UNIVERSITY OF PAVIA – IUSS SCHOOL FOR ADVANCED STUDIES PAVIA

Department of Brain and Behavioral Sciences (DBBS)
MSc in Psychology, Neuroscience and Human Sciences

A mixed-methods investigation of clinician attitudes towards explainable AI in medical decision making

Supervisor:
Prof. Dr. Enea Parimbelli

Co-supervisor:
M.Sc. Laura Bergomi

Thesis written by:
Marta Anna Orłowska

Academic year 2023/2024

1. Index

Abstract

This study explores human-XAI (Explainable Artificial Intelligence) collaboration in the medical setting, focusing on clinicians' perceptions and preferences. Ten clinicians from I.R.C.C.S. Policlinico San Matteo Foundation of Pavia, Italy, participated in the survey, and two of them took part in the think-aloud session. The aim of the study is to assess and compare clinicians' perceptions of three tools: an explainable-by-design Bayesian network and two local XAI methods – Shapley values (Shap) and Araucana tree. The explanations were designed as the extension of the ALFABETO project, classifying COVID-19 patients for either discharge or hospitalization. Perceptions were assessed on usability dimensions: self-reported helpfulness, comprehensibility, and cognitive load. Sentiment analysis was also used to gauge emotional tone.

Results show clinicians generally trusted XAI explanations, with high compliance rates of 86%, though only 50% of predicted cases showed correct classification, indicating potential over-reliance. Compliance correlated with experience and survey completion time. Shap was perceived as the most comprehensible, helpful, and requiring the least cognitive effort due to its additive nature. Araucana required higher cognitive load and had slightly lower scores mirroring its higher complexity. The Bayesian network was neither comprehensive nor helpful, requiring too much cognitive effort. Sentiment analysis mirrored survey results, but more data is needed for conclusive findings.

Significant differences in tool preferences were found between ER (Emergency room) and ID (Infectious diseases) departments, with ID clinicians preferring Shap and ER clinicians preferring both Shap and Araucana. The study highlights the need for theoretical and empirical studies run together, by fitting results into a four-dimensional explainability framework. Overall, fine-tuning cognitive load and usability based on specific user needs makes Shap and Araucana strong candidates for effective human-XAI collaboration in healthcare.

Key words: Explainable AI (XAI), Human-XAI collaboration, Medical decision-making, Usability, Shapley values, Araucana Tree, Bayesian network.

Introduction

The integration of Artificial Intelligence (AI) in healthcare can revolutionize medical decision-making by providing advanced tools for diagnosis, treatment planning and lowering costs (Hudda et al., 2024). AI can analyse vast amounts of medical data quickly and accurately, identifying patterns and insights that might be missed by human practitioners. This can lead to more accurate decisions and treatments resulting in improved patient outcomes.

Artificial Intelligence is a branch of computer science that focuses on creating systems that can perform tasks traditionally thought of as requiring human intelligence. These tasks include learning from data, pattern recognition, making decisions and understanding natural language. The key of AI algorithms is their ability to learn from experience and adapt to new inputs. In consequence they are able to perform many human-like tasks previously thought of as impossible to achieve for a machine.

AI applications in healthcare encompass medical imaging analysis, predictive analytics, natural language processing, personalized treatment plans and the automation of administrative tasks (Hudda et al., 2024; Talati, 2023). AI is particularly beneficial in fields such as radiology, cardiology and ophthalmology (Anwar & Khan, 2023). It offers significant advantages over traditional analytics and clinical decision-making tools. However, to fully realize its potential, issues related to transparency, data privacy, ethical considerations, and regulatory frameworks must be addressed.

The global awareness of this technological change is growing. The legislators and various global organizations are highlighting the great potential but also multiple risks of AI in healthcare. The concerns raised are mainly due to the fact that, unlike many other domains, healthcare is considered a high-stake decision making environment (Zytek et al., 2021). Therefore, ensuring that all decisions are made as best as they can is of the highest importance. Many experts highlight the need for making the systems transparent and safe. However, few practical solutions have been suggested openly.

Unfortunately, great performance of AI algorithms is compromised with intractability of those systems. This lack of traceability, commonly known as the “black box” problem, means that while

AI can deliver highly accurate results, it is difficult to discern how these results were achieved. This opacity can lead to issues with trust, accountability, and biases, as users and stakeholders struggle to understand and verify the AI's decisions. Efforts to address this include developing explainable AI (XAI) techniques, which aim to make AI systems more interpretable and their decisions more transparent.

Explainability is recognized as a key ethical principles for trustworthy AI (European Comission, 2019; European Comission (ALTAI), 2020), addressing need for human agency and oversight, transparency, accountability as well as fairness. However, the precise relationship between XAI and these principles remains unclear, making it challenging to fully evaluate XAI's overall impact on the development of trustworthy AI systems. Many global leaders highlight the need for transparency and appropriate validation of the systems. For example WHO Director-General Dr Tedros Adhanom Ghebreyesus highlighted that 'Artificial intelligence holds great promise for healthcare, but also comes with serious challenges, including unethical data collection, cybersecurity threats and amplifying biases or misinformation' (World Health Organization, 2023). However, the means to ensure how those challenges should be addressed are not mentioned. The Future of Health (FOH), an international organization of senior health leaders highlights the need for making AI in healthcare trustworthy as one of top priorities (Silcox et al., 2024). However, they do not suggest or quote practical measures with exception of the articles that focus on prevention of racial or subgroup biases (Introducing HealthAI, 2023; Shah et al., 2024; Trevan et al., 2022). These kinds of biases are of great importance but they are not the core issue as they are not the cause but one of many consequences of lack of transparency. In those reports the 'black-box nature' of AI is usually not explained which suggests that the transparency might be considered as something that can be achieved if sufficient time and resources are invested in the creation of the system. It is a misconception as even small AI systems inherently lack transparency (Lipton, 2017).

Currently XAI techniques are considered to have great potential to combat 'the black-box' nature of AI. However, they can serve as only a proxy for transparency. To date there is no technique on the horizon that would have any chance at fully combating intractability of AI algorithms (Yang et al., 2023). Therefore, the awareness of the limitations of AI systems and XAI is not developed yet. Fortunately, the growing literature and studies on transparency can raise the awareness, estimate the potential and limitations of XAI and provide practical solutions to aforementioned concerns.

Although AI technology is making significant strides and revolutionizing various aspects of healthcare, it is improbable that doctors will be replaced by AI. The future of healthcare is more likely to involve a collaborative effort between AI and medical professionals. AI can support doctors by swiftly analysing large datasets, offering diagnostic recommendations, and predicting patient outcomes. However, the human touch, empathy, and deep understanding that doctors provide are irreplaceable (European Comission, 2019; Holzinger et al., 2019; Longo et al., 2020; Schneeberger et al., 2023). This partnership will allow doctors to concentrate more on patient interactions and complex decisions, while AI can manage data-driven tasks, resulting in a more effective and compassionate healthcare system.

Human-AI and human-XAI interaction or collaboration encompasses specifically the ways in which humans communicate and collaborate with AI and XAI systems. This field prioritizes the development of AI systems that are user-friendly, transparent, and fair, ensuring that they augment human abilities rather than replace them. The objective is to create tools that not only perform tasks as instructed but also are adapted to the users and evolve alongside them, while maintaining safety, ethical standards, and privacy. Ultimately, the goal is to foster a seamless partnership where support systems enhance human capabilities (Reverberi et al., 2022).

The present will work investigate the collaboration of the doctors with XAI in a user-centred study. Through a questionnaire-based survey, we collect clinicians' perception of three different types of XAI methods and evaluate the benefits and drawbacks of each method. This study's originality lies in its holistic evaluation of clinicians' perceptions of different XAI tools in a medical setting, a relatively underexplored area in recent literature (Gupta & Seeja, 2024; Manresa-Yee et al., 2022). By comparing an explainable-by-design Bayesian network with local XAI methods like Shap and Araucana tree, it provides nuanced insights into usability and cognitive load, aligning with recent findings on the importance of user-friendly explanations in healthcare. The integration of sentiment analysis to gauge emotional tone adds a novel dimension to understanding clinicians' interactions with XAI, echoing the call for more human-centred approaches and user experience considerations in XAI research (Liao & Varshney, 2022). Additionally, the study's focus on specific departmental preferences highlights the need for tailored XAI solutions, contributing to the growing body of work advocating for context-specific explainability in medical AI. It is crucial as XAI techniques have been developed primarily to serve model developers and AI researchers, not end users (Liao & Varshney, 2022).

