000) 409–426. doi:10.1016/S0953-5438(99)00006-5.

[21] P. C. Sen, M. Hajra, M. Ghosh, Supervised classification algorithms in machine learning: A survey and review, in: Emerging technology in modelling and graphics, Springer, 2020, pp. 99–111. doi:https://doi.org/10.1007/978-981-13-7403-6_11.

[22] D. Maulud, A. M. Abdulazeez, A review on linear regression comprehensive in machine learning, Journal of Applied Science and Technology Trends 1 (2020) 140–147.

[23] J. Amutha, S. Sharma, S. K. Sharma, Strategies based on various aspects of clustering in wireless sensor networks using classical, optimization and machine learning techniques: Review, taxonomy, research findings, challenges and future directions, Computer Science Review 40 (2021) 100376. doi:https://doi.org/10.1016/j.cosrev.2021.100376.

[24] W. Silva, K. Fernandes, M. J. Cardoso, J. S. Cardoso, Towards complementary explanations using deep neural networks, in: Understanding and Interpreting Machine Learning in Medical Image Computing Applications, Springer, 2018, pp. 133–140. doi:https://doi.org/10.1007/978-3-030-02628-8_15.