Exploring Commonalities in Explanation Frameworks: A Multi-Domain Survey Analysis

Eduard Barbu[1,*], Marharytha Domnich[1], Raul Vicente[1], Nikos Sakkas[2] and André Morim[3]

[1]Institute Of Computer Science, Tartu, Estonia
[2]Apintech Ltd, POLIS-21 Group, Limassol, Cyprus
[3]LTPlabs, Avenida da Senhora da Hora,459, Porto, Portugal

Abstract

This study presents insights gathered from surveys and discussions with specialists in three domains, aiming to find essential elements for an explanation framework that could be applied to these and possibly other use cases. The applications analyzed include a medical scenario (involving predictive ML), a retail use case (involving prescriptive ML), and an energy use case (also involving predictive ML). We interviewed professionals from each sector, transcribing their conversations for further analysis. Additionally, experts and non-experts in these fields filled out questionnaires designed to probe various dimensions of explanatory methods. The findings indicate a universal preference for sacrificing a degree of accuracy in favor of greater explainability. Additionally, we highlight the significance of feature importance and counterfactual explanations as critical components of such a framework. Our questionnaires are publicly available to facilitate the dissemination of knowledge in the field of XAI.

Keywords

machine learning, expert surveys, explainability framework

1. Introduction and Related Work

This paper explores the role of AI in data-driven decision-making across sectors like healthcare, retail, and energy, highlighting the challenges of ML models' complexity and opacity. It focuses on improving explanation understandability and trust through a study involving expert and layman feedback on different explanation types. Although the study focuses on developing a genetic programming (GP) tool to aid decision-making in these fields, the findings are relevant for any machine learning algorithm. This strategy enhances user trust and transparency across various ML models, providing applicable insights for AI applications.

Research in explainable AI (XAI) aligns AI system explanations with user expectations and needs. Key studies, such as [1], highlight identifying crucial stakeholders in AI explainability and the development of a framework to meet these needs. Tools like the System Causability Scale

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

*Corresponding author.
eduard.barbu@ut.ee (E. Barbu); marharyta.domnich@ut.ee (M. Domnich); raulvicente@gmail.com (R. Vicente); sakkas@apintech.com (N. Sakkas); andre.morim@ltplabs.com (A. Morim)
0000-0002-3664-5367 (E. Barbu); 0000-0001-5414-6089 (M. Domnich); 0000-0002-2497-0007 (R. Vicente); 0000-0003-4724-1322 (A. Morim)

[2] and the System Usability Scale [3] have been introduced to assess ML explanation interfaces and their effectiveness. Furthermore, a novel questionnaire leveraging psychometrics [4] aims to reliably evaluate XAI method explanations, addressing explainability's complex nature. This body of work underpins our effort to craft AI tools that meet the diverse requirements of professionals in fields such as medicine, retail, and energy, proposing a cross-disciplinary approach to enhance user satisfaction and trust in AI applications. In their literature review, the authors in [5] define five primary goals for AI system interactions with end users: understandability, trustworthiness, transparency, controllability, and fairness. They recommend designing XAI systems to achieve these objectives and suggest guidelines for creating explanations focusing on crucial system components. Additionally, they highlight the necessity for compromises in AI explanations, underlining the absence of a one-size-fits-all solution.

The paper is organized as follows: we begin with an overview of related work. This is followed by introducing the three distinct use cases and their unique characteristics. In Section 3, we elaborate on the methodology employed in conducting the surveys. The paper concludes with a discussion of our findings and presents conclusions, including recommendations for developing a GP tool to support practitioners across three use cases. The developed questionnaires are publicly available to facilitate the dissemination of knowledge in the field of XAI.

2. The use cases

Medical Scenario The medical scenario explores GP models for paraganglioma and diabetes, aiming to predict the tumor's progression and diabetes presence. The model for paraganglioma seeks to guide physicians on treatment timing, enhancing shared decision-making, optimizing treatments, and reducing unnecessary interventions without substituting clinical judgment. For diabetes, the model uses a well-known dataset [6] to predict if a patient has or does not have diabetes.