Background

Technology and leveraging algorithmic data processing has been used for supporting human decision making since 1960s, around 30 years after the development of first digital computers in 1940s (Ferguson & Jones, 1969). First implementations and tests in the healthcare context are dated to 1970s (Chen et al., 2023). It means these types of systems have been around for only 50 years since the initial idea. These systems are usually called Decision Support Systems (DSSs). They are designed to assist in the decision-making process by analysing large volumes of data and presenting actionable insights. A DSS acts like as an intelligent guide, processing all available information to show possible solutions. They integrate data, analytical models, and user-friendly interfaces to help decision-makers evaluate different scenarios and predict outcomes with greater accuracy (Leong, 2003). DSS can enhance the quality and efficiency of decisions, ultimately leading to better healthcare functioning both for the patients and for the medical staff. However, since the initial idea has been introduced, the success at least in the domain of healthcare has been limited. The digitalisation and introduction of the electronic health records (EHRs) largely improved healthcare systems but not as the direct decision-making aid. The capabilities of DSSs were limited to risk assessment and simple medical decision-making cases, not creating a significant impact (Abell et al., 2023). Only the increase in processing speed and ability to handle high volumes of data and using more advanced statistical tools created a possibility of creating a significant impact in individual decision-making cases in healthcare. Nevertheless, more advanced systems had led to another problem, i.e., the lack of transparency undermining trust in the DSS (Jones et al., 2021).

Traditional DSS are tractable: they rely on clear logical rules and the decision process can be traced directly from the input data to the final output. On the other hand, the new generation of support systems relying on AI operates on what is usually called a ‘black-box’ reasoning. There is no clear and certain way of seeing why particular input results in a specific output of the system. This in turn has led to yet another field which tries to trace back and explain why a particular decision has been taken: Explainable AI (XAI).

In 2004, Van Lent and colleagues introduced the term XAI to describe systems capable of explaining the actions of AI-controlled entities (van Lent et al., 2004). XAI is crucial in the medical field because it addresses the ‘black-box’ problem, where AI systems make high-stakes decisions without providing reasoning behind it (Zytek et al., 2021). In healthcare, where decisions can have significant consequences, it is essential for clinicians to understand and trust the AI tools they use. XAI can enhance clinician confidence, improve patient outcomes, and ensure ethical standards are maintained by providing insights into how AI systems arrive at their conclusions, ultimately fostering better human-AI collaboration.

Medical decision making and cognitive processes

Clinical reasoning is essential for evaluating a patient’s condition and making medical decisions, enabling doctors to administer appropriate treatments. Effective clinical reasoning relies on three key areas of knowledge: diagnostic, etiological, and treatment knowledge (Elstein & Schwarz, 2002). Moreover, additional factors such as years of experience, both individually and within a department, significantly enhance clinical reasoning by improving all three aforementioned key areas (Dobber et al., 2023). Experienced clinicians can quickly recognize patterns and generate accurate hypotheses, while experienced departments benefit from collaborative knowledge and refined protocols. This collective expertise helps mitigate cognitive biases, ensuring more accurate and effective medical decisions (Dobber et al., 2023).

The interplay of knowledge and the reasoning processes in medical decision-making can be understood through multiple frameworks. Researchers distinguish different types of clinical reasoning a simple example being deductive and inductive reasoning (Shin, 2019). Clinical reasoning can be also understood and studied by considering more abstract multiple decision-making streams as well as their consequences and pitfalls.

Here one of the most influential and relevant frameworks in the context of decision-making will be reviewed, i.e., the dual-process theory. Interestingly, this theory can be applied in three different ways: to analyse the reasoning of doctors on their own, the decision-making of the AI-DSSs as parallel to human cognition, and the joint (clinician and AI-DSS) decision-making process. Below, the first - the decision-making of doctors – the traditional application of the theory will be summarised followed by the interpretation of joint decision-making of the clinician and AI-DSS.

The dual-process theory, described by Kahneman and Tversky, proposes two distinct cognitive systems that influence human decision-making. System 1 is fast, intuitive, and emotional - it is a parallel to the inductive reasoning, whereas System 2, which is slower, more deliberative, and more logical is more aligned with deductive reasoning (Kahneman, 2011). The clinician's intuitive expertise (System 1) rapidly assesses a situation based on learned patterns and experiences, while System 2 provides a deliberative analysis and logical explanations why a certain decision should be taken. According to Kahneman's theory, these systems interact and can sometimes conflict, explaining in simple terms many pitfalls in the clinical decision-making process.

System 1 is typically relied upon when decisions need to be made quickly, based on pattern recognition and heuristics (decision-making proxies and shortcuts), which can be beneficial in emergency situations. However, this system is also more susceptible to cognitive biases, such as the availability heuristic, where clinicians may judge the probability of an event by the ease with which instances come to mind, potentially leading to overestimation of common conditions and underestimation of rare ones. System 2, on the other hand, is engaged when more complex decisions are required, involving data analysis and critical thinking, such as weighing the risks and benefits of a particular treatment plan. This system is typically engaged when a DSS is used, providing a structured framework for analysing clinical data and guiding choices. Yet, even with DSSs, biases can infiltrate the decision-making process. For instance, confirmation bias may lead to seeking out information that supports the preconceived notions or diagnoses, while neglecting data that contradicts them. Similarly, anchoring bias can cause clinicians to rely too heavily on the first piece of information encountered - often initial test results or patient presentation when making decisions.

The interaction between these systems and DSSs can sometimes lead to suboptimal clinical decisions. For example, when under time pressure or cognitive overload, clinicians may default to System 1, despite the availability of a DSS designed to engage System 2 processes. This can result in diagnostic shadowing, where the presence of a salient feature (such as a patient's known medical condition) overshadows the accurate interpretation of symptoms, leading to misdiagnosis or inappropriate treatment. Moreover, the framing effect, where the same information presented in different ways can lead to different decisions, can be exacerbated by the design and output of DSS, which may present data in a way that inadvertently influences the clinician's judgments.

To mitigate these biases, it is crucial for DSS to be designed with an understanding of dual-process theory, ensuring that they support System 2 processing without inadvertently reinforcing System 1 biases. This includes providing balanced information presentation, prompts for critical thinking, and preventing strategies for common cognitive biases. Additionally, training clinicians to be aware of these biases and to consciously engage System 2 thinking when using DSS can further enhance decision-making quality. Ultimately, the dual-process theory provides a valuable framework for understanding the cognitive underpinnings of clinical decision-making and the potential pitfalls that can arise when interacting with DSSs. By acknowledging and addressing the interplay between intuitive and analytical thinking, healthcare providers can improve the accuracy and effectiveness of their decisions, leading to better patient outcomes.

AI and predictive modelling in medicine

Predictive AI in medicine involves using advanced algorithms and machine learning models to anticipate patient outcomes, diagnose diseases, and personalize treatment plans. There are several types of predictive AI models. Supervised learning involves training models on labelled data to predict outcomes, such as regression for continuous values and classification for discrete labels (Sarker, 2021). In contrast, unsupervised learning uses unlabelled data to uncover hidden patterns, with clustering being a common technique. Deep learning, a subset of machine learning, employs neural networks for complex tasks and can be used in both supervised and unsupervised learning. Convolutional Neural Networks (CNNs) are typically used for image recognition, while Recurrent Neural Networks (RNNs) handle sequential data and time series prediction (Sarker, 2021).

These models analyse vast amounts of medical data, including electronic health records and imaging, to identify patterns and make predictions. However, challenges persist, including data quality and heterogeneity, model interpretability, and the integration of AI systems into clinical workflows. Ensuring the ethical use of AI and preventing cognitive biases mentioned in the previous chapter (Medical decision making and cognitive processes) are of utmost importance especially in high-risk medical decision-making (Maleki Varnosfaderani & Forouzanfar, 2024).

Traditionally, predictive AI is viewed as an impersonal process, with algorithms operating autonomously on passively collected data. However, it should also be seen as a tool for achieving human goals. Human-centred AI acknowledges the essential role of people in the process and adapts workflows to fit human working practices. This approach highlights that machine learning systems are not just autonomous entities but are designed to be used by people to achieve specific objectives. By integrating human insights and contextual understanding of DSSs, the development and deployment of machine learning models can result in systems that are more intuitive and effective (Gillies et al., 2016). The design process should focus on making the tools comprehensible and helpful, ensuring that users can easily understand and effectively utilize the technology. This perspective conceptualizes computational learning, emphasizing the importance of usability especially in high-stakes environments like healthcare.

Newest reports indicate that human-AI collaboration is more effective and can outperform the use of the AI prediction itself and the clinician working alone (Reverberi et al., 2022). Various protocols are being investigated at the moment (Cabitza et al., 2023).

By focusing on human-AI collaboration, the usability of DSSs can be enhanced leading to better diagnostic accuracy and efficiency, early detection of diseases, reduced human error, and continuous improvement of care.

Explainable AI

Explainable AI (XAI) originated from the need to make complex AI models more transparent and understandable, especially in critical decision-making settings. Traditional AI models, often referred to as ‘black boxes’, usually provide no insight into how they arrive at their predictions. It can be problematic in medical decision-making where understanding the rationale behind a diagnosis or treatment recommendation is crucial for the patient and as feedback for doctors. XAI addresses this by offering clear, interpretable explanations of AI decisions, enhancing trust and accountability. The benefits of XAI in healthcare are significant: it helps clinicians understand and validate AI-driven insights, supports informed decision-making, and ensures compliance with regulatory standards. Moreover, XAI can enhance patient and doctor's trust by providing transparent explanations of AI-assisted diagnoses and treatments, ultimately leading to better patient outcomes and safer use of AI DSSs (Longo et al., 2020).