Retail use case Grocery stores use Dynamic Timeslot Pricing to balance customer satisfaction with efficiency in home delivery. They offer flexible delivery times while keeping costs low. This AI-based approach sets fair and clear prices by looking at customer data and delivery logistics to estimate how much customers are willing to pay and the cost to serve. An algorithm then matches customer preferences with delivery efficiency to find the best times and prices.

The method, which sets slot prices using a specific formula (Prescriptive Model), depends on two support models—the Willingness to Pay (WTP) and Cost to Serve (CTS) models.

Energy use case To recommend savings, the energy use case predicts household energy consumption by analyzing weather, historical usage, building dynamics, pricing, and indoor temperatures. It aims to offer users clear explanations to support informed decisions and to integrate these insights into business strategies for improved energy efficiency. Key considerations include weather conditions, past consumption patterns, building characteristics, pricing strategies for managing demand, and indoor temperature monitoring for energy conservation. The challenge is making these forecasts understandable and actionable, facilitating efficient energy use and decision-making in practical settings.

3. Survey methods

This section outlines the survey methodologies applied to the three investigated use cases. Our approach incorporated two methods: conducting interviews with domain experts and distributing questionnaires to practitioners who may not have expert knowledge.

Details of the surveyed experts are available at this link: Interviewed Experts Document. Links to the questionnaires for each use case can be found in the following subsections. Three medical doctors completed the medical use case questionnaires, while the retail questionnaires were filled out by the interviewed expert and six additional respondents. For the energy case, six respondents completed the questionnaires, four of whom were the experts interviewed.

3.1. Survey methods for the Medical Scenario

The questionnaire, which focused on diabetes risk estimation and was developed for the medical scenario, aimed to explore the type of AI model explanations doctors need. Key areas explored included the trade-off between accuracy and explainability, various presentation formats (such as symbolic regression graphs, genetic programming protocols, SHAP feature importance graphs, coefficients tables, and textual explanations), and their impact on understandability and decision-making effectiveness. Doctors were asked to rate each format's interpretability and effectiveness on a 1 to 5 scale.

Additionally, an interview focusing on the paraganglioma case collected insights on tumor identification, statistical prediction models, genetic factors, training protocols for new doctors, expectations from AI tools in managing paraganglioma, and the specific explanations needed for comprehending this condition. The questionnaire and interview outcomes are intended to guide the development of AI tools that effectively meet doctors' informational needs and preferences.

The questionnaire for the medical scenario can be explored here: Diabetes Questionnaire

3.2. Survey methods for the Retail Use Case

The retail use case questionnaire was designed to delve into several key areas. First, they explored price breakthroughs to gauge the significance of location and demand and how clear the explanations were to customers. Next, the questionnaire sought to identify which types of explanations customers preferred and how well they understood them. Lastly, there was a focus on summarization assessment to evaluate the need for summaries in conjunction with detailed pricing information. This part aimed to assess how these summaries affected clarity and influenced decision-making. Participants rated explanations on interpretability and effectiveness from 1 (least) to 5 (highest), aiming to understand the extent to which explanations helped in decision-making and their clarity to customers. For this use case, two questionnaires have been devised for two categories of users.

1. **Decision-makers** Seek a comprehensive understanding of feature contributions to model predictions for system optimization. With their expert background, they prefer detailed, technical explanations to build trust and validate the model's use based on its accuracy. Decision-Makers Questionnaire

2. **Customers** Favor straightforward, accessible explanations that still convey essential information, aiding in understanding the rationale behind received offers without overwhelming technical detail.Customers Questionnaire

The interview, which was recorded as a video file, explored issues such as finding a balance between accuracy and explainability in e-commerce models, the incorporation of graphs and mathematical formulas into explanations, understanding customer behavior through the dynamic relationship between slot availability and pricing, and designing a dynamic dashboard to manage the interaction between operational efficiency and customer behavior effectively.

3.3. Survey methods for the Energy Use Case

The questionnaire targets operational managers and customers, aiming to identify their preferred formats (tables, charts, interactive graphics, text) and types of explanations (causal, contrastive, counterfactual) for model predictions. Operational managers, the primary audience, must provide detailed feedback based on their expertise. They will focus on how model features affect predictions and optimization opportunities to enhance their trust and model endorsement through accurate and complex explanations. In contrast, customers likely prefer simpler, straightforward explanations that clarify the rationale behind offers. The energy questionnaire delves into key areas like the accuracy-explainability trade-off, the value of explanations in forecasting, the role of what-if scenarios in understanding model outcomes, and the specific needs of facility managers for detailed explanations and visualization tools such as SHAP graphs, highlighting preferences for explanation frequency and detail level.

All interviewed experts and five additional energy experts have completed the questionnaire. Energy Questionnaire

The interviews explored the energy problem from various angles, each tailored to the interviewee's expertise. Discussions ranged from addressing market challenges in energy solutions and the importance of clear explanations for end-users to exploring energy consumption disaggregation and the role of genetic programming in enhancing analysis. Insights were also shared on leveraging machine learning for water consumption monitoring to optimize resource management and identify inefficiencies. Additionally, the design and usability of user interfaces for energy management systems were examined, emphasizing the need for intuitive and engaging interfaces to manage energy consumption better.

4. Results

4.1. Medical scenario

Figure 1 summarizes key findings from the diabetes questionnaire.

Doctors prefer AI explanations that balance a slight decrease in accuracy for better clarity, find complex graphs challenging, and favor clear, intuitive details like protocols and SHAP graphs. Simplification and clarity were highlighted as essential for effectively conveying model logic, with counterfactual explanations being particularly valued for their potential to improve patient understanding and therapy compliance.

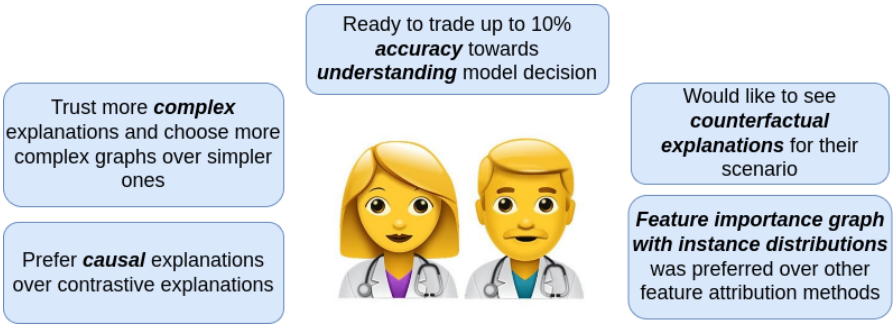

Figure 1: Insights into doctors' preferences for medical scenario derived from the questionnaire.

Feature importance graphs were most favored, followed by textual explanations and rule-based protocols. Graphs and coefficient tables were least preferred due to concerns about understandability.

Interview insights highlight the novelty of our paraganglioma models due to a lack of benchmarks to measure the accuracy of our models, the critical role of genetic data in personalized medicine, and the need for tools to monitor tumor growth. The value doctors place on model predictions for patient communication emphasizes the importance of accurate, explainable models to foster trust and informed decisions. Initial tests on GP models for paraganglioma are documented in [7], providing detailed outcomes.

4.2. Retail use case

The decision-makers seek explanations across various dimensions: customer behavior, transportation costs, and strategies for maximizing profits. The questionnaires findings are summarized in the figure 2

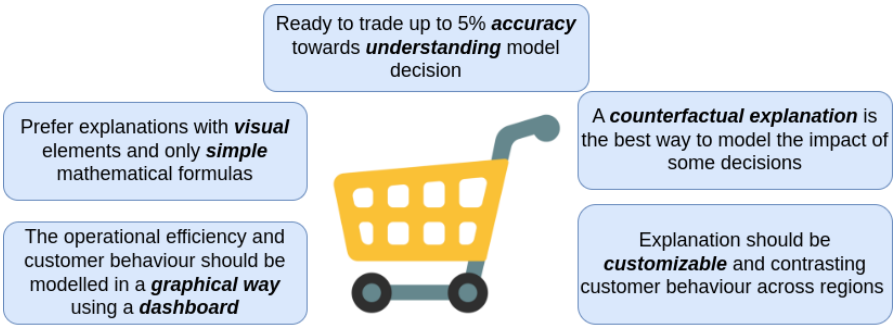

Figure 2: Insights into online retail decision-makers preferences derived from the questionnaire.