The recent and fast advancement of AI implementation has led to development of multiple XAI tools. They can be categorized in terms of the stage of applicability (ante-hoc vs post-hoc), scope of the explanation (global vs local) and the general model applicability of the technique (model-specific vs model-agnostic) (Figure 1). A description of the categories is outlined below.

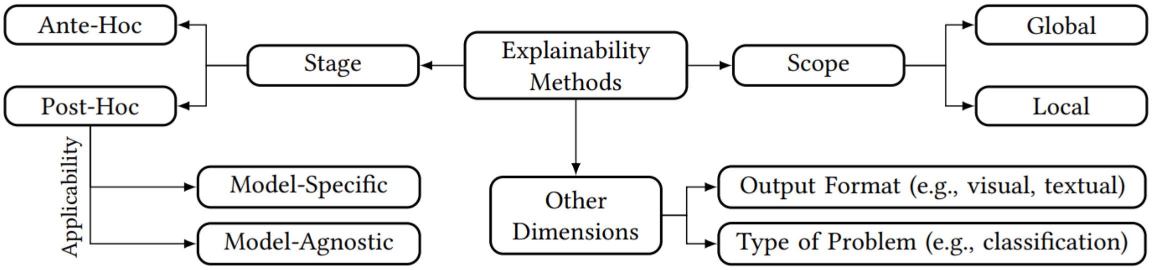


Figure 1. Taxonomy of XAI. Three main categories of classification include the stage during which the technique is applied, scope of the explanation and applicability (Speith, 2022).

Ante-hoc vs post-hoc

Ante hoc (intrinsic) methods are designed to be inherently interpretable from the outset, meaning the models are built with transparency in mind. These methods provide clear and understandable decision paths, ensuring that the reasoning behind predictions is accessible and straightforward. Models such as linear regression, decision trees, k-nearest neighbours, rule-based learners, general additive models, and Bayesian learners are typically considered ante-hoc explainable, if they are not excessively large (Speith, 2022). However, these models are not suitable for all machine learning problems due to their lower performance. They usually do not achieve the same level of accuracy as opaque models like deep neural networks. This is commonly referred to as the performance-explainability trade-off (Speith, 2022).

On the other hand, post hoc (black-box) methods are applied after a model has been trained. These methods aim to explain the decisions of complex, often opaque models like deep neural networks. Techniques such as LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016) and Shap (Lundberg & Lee, 2017) fall into this category, offering insights into model behaviour without altering the original model. Another example is rule-based decision tree - Araucana used in this study (Parimbelli et al., 2023).

Both approaches are crucial for enhancing the transparency and trustworthiness of AI systems, particularly in high-stakes fields like healthcare. Ante-hoc methods are integrated into the model and inherently interpretable but are not easily transferable. Post-hoc methods are independent of the model architecture and can explain any trained model without affecting its accuracy. Both approaches bring some unique advantage to explainability (Vale et al., 2022).

Local vs global

Local explanations interpret decisions based on specific data instances, while global explanations provide insights into the overall model behaviour.

Local explainability methods are believed to likely be more important to those affected by AI system outputs than to regulators creating AI legislation. While individuals may want to determine if they have experienced discriminatory decisions, regulators are focused on reducing discrimination overall (Langer et al., 2021). Techniques for local explainability, such as LIME and Shap, are vital for individuals aiming to comprehend particular AI-driven decisions, like a patient exploring the reasons behind a particular medical treatment chosen. One example is the use of Shap for assessment of the whole brain radiotherapy treatment (Wang et al., 2022).

In contrast, healthcare regulators might employ global explainability approaches to verify that AI systems do not systematically discriminate against any demographic group. Some examples include introduction of a model that combined a binary decision tree with a clustering algorithm (Eiras-Franco et al., 2019). Another example is GLocalX, which is a ‘local-first’ model agnostic explanation method (Setzu et al., 2021). It starts with local explanations in the form of local decision rules and iteratively generalizes them into global explanations through hierarchical aggregation. This approach aims to create interpretable models that can simulate or replace black box models, achieving high accuracy and comprehensibility without trading the performance for transparency. The scope distinction of XAI emphasizes the differing priorities between personal and systemic view in the use of AI.

Model-specific vs model-agnostic

Lastly, model-specific methods are tailored to particular model architectures, whereas model-agnostic methods can be applied across different models and domains.

The selection between model-specific and model-agnostic methods is contingent upon the particular explainability goals and the type of model employed (Langer et al., 2021). Developers might have a preference for exploring model's parameters using model-specific methods instead of model-agnostic ones for greater precision. Whenever a profound understanding of model behaviour is essential like in research setting, model-specific methods are often prioritized. For instance, when the objective is to investigate the internal parameters of a complex neural network, a model-specific method like Layer-wise Relevance Propagation is preferred due to its precision (Binder et al., 2016). Similarly, researchers or developers examining the functionality of convolutional neural networks in image recognition tasks might utilize Integrated Gradients to attribute predictions to input features (Sundararajan et al., 2017). Conversely, if the goal is to provide generalizable explanations across various models a model-agnostic approach such as Shap is more suitable due to its flexibility. It holds true for industrial setting and commercial applications, especially if the end-users do not have AI expertise. Model-agnostic methods offer consistent explanations across different models, thereby enhancing transparency and building trust in AI systems by comparative studies (Gupta & Seeja, 2024).

The aforementioned taxonomies constitute only the most general part of the XAI categorizations. More detailed taxonomy can be found in the reviews that focus specifically on the XAI classification (Arrieta et al., 2019; Schwalbe & Finzel, 2023; Speith, 2022).

In this study, three XAI methods were chosen for the assessment. They are explained further in the following sections – Bayesian network, tree-based Araucana and Shap. Table 1 shows their categorization according to the XAI taxonomy described in this section.

Table 1. The summary of the taxonomy of the explanations used in this study.

XAI TECHNIQUE	STAGE	SCOPE	APPLICABILITY
ARAUCANA	Post-hoc	local	Model-agnostic
BAYESIAN NETWORK	Ante-hoc	global	Model-specific
SHAP	Post-hoc	local	Model-agnostic

The InXAID project

This thesis is part of the Italian project PRIN PNRR 2022 ‘InXAID - Interaction with eXplainable Artificial Intelligence in (medical) Decision making’ CUP: H53D23008090001, funded by the European Union - Next Generation EU.

This project focuses on enhancing the explainability of AI models in DSSs, especially in critical healthcare settings. It aims to develop a methodological framework for designing and validating Human-AI collaboration protocols, ensuring AI systems are fair, ethical, and trustworthy. The project will investigate XAI approaches to improve decision quality, reduce biases, and increase user confidence. Ultimately, it seeks to foster effective human-AI collaboration, leading to better decision-making outcomes. The inXAID project aims to enhance healthcare DSS by developing and evaluating AI-based models and methods. The project will adopt a model-agnostic approach, emphasizing the interaction between DSS and users to maximize the positive influence and appropriateness of DSS support. Expected outcomes include protocols for efficient and safe data input and result interpretation, ensuring users can make informed decisions based on AI system outputs.

Objectives and hypotheses

Most of the literature currently focuses on theoretical predictions identifying potential pitfalls and creating explainability frameworks. It leads to multiplication of terminology and speculation both in the scientific literature and in legislation (Schneeberger et al., 2023). However only experimental approaches can provide evidence as to which solutions are viable and feasible in real settings.

In the present study, we aim to combine usability research and cognitive psychology to evaluate the effectiveness and empirically test different XAI explanations. Our work focuses on three XAI techniques, built on top of the ALFABETO project (Nicora et al., 2021) - Araucana, Bayesian network and Shap.

The following objectives are set and addressed with corresponding research questions and hypotheses. The first objective of the thesis is the investigation of human-XAI collaboration in the medical setting by high-level analysis and integration of XAI research, psychological and usability studies (Table 2 RQ1 & RQ2). The second objective is quantitative assessment of the human-XAI collaboration to empirically compare selected XAI methods (Table 2 RQ3 & RQ4). The analysis is supported by complementary objective to compare Emergency room (ER) and Infectious diseases (ID) departments as well as clinician pairs to provide further context for the results (Table 2 RQ5 & RQ6). The final objectives are: the assessment of emotional tone as an approach in XAI-interaction research and positioning of the results in the theoretical explainability frameworks (Table 2 RQ7 & RQ8). The objectives are reached through the aforementioned research questions and corresponding hypotheses (Table 2; Table 3).

Table 2. Research questions. The abbreviations refer to two studied departments: ER – Emergency room, ID – Infectious diseases.

RQ1: General Perception: What are the perceptions of clinicians regarding the use of human-AI collaboration tools in a medical setting?
RQ2: Compliance: What are the compliance rates and how do they compare to established frameworks?
RQ3: Method Comparison: How do clinicians' perceptions of comprehensibility, helpfulness and cognitive load differ among the three XAI methods (Shap, Araucana tree and Bayesian network)? Do they align with directly expressed method preferences?
RQ4: Sentiment Analysis (emotional tone): What is the emotional tone of clinicians' attitudes towards each XAI explanation assessed through sentiment analysis?
RQ5: Department Comparison: Are there any differences in terms of method perception and preference between ER and ID departments on average?
RQ6: Pair Comparison: Are there any similarities or consistent patterns of comprehensibility and helpfulness ratings within allocated ER - ID clinician pairs?
RQ7: Sentiment Analysis (suitability): Are general purpose sentiment analysis models a suitable tool for detecting emotional tone while assessing human-XAI interactions in the medical setting.
RQ8: Explainability assessment: How does selected research design: survey, think-aloud and sentiment analysis fit into existing theoretical explainability frameworks?

The first objective is reached through the evaluation of the literature, initial questionnaire and analysis of compliance (Table 3 H1 & H2). The second objective is addressed with the analysis of reported assessment of comprehensibility, helpfulness and method preferences as well as think-aloud sessions assessing the cognitive load through time measurement (Table 3 H3). Sentiment analysis is performed to assess XAI methods in terms of emotional tone (Table 3 H4). Comparison between departments and pairs is achieved through additional analyses of the survey answers (Table 3 H5 & H6). Finally, evaluating the relevance of sentiment analysis and the overall study design in terms of theoretical explainability is performed (Table 3 H7 & H8).

Table 3. Study Hypotheses. The abbreviations refer to two studied departments: ER – Emergency room, ID – Infectious diseases.

H1: General Perception: Clinicians generally perceive human-AI collaboration tools as positive and trustworthy in the medical setting.
H2: Compliance: Significant proportion of the compliant decisions are incorrect, indicating potential over-reliance on the XAI system.
H3: Method Comparison: There are significant differences in comprehensibility, helpfulness and cognitive load measures among Shap, Araucana tree and Bayesian network.
H4: Sentiment Analysis (emotional tone): Sentiment analysis reveal differences in the emotional tone between Shap, Araucana tree and Bayesian network mirroring the perceptions found with the use of the survey.
H5: Department Comparison: There are significant differences in tool perceptions and preferences between ER and ID departments, with ER clinicians preferring tools that provide rapid, comprehensible explanations and ID clinicians preferring tools that offer more detailed, in-depth explanations.
H6: Pair Comparison: Within allocated clinician pairs, there are consistent patterns in comprehensibility and helpfulness ratings, reflecting similar perceptions and preferences of clinicians from ER and ID departments while assessing same patients and corresponding XAI explanations.
H7: Sentiment Analysis (suitability): General purpose sentiment analysis models are suitable for assessing emotional tone towards XAI tools in the medical setting.
H8: Explainability Assessment: Studied tools can be fit into theoretical frameworks and highlight the need for theoretical and empirical studies being conducted together.

Methods

Related work

Alfabeto project

The study was done as a part of the ALL FAster BEtter Together - ALFABETO project which aims to explore clinicians' perceptions of various explanations for AI classifications generated by different XAI methods and evaluate their impact on users (Catalano et al., 2023; Nicora et al., 2021).

Within the ALFABETO project, machine learning models were developed to predict whether a COVID-19 patient in electronic health records (EHR) requires hospitalization, classifying patients into 'Home' or 'Hospital' categories based on clinical features. The ALFABETO project was focusing on clinical and chest X-ray data. To address the interpretability issue, explainable-by-design Bayesian networks were used as they incorporate medical knowledge and patient data in a complete way. They were investigated and found to be both explainable and effective, performing comparably to less interpretable models across various COVID-19 waves. The training data was collected from 660 COVID-19 patients treated at the I.R.C.C.S. Policlinico San Matteo Foundation of Pavia, Italy. Half of the patients required hospitalization, while the rest had a better prognosis and were managed at home. For each patient, additional clinical characteristics were collected such as age, gender, and comorbidities. Deep Learning was utilized to extract features from chest radiographs (RX): a deep network X-RAIS was used to analyse different types of medical images and extract relevant diagnostic information. Here, X-RAIS transformed the RX images into five numerical features: Consolidation, Infiltration, Edema, Effusion, and Lung Opacity. These five features, combined with 19 clinical features, served as inputs for a machine learning model to predict whether a patient should be hospitalized (class 1) or not (class 0). 90% of the patients were randomly selected for the training set, and 10% were used for testing and identifying the best-performing model.

To create the Bayesian Network, initially a graph was designed based on existing knowledge to reflect clinicians' decision-making. This graph included a target node for hospitalization decisions, considering factors like age, gender, and breathing difficulties. then the graph was enhanced using the hill climbing search algorithm on training data, which adjusts edges to maximize a fitness score until a local maximum is achieved (Nicora et al., 2021).

Bayesian network for the prognostication of COVID-19 ER cases

Bayesian networks play a significant role in constructing predictive systems by offering both local and global explanations through their probabilistic graphical models. Bayesian networks are explainable-by-design. They are probabilistic graphical models that represent a set of variables and their conditional dependencies via a directed acyclic graph (DAG). This structure inherently provides transparency in the model, reasoning, and evidence (Derks & de Waal, 2020). These networks use DAGs to represent variables and their conditional dependencies, making them inherently interpretable. On a local level, Bayesian networks can clarify individual predictions by showing the probabilistic relationships and dependencies that led to a specific outcome. On a global level, they provide insights into the overall structure and behaviour of the model, showing how different variables interact and influence each other across the entire dataset.

A DAG is a type of graph used in various fields like mathematics and computer science to represent objects connected by directed edges, ensuring no cycles are present (Figure 2-left). This means it is not possible to start at one node and follow a path that leads back to the same node. DAGs consist of nodes (vertices) and directed edges (arcs), which prevent the formation of loops. They are particularly useful for modelling processes with a clear directional flow by visualizing dependencies and sequences of events. In Bayesian networks, DAGs represent probabilistic relationships among variables, ensuring the network can be interpreted as a series of conditional dependencies, making them valuable for understanding complex systems.

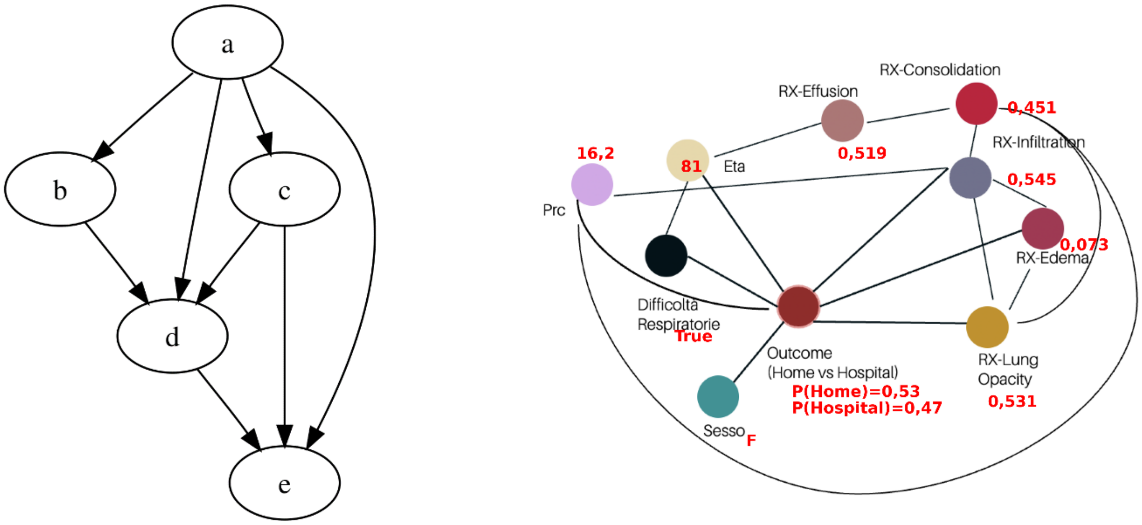


Figure 2. Examples of a directed acyclic graph ('Directed Acyclic Graph', 2024) (left) and the example of the Bayesian network used in the survey (right).

The process of constructing a DAG Bayesian network involves few steps. First, relevant variables that represent different aspects of the system being modelled need to be defined, including the possible values for each variable, which can be either discrete or continuous. In this study the parameters of the systems are blood test results and relevant details of the patients collected during the COVID-19 pandemic. Here, the parameters are marked as circles and red values represent corresponding conditional probabilities (Figure 2). Then, the network structure is created by connecting the variables with directed edges that represent causal or conditional dependencies, ensuring the graph remains acyclic. After that, the Conditional Probability Tables (CPTs) for each variable, which quantify the probabilities of the variable's values given the values of its parent variables in the DAG, have to be specified. Finally, the network should be validated by comparing its predictions with known data or expert knowledge to ensure it accurately represents the system. This process results in a Bayesian network that can be used for tasks such as diagnostics, reasoning, and decision-making under uncertainty (Boettcher & Dethlefsen, 2003).