In feedback from decision-makers on AI system explanations, there's an openness to sacrificing a portion of model performance for enhanced explainability, with preferences for detailed yet intuitive insights into model workings. This encompasses a broad interest in customer

behavior, cost analysis, and profit strategies, highlighting a desire for interactive tools and visualizations that facilitate deeper understanding and strategic adjustments. There's a notable emphasis on practical application, with decision-makers valuing features like counterfactual explanations and the ability to interpret and act upon complex information, all aimed at optimizing operational efficiency and customer engagement.

The interview highlighted a preference for explainability over accuracy, with caution advised due to limited machine learning expertise. Simple visual explanations and mathematical formulas are preferred to avoid complexity. Graphical dashboards are recommended for assessing operational efficiency and customer behavior, enhancing interpretability and interaction. Counterfactual explanations are valued for demonstrating the impact of decisions such as new scheduling slots. Developing models that identify customer characteristics and behaviors by region is essential for deeper business insights.

4.3. Energy use case

The insights from operational and facility managers are summarized in figure 3.

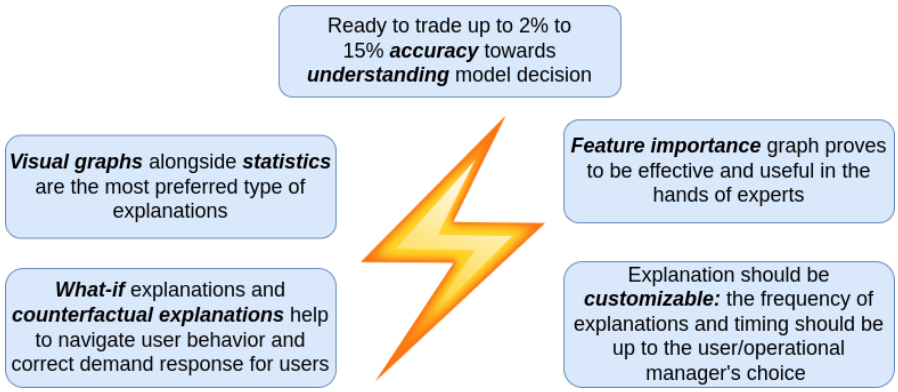

Figure 3: The insights from the energy questionnaire from operational and facility managers

Operational managers favor a balance between accuracy and transparency, adjusting the trade-off based on the audience. They prefer visual and simple mathematical explanations to suit various stakeholder technical levels. Graphical dashboards are effective for insights into efficiency and customer behavior, with counterfactual explanations providing useful scenario analysis. Strategic analyses, such as regional behavior modeling and what-if scenarios, highlight the value of feature importance graphs and counterfactuals in delivering clear, actionable insights for decision-making and management.

Insights from the interviews demonstrate a preference for explanatory forecasting models over basic ones, with methods applicable across sectors like gas and energy. Ease of use and interactive elements are advised for the graphical interface, alongside a smartphone component for energy applications to enable notifications. For detailed analyses of GP models in energy, see [8] and [9].

4.4. General guidelines

The table 1 summarizes the overarching guidelines derived from the survey findings.

Table 1
Guidelines and Insights from User Studies on Explanatory Tool's Architecture

Domain	Insight	Recommendation
All	Preference for explainability over perfect accuracy, feature importance graphs as effective communication tools, and value of counterfactual explanations.	Balance explainability and accuracy, utilize feature importance graphs, and supplement counterfactuals for comprehensive understanding.

Drawing from these insights, the design of the explanatory tool should incorporate two essential modules: a Counterfactual Module, which calculates the minimal changes required to shift the model's decision towards a desired outcome, thereby enabling "What-if" scenarios based on user queries, and a Global Importance Module, which provides visualization of the significant feature contributions to the model's predictions, in line with findings from the user studies. Both modules should be integrated within the tool, ensuring that the inputs, outputs, and connections between modules are well-defined.