Bayesian networks and other artificial intelligence algorithms differ significantly in their theoretical foundations, operational mechanisms, and applications. Bayesian networks are probabilistic graphical models that use DAGs to represent variables and their conditional dependencies. They are highly transparent and interpretable, making them ideal for fields like healthcare where understanding the reasoning behind predictions is crucial. In contrast, neural networks and other AI algorithms, such as neural networks and decision trees, often prioritize predictive accuracy over interpretability. Neural networks, for example, are far superior at handling large, complex datasets and learning intricate patterns but are often seen as "black boxes" due to their lack of transparency. Additionally, Bayesian networks can incorporate both data-driven and expert knowledge, offering flexibility and modularity, while other AI algorithms may not provide the same level of insight into their decision-making processes (Kitson et al., 2023).

AraucanaXAI – tree-based explanations of ML predictions

A decision tree is a tool used for making decisions and predictions. It resembles a flowchart, where each node represents a decision point or a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a final decision or classification (Figure 3) (*What Is a Decision Tree?*, 2021).

The AraucanaXAI (Araucana) (Parimbelli et al., 2023) method is an innovative technique for providing local explanations of machine learning model predictions. It utilizes decision trees to create interpretable models capable of managing non-linear decision boundaries, ensuring high fidelity to the original model's predictions. By growing unpruned trees, the Araucana enhances the accuracy of its explanations, making it particularly valuable in critical applications like medicine where transparency and reliability are essential (Figure 4). Furthermore, this model-agnostic approach is compatible with a variety of machine learning models, which makes it a flexible tool for improving the interpretability of complex AI systems.

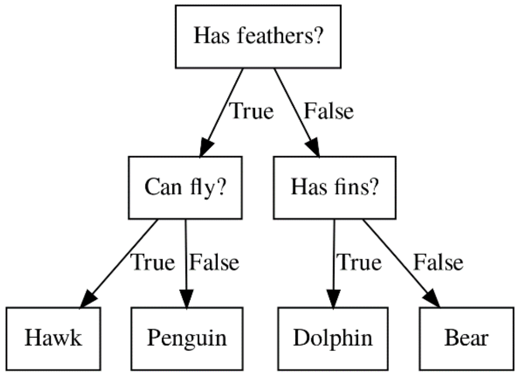


Figure 3. A simple example of a deterministic true/false decision tree.

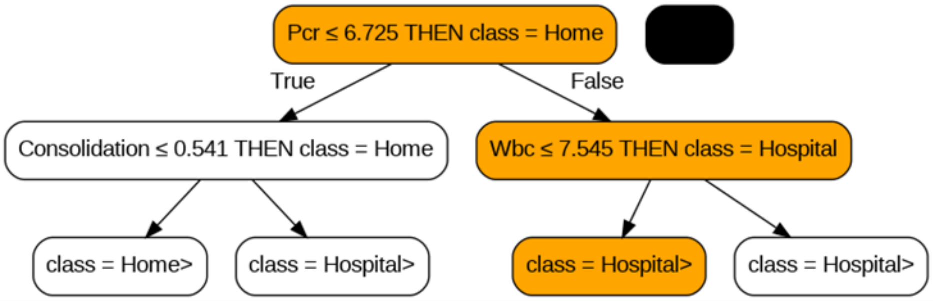

Pcr ≤ 6.725 THEN class = Home
True
False
Consolidation ≤ 0.541 THEN class = Home
Wbc ≤ 7.545 THEN class = Hospital
class = Home>
class = Hospital>
class = Hospital>
class = Hospital>

Figure 4. An example of Araucana output tree. In the example of the Araucana tree for this patient PCR = 16.2 and Wbc = 6.4, so the first condition is false (16.2 > 6.725) and after moving down the tree the second condition is true (6.4 < 7.545), resulting in 'Hospital' classification. The explanation path is highlighted in orange.

Araucana can be more suitable over other local, model-agnostic, post-hoc explanation methods, and improve fidelity compared to LIME (Ribeiro et al., 2016), and the ability to handle both classification and regression problems. The approach reuses the original training set, which helps uncover biases and unexpected model behaviours. However, it has limitations, including the complexity of unpruned explainer trees and the need for the original training set.

Constructing a decision tree involves a series of steps aimed at creating a model that can make decisions based on input data. The process begins with selecting the best attribute to split the data, which is typically done using criteria like information gain. Information gain measures the reduction in entropy after splitting a dataset based on a specific attribute. The attribute that achieves the highest information gain is the most effective at classifying the training data according to its target classification, as it best reduces uncertainty. The data is then divided into subsets based on the chosen attribute, and this process is recursively applied to each subset. This recursive splitting continues until a stopping condition is met, such as all data points in a subset belonging to the same class or reaching a maximum tree depth. The final result is a tree where each internal node represents a decision based on an attribute, each branch represents the outcome of that decision, and each leaf node represents a final classification or decision.

One advantage of the Araucana tree is its completeness because the trees are constructed using all the parameters (Parimbelli et al., 2023). However, in practical applications such as DSSs in healthcare the trees have to be pruned to include fewer parameters in favour of comprehensibility and understandability.

Shapley values

Shapley values or SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017) here referred to as Shap, originating from cooperative game theory, are pivotal in enhancing the explainability of AI systems, particularly in medicine. They provide an effective method to attribute the overall outcome of a predictive model to individual features (Figure 5). By calculating the marginal contribution of each feature across all possible combinations, Shap helps interpret complex AI models, making them more transparent. Shap is an example of local explanations. In medical applications, this means understanding which patient characteristics (e.g., age, genetic markers, lifestyle factors) most significantly influence predictions, such as disease risk or treatment efficacy.

This transparency not only aids in personalized medicine by tailoring treatments to individual patients but also ensures that healthcare providers can trust and effectively utilize AI-driven insights for better patient outcomes (Feretzakis et al., 2024) (Ter-Minassian et al., 2023).

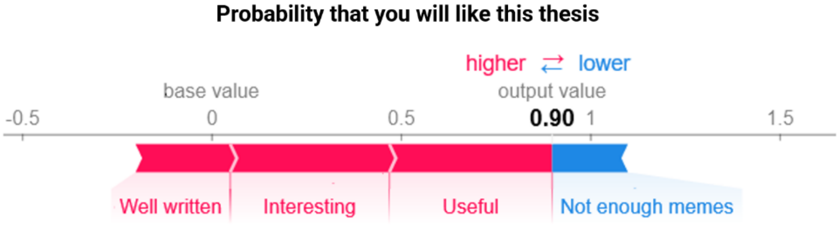


Figure 5. A simple example of summation of positive and negative Shapley values which results in the final prediction probability 0.9 (E. Gerber, 2019).

In the context of the study, Shap explanation was providing the clinicians with the presentation of seven variables and their contributions to the overall prediction for each patient (Figure 6).

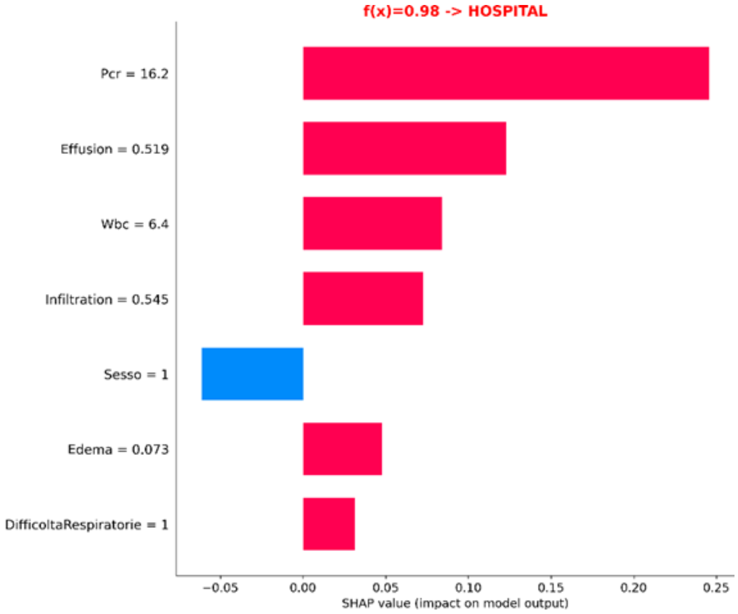


Figure 6. An example of Shap explanation used in the survey. The bars show the parameters and their contribution to the final prediction. Red bars show the contribution that pushes the prediction towards hospitalization, blue bar shows the contribution that pushes the prediction towards discharge from the hospital. (y-axis parameters – PCR (Polymerase chain reaction), Effusion, White blood cells (WBC), Infiltration, Sex, Edema, Respiratory difficulty).

Studies evaluating XAI in medical decision making

Many of empirical studies highlight the fact that human-AI and human-XAI collaboration is more fruitful and achieves better results than clinicians or the AI alone (Cabitza et al., 2023; Reverberi et al., 2022). The existing literature on XAI emphasizes theoretical frameworks and implications, often highlighting potential pitfalls and proposing various solutions. However, as highlighted by (Holzinger et al., 2019), achieving explainable medicine requires measurement of the quality of explanations, similar to how usability measures the quality of use. Importantly, it differentiates causability – a human attribute, from explainability – a system's attribute. It distinguishes explainability as a technical term related more to model engineering which rely predominantly on statistical learning as opposed to the human cognition wired for making causal inferences (Pearl, 2018). The proposed human-centered approach states that successful explanations have to leverage the ways in which human experts make decisions in contrast to just providing an explanation of a predictive model (Holzinger et al., 2019).

For example, two XAI user studies: one with 12 specialist radiologists and another with 44 ECG readers who assessed 240 and 20 cases, respectively, explored different collaboration setups (Cabitza et al., 2023). The results confirm the utility of AI support but also reveal that XAI can sometimes lead to a "white-box paradox," resulting in neutral or negative effects (Cabitza et al., 2023). This paradox underscores that although explanations aim to make AI systems more transparent and comprehensible, they can inadvertently foster a false sense of security, leading users to depend too much on the AI's recommendations, even when they are incorrect. Additionally, Cabitza et al., 2022 puts much emphasis on the importance of the presentation sequence. For example, AI recommendation can either be seen first before the clinician makes the decision, in contrast to clinician making the decision after seeing AI recommendation. According to the study decision-making after seeing AI recommendation leads to higher diagnostic accuracy (Cabitza et al., 2022).

Another study (Lesley & Kuratomi Hernández, 2024) using mixed-methods approach evaluating XAI concluded that XAI explanations often do not meet physicians' expectations for reducing uncertainty and providing detailed explanations. Interestingly, the areas deemed most important by physicians performed the worst, while it excelled in less critical areas. Physicians also emphasized the need for better AI and XAI training for clinical staff, integration of these tools into routine healthcare and the ability to customize XAI explanations (Lesley & Kuratomi Hernández, 2024).

However, many systems already in practice have not been widely studied in terms of the collaboration with the clinician (Troya et al., 2022) such as computer-aided polyp detection (CAD) which is already adopted in clinics. When under investigation, CAD systems outperformed human examiners in terms of the speed of polyp detection. However, the integration of CAD did not lead to a quicker detection time for human examiners. The use of CADe systems resulted in more frequent misinterpretations of the mucosa and a decrease in the distance the eye traveled during mucosal inspections. Analysis of false positives and eye-tracking data indicated significant changes influenced by CADe, suggesting a potential risk of overdependence and skill degradation when these systems are employed (Troya et al., 2022). It suggests that the systems should be widely studied before the implementation into clinical practice.

Overall, the studies call for human-centered XAI research that respond to the needs of particular users preferably before the implementation of the systems in clinics (Kong et al., 2024). One major literature gap is the lack of comprehensive user studies that incorporate insights from cognitive and social sciences to evaluate the effectiveness of XAI methods (Kong et al., 2024; Liao & Varshney, 2022). While many studies focus on developing new algorithms and technical solutions, there is a shortage of research that examines how these solutions are perceived and utilized by end-users, particularly non-experts (Severes et al., 2023). Additionally, existing studies often fail to address the diverse needs of different stakeholders in healthcare settings. This gap highlights the need for more interdisciplinary approaches that combine technical, cognitive, and social perspectives to create XAI systems that are truly user-friendly and effective in real-world applications (Kong et al., 2024). Furthermore, there is a need for standardized evaluation metrics and frameworks that can consistently measure the usability and trustworthiness of XAI systems across various domains (Severes et al., 2023). Addressing these gaps is crucial for advancing the field of human-centered XAI and ensuring that AI technologies are both explainable and beneficial to all users.

These points are addressed in the present study by integration of cognitive psychology, usability and explainability frameworks. It is focused on end-users which are not familiar with neither AI nor XAI. It provides further context by identifying differences between two different medical wards showing the diversity and complexity of the medical setting. The originality of the study lies in the diversification of the techniques used for evaluation of XAI to provide stronger and more contextualized evidence. Finally, it positions the study in the theoretical explainability frameworks which is a crucial step towards creation of standardized evaluation metrics for XAI usability and trustworthiness.

Study design

Study participants

10 experienced clinicians from two departments (5 from emergency room, i.e. ER, and 5 from the infectious disease, i.e. ID) took part in the survey. All of them belong to I.R.C.C.S. Policlinico San Matteo Foundation of Pavia, Italy. Collecting the data from clinicians from two different departments provides additional insights because their decision-making is guided by different environments. ER requires often fast decision-making and action followed by immediate consequences. On the other hand, ID department allows for more deliberate decision-making that take into account more data, including imaging, careful examination and interview with the patient.

The years of experience among clinicians, collected through an initial questionnaire (explained in detail in the following section), ranged from 1 to 30. The study involved 3 female and 7 male clinicians.

The clinicians from the two departments were matched in pairs in order to compare the responses of two departments – ER and ID. The pairs of clinicians were created to best match the years of experience within the pair (Table 4).

Table 4. Table showing years of experience (Exp.) and pairing of clinicians from two departments. ID – Infectious diseases, ER – Emergency room.

	ID		**ER**	
PAIR INDEX	Clinician index	Exp.	Clinician index	Exp.
1	5	1	8	2
2	1	2	9	6
3	3	5	2	10
4	6	15	7	24
5	4	30	10	30

Survey instrument

In this study, KoboToolbox was employed to conduct the survey. This instrument facilitated the systematic and accurate collection of data, ensuring that all responses were securely recorded and readily accessible for subsequent analysis.

Study protocol

The study was initiated with a short questionnaire designed to gauge the general attitudes of the clinicians towards AI in medicine. It was composed of 5 yes/no questions (Table 5).

The information collected included also years of experience, department and sex. The structure and layout of the initial questionnaire was included in the Appendix in section 1.1. The questions were used to assess the general familiarity with AI of the clinicians and the general attitude – skeptical or positive.

Table 5. Five questions used in the initial questionnaire in Italian and their corresponding translation to English. Each question assesses either familiarity – Fam. or attitude – Att. towards AI in healthcare.

NO.	QUESTION (IT)	QUESTION (EN)	GROUP
1	Ho una buona conoscenza nell'ambito dell'Intelligenza Artificiale.	I have good knowledge in the field of Artificial Intelligence.	Fam.
2	Sono stato in contatto e/o ho usato sistemi di Intelligenza Artificiale nel mio lavoro.	I have been in contact with and/or used Artificial Intelligence systems in my work.	Fam.
3	Sono convinto che un'Intelligenza Artificiale possa aiutarmi a rispondere piu orrettamente e velocemente a domande di cui non conosco le risposte con certezza.	I am convinced that Artificial Intelligence can help me answer questions more correctly and quickly when I am uncertain.	Att.
4	Sono convinto che farmi aiutare da un'Intelligenza Artificiale (ad esempio un assistente virtuale) nel mio lavoro o nello studio possa aumentare la mia produttivita.	I am convinced that being assisted by Artificial Intelligence (e.g., a virtual assistant) in my work or study can increase my productivity.	Att.
5	Sono convinto che farmi aiutare da un'Intelligenza Artificiale possa migliorare l'efficacia di quello che faccio.	I am convinced that being assisted by Artificial Intelligence can improve the effectiveness of what I do.	Att.

A few weeks after the initial questionnaire, clinicians were presented with the main survey. The completion dates varied, as clinicians could fill out both the initial questionnaire and the main survey at their convenience. Each clinician was presented with ten patient cases. These patients were selected from a larger dataset of 50 patients (ALFABETO test set), ensuring that each case was evaluated twice.

Half of the 50 cases represented correct predictions by the algorithm, with accurate explanations: true negative (TN) and true positive (TP). These refer to correctly predicted home discharges and hospitalizations, respectively. The other half represented incorrect predictions (with consistent explanations): false positive (FP) and false negative (FN). Consequently, each clinician evaluated an equal number of TP, TN, FP, and FN cases.

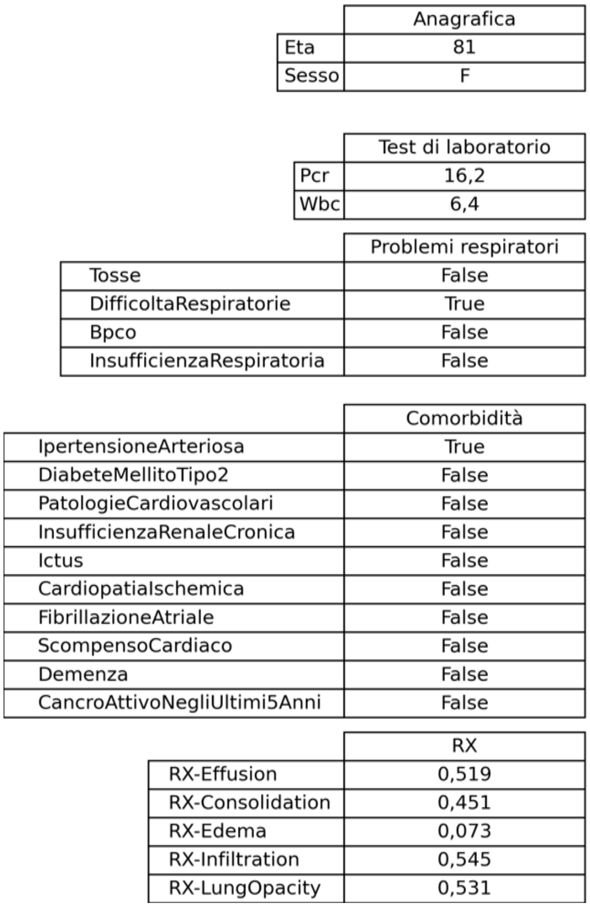
To prevent negative bias against the explanations and predictions, TP and TN cases were always shown first (Cabitza et al., 2022b; Kim et al., 2020). After these correct classifications, the FP and FN cases were presented.