5. Conclusions

This study identifies foundational components for an XAI framework intended for various applications through comprehensive questionnaires and interviews with domain experts in three distinct use cases. The envisioned XAI tool incorporates a Counterfactual Module to facilitate "What-if" scenarios, allowing users to see how minimal changes could lead to desired outcomes. Additionally, a Global Importance Module is designed to visually represent the most influential features in model predictions, resonating with the XAI literature emphasizing the critical role of feature importance and counterfactual explanations. While aiming for shared applicability, the framework also acknowledges the unique requirements of each specific case, although the detailed exploration of these unique case aspects was beyond this paper's scope. This approach informs the ongoing development of the AI tool, leveraging insights gathered from user studies to ensure the tool's effectiveness across different domains. Our tool is now prepared for evaluation by experts across the three fields. We will integrate their feedback into an updated version of the tool. For future research, the interest in online retail and energy sectors for customizable and user-specific explanations points towards a growing trend. This trend leans towards integrating NLP interactivity into explanations, an area we are beginning to explore.

Acknowledgments

This research was conducted under the Transparent, Reliable, and Unbiased Smart Tool for AI (Trust-AI) project, with Grant Agreement ID: 952060, funded by the EU Commission.

References

[1] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, K. Baum, What do we want from explainable artificial intelligence (xai)? – a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research, Artificial Intelligence 296 (2021) 103473. URL: https://www.sciencedirect.com/science/article/pii/S0004370221000242. doi:https://doi.org/10.1016/j.artint.2021.103473.

[2] A. Holzinger, A. M. Carrington, H. Müller, Measuring the quality of explanations: The system causability scale (SCS). comparing human and machine explanations, CoRR abs/1912.09024 (2019). URL: http://arxiv.org/abs/1912.09024. arXiv:1912.09024.

[3] M. Dragoni, I. Donadello, C. Eccher, Explainable ai meets persuasiveness: Translating reasoning results into behavioral change advice, Artificial Intelligence in Medicine 105 (2020) 101840. URL: https://www.sciencedirect.com/science/article/pii/S0933365719310140. doi:https://doi.org/10.1016/j.artmed.2020.101840.

[4] G. Vilone, L. Longo, Development of a human-centred psychometric test for the evaluation of explanations produced by xai methods, in: L. Longo (Ed.), Explainable Artificial Intelligence, Springer Nature Switzerland, Cham, 2023, pp. 205–232.

[5] S. Laato, M. Tiainen, A. Najmul Islam, M. Mäntymäki, How to explain ai systems to end users: a systematic literature review and research agenda, INTERNET RESEARCH 32 (2022) 1–31. doi:10.1108/INTR-08-2021-0600, funding Information: The initial literature search upon which this article develops was done for the following Master's thesis published at the University of Turku: Tiainen, M., (2021), To whom to explain and what?: Systematic literature review on empirical studies on Explainable Artificial Intelligence (XAI), available at: https://www.utupub.fi/handle/10024/151554, accessed April 2, 2022. Publisher Copyright: © 2021, Samuli Laato, Miika Tiainen, A.K.M. Najmul Islam and Matti Mäntymäki.

[6] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, R. S. Johannes, Using the adap learning algorithm to forecast the onset of diabetes mellitus, in: Proceedings of the Annual Symposium on Computer Application in Medical Care, 1988, pp. 261–265.

[7] E. M. C. Sijben, J. C. Jansen, P. A. N. Bosman, T. Alderliesten, Function class learning with genetic programming: Towards explainable meta learning for tumor growth functionals, 2024. arXiv:2402.12510.

[8] N. Sakkas, S. Yfanti, P. Shah, N. Sakkas, C. Chaniotakis, C. Daskalakis, E. Barbu, M. Domnich, Explainable approaches for forecasting building electricity consumption, Energies 16 (2023). URL: https://www.mdpi.com/1996-1073/16/20/7210. doi:10.3390/en16207210.

[9] N. Sakkas, S. Yfanti, C. Daskalakis, E. Barbu, M. Domnich, Interpretable forecasting of energy demand in the residential sector, Energies 14 (2021). URL: https://www.mdpi.com/1996-1073/14/20/6568. doi:10.3390/en14206568.