To compare the two departments, clinicians participating in the survey were paired based on their years of experience prior to the experiment. Each pair of clinicians was presented with the same patients in the same order to directly compare their responses. Thus, 10 clinicians resulted in 5 pairs (Table 4).

Each patient's case was presented as follows: (1) the patient's characteristics, (2) the predicted class and its explanations, and (3) the final considerations. The exact structure and layout of each section can be viewed in section 1.2 of the Appendix.

For each patient, the first piece of information shown was a table summarizing the patient's characteristics (Table 6). These characteristics represent standard tests for patients admitted to the hospital during the COVID-19 pandemic and serve as the machine learning input features. The results should provide enough information to decide if the patient should be hospitalized or sent home.

Table 6. An example of a table that represents the initial presentation of the patient's characteristics to the clinician, i.e. personal data (age, sex), laboratory test results (PCR, WBC - White Blood Cells), breathing problems (cough, respiratory difficulty, COPD - chronic obstructive pulmonary disease, respiratory failure), comorbidities (arterial hypertension, diabetic mellitus type II, Cardiovascular pathology, Chronic renal failure, Ictus, Ischemic heart disease, Atrial fibrillation and Heart failure) and RX-features (Effusion, Consolidation, Edema, Infiltration, Lung Opacity).



	Anagrafica
Eta	81
Sesso	F

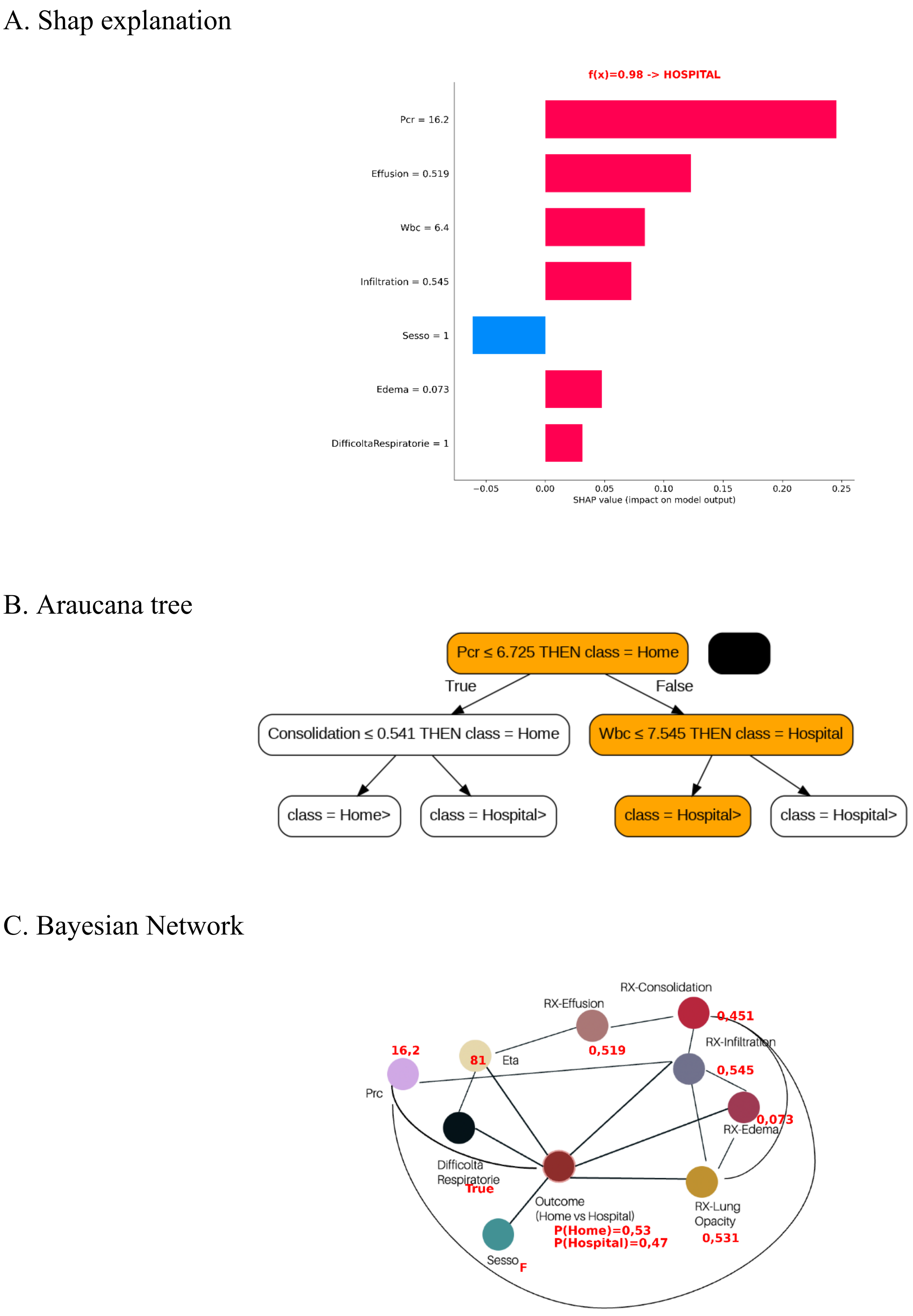
	Test di laboratorio
Pcr	16,2
Wbc	6,4

	Problemi respiratori
Tosse	False
DifficoltaRespiratorie	True
Bpco	False
InsufficienzaRespiratoria	False

	Comorbidità
IpertensioneArteriosa	True
DiabeteMellitoTipo2	False
PatologieCardiovascolari	False
InsufficienzaRenaleCronica	False
Ictus	False
CardiopatiaIschemica	False
FibrillazioneAtriale	False
ScompensoCardiaco	False
Demenza	False
CancroAttivoNegliUltimi5Anni	False

	RX
RX-Effusion	0,519
RX-Consolidation	0,451
RX-Edema	0,073
RX-Infiltration	0,545
RX-LungOpacity	0,531

Next, (2) the predicted class ('Home' or 'Hospital') was accompanied by three explanations (Shap, Araucana tree, and Bayesian network) (Figure 7). The order of the presentation of the three explanations was randomised in order to avoid the preferences for a method due to a particular order.



A. Shap explanation

B. Araucana tree

C. Bayesian Network

Figure 7. Example of the three explanations shown during the survey. They all represent the explanation of the classification of the same patient summarised in Table 6.

The study under consideration involves the analysis of real-world retrospective cases, which were presented in a random sequence. This methodological approach is consistent with the criteria established for Levels 4 and 5 within the evidence hierarchy relevant to empirical research in AI and XAI. According to (Famiglini et al., 2024), these levels are indicative of robust empirical validation and rigorous methodological standards. The alignment with these levels underscores the study's adherence to high standards of evidence-based research, thereby enhancing the credibility and applicability of its findings in the context of AI and XAI.

The graphical explanation was followed by two statements (Hoffman et al., 2018). The clinician had to assign the value on a Likert scale from 1-strongly disagree to 6-strongly agree for each explanation. The statements were chosen to represent two important usability measures – comprehensibility and helpfulness:

- The explanation is intuitively understandable (1-6)
  IT: La spiegazione è intuitivamente comprensibile

- The explanation helps me take an appropriate decision on the case at hand (1-6)
  IT: La spiegazione mi aiuta a prendere una decisione appropriata sul caso in questione

Comprehensibility is established in the machine learning research as the capacity of a learning algorithm to convey its acquired knowledge in a way that humans can easily grasp (Arrieta et al., 2019; Fernandez et al., 2019). It aligns with the usability principles of ease of learning and understandability. A system that is comprehensible allows users to quickly grasp how to use it, which is a key aspect of usability. This ensures that users can efficiently learn and navigate the system without extensive training or support. On the other hand, helpfulness fits with the principles of effectiveness and decision support. A supportive system provides users with the necessary information and tools to make informed decisions, thereby enhancing their overall experience. By ensuring that users have access to clear and helpful information, the system improves its usability and user satisfaction. It assists users in achieving their goals efficiently and effectively. Some synonyms of helpfulness include usable or useful. (*What Is Usability - The Ultimate Guide*, n.d.).

After the first explanation and Likert scale evaluation the remining two explanations were presented and their corresponding Likert evaluation was taken.

At the end of the case, (3) the clinicians were asked if they agree with the prediction and corresponding explanations and which explanation was the most useful in their opinion:

- Overall, do you agree with the class predicted for the Patient?
  (IT: Nel complesso, sei d'accordo con la classe predetta per il Paziente?)

- Overall, which type of explanation did you find most suitable and intuitive for this patient?
  (IT: Nel complesso, quale metodo di spiegazione hai trovato più adatto e intuitivo nel compito di classificazione del Paziente?)

Think-aloud protocol

Think-aloud is an established protocol in software engineering and cognitive psychology, however it is also very relevant in any case of decision-making. It has been used before in order to assess decision-making in healthcare, for example to compare clinician and nurse reasoning (Thompson et al., 2017) or decision-making at emergency departments (Gamborg et al., 2023; Press et al., 2015). Some recent reports of clinical XAI DSSs have already used think-aloud it for assessment (Anjara et al., 2023).

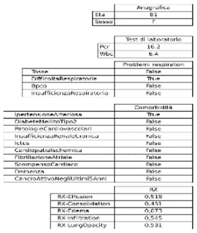
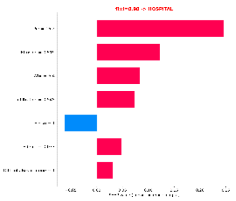
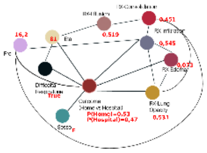
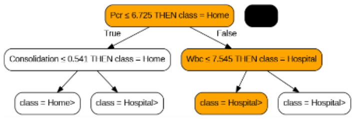
During the think-aloud session the user is supposed to verbalize all the thoughts, feelings and behaviours that occur while performing a task. The unstructured format allows for exploration and discovery of the issues and points that were not foreseen by the investigator. It allows for feedback that cannot be substituted by a questionnaire or an interview because the decision-making details and intricacies can be captured only while the decision is being made (Noushad et al., 2024). There is a vast literature on the biases of memory formation that prove that the retrospective account of events is largely distorted and influenced for example by additional information being gathered (Leighton, 2017).

Two clinicians took the survey while following the think-aloud protocol. They were taking the survey in their regular clinical setting. They were speaking their first language – Italian in order to assure easy flow of verbalizing and not introduce differences between clinicians which might have different levels of the English language (Noushad et al., 2024). They are referred to in the text as Clinician 1 (C1) belonging to the ER team and Clinician 2 (C2) working at the ID department. The whole session was recorded with written consent. At the beginning the clinicians were encouraged to verbalize every thought, feeling and impression that is on their mind while they take part in the survey. The protocol specified no interaction with the clinician while they were taking the survey unless they stop verbalizing their thoughts for around 10 seconds.

After data collection the recording allowed for additional time measurement and segmentation. Therefore, time spent on each patient and explanation was measured and compared with the questionnaire answers (Figure 8).

The recording was transcribed with the use of advanced online translator - TurboScribe with timestamps which allowed for further analysis of the sentiments. The transcript was proof-read and translated to English with the Helsinki-NLP (Natural Language Processing) translation model, available on Hugging Face (Helsinki-NLP/opus-mt-it-en), which is a state-of-the-art tool designed for translating text from Italian to English. It leverages advanced neural machine translation techniques to provide accurate and contextually appropriate translations

Screen views:

Patient	Explanation 1	Explanation 2	Explanation 3
			
1	Shap	Bayesian network	Araucana

	Anagrafica
Eta	81
Sesso	F

	Test di laboratorio
Pcr	16.2
Wbc	6.4

	Problemi respiratori
Tosse	False
DifficoltaRespiratorie	True
Bpco	False
InsufficienzaRespiratoria	False

	Comorbidità
IpertensioneArteriosa	True
DiabeteMellitoTipo2	False
PatologieCardiovascolari	False
InsufficienzaRenaleCronica	False
Ictus	False
CardiopatiaIschemica	False
FibrillazioneAtriale	False
ScompensoCardiaco	False
Demenza	False
CancroAttivoNegliUltimi5Anni	False

	RX
RX-Effusion	0,519
RX-Consolidation	0,451
RX-Edema	0,073
RX-Infiltration	0,545
RX-LungOpacity	0,531

Pcr ≤ 6.725 THEN class = Home
True
False
Consolidation ≤ 0.541 THEN class = Home
Wbc ≤ 7.545 THEN class = Hospital
class = Home>
class = Hospital>
class = Hospital>
class = Hospital>

Example timeline:

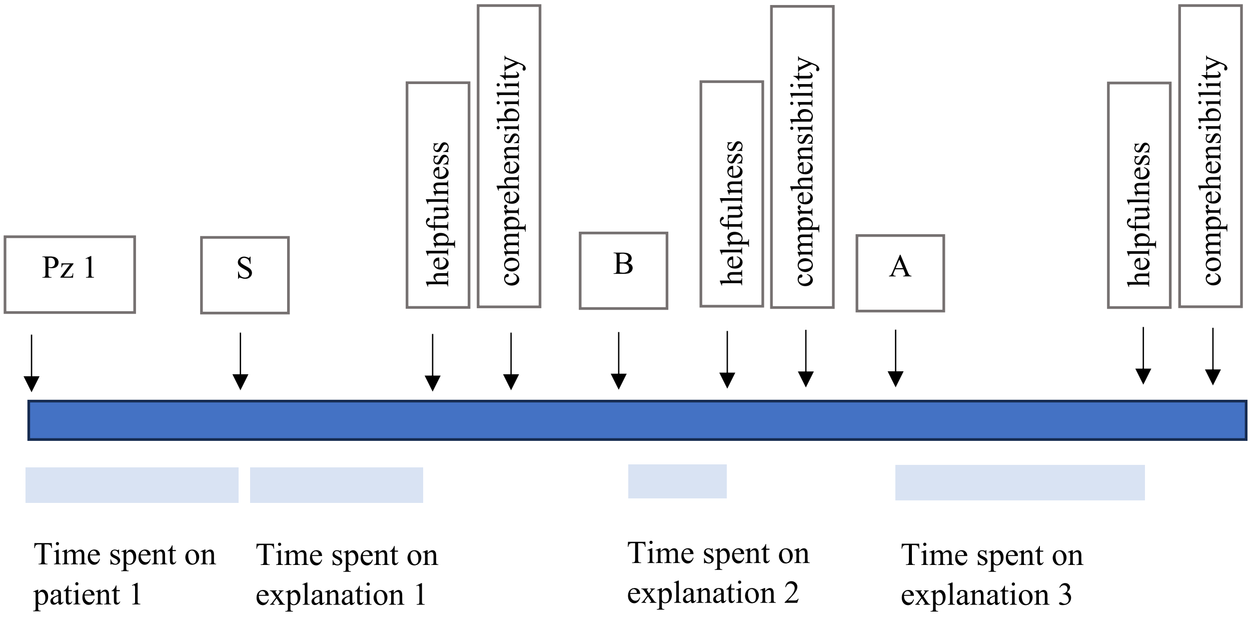


Figure 8. Example of time measurement during the survey – Patient 1 (Pz 1). Time spent on the patient starts at the initial presentation of the patient and stops as the first explanation is presented. Time spent on the explanation starts with the initial presentation of the explanation and finishes as the clinician starts to evaluate the method in terms of helpfulness.

Sentiment Analysis models

The transcript was annotated with the utterances related to four groups – Araucana, Bayesian network, Shap or neutral (description of the patients and technical or general comments). Due to scarcity of pre-trained sentiment models for the medical field, three established sentiment models were used to evaluate if they could be useful in the context of think-aloud in the medical setting. The models were not fine-tuned because of relatively small and diverse amount of data points from the think-aloud sessions.

Three models were utilized and compared – Vader, Roberta and Bert. Vader and Roberta are well established for analysis in the English language while Bert can return the sentiment score for multiple languages including Italian.

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool trained on Tweeter database. It is open-sourced under the MIT License (VaderSentiment 3.3.1 Documentation). It outputs positive, negative, neutral, and compound sentiment values, with the overall sentiment determined by the compound score. The calculation of the compound score involves several key steps. Each word in the sentiment lexicon is assigned scores for positive, negative, and neutral sentiments, ranging from -4 (most negative) to 4 (most positive). Heuristic rules are applied to account for punctuation, capitalization, degree modifiers, contrastive conjunctions, and negations, which adjust the compound score of a sentence. The scores of all words in the text are then standardized to a range of -1 to 1.

BERT (Bidirectional Encoder Representations from Transformers) is an NLP model developed by Google AI (Devlin et al., 2019). Unlike traditional models, BERT processes text bidirectionally, considering context from both directions. This approach allows BERT to understand the nuanced meaning of words based on their surrounding context. Pre-trained on a vast corpus, including Wikipedia and the Toronto Book Corpus, BERT uses Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) to learn language patterns. Its versatility allows it to be fine-tuned for various NLP tasks, such as question answering and sentiment analysis, with minimal modifications. This adaptability has made BERT a benchmark for performance in multiple NLP applications. The multilingual version of Bert-base-multilingual-uncased-sentiment (nlptown/bert-base-multilingual-uncased-sentiment · Hugging Face) is a model fine-tune for sentiment analysis trained on reviews training data. It is available in 6 languages including Italian.

This model processes text input by tokenizing it and passing it through BERT model to generate contextualized embeddings. These embeddings are then fed into a classification layer that outputs a sentiment rating from 1 to 5 stars based on the highest probability class. The output can then be converted to the [-1, 1] range to be more easily compared and visualized. In the following section using the multilingual version of BERT will be abbreviated as Bert.

RoBERTa (Robustly Optimized BERT Approach) here referred to as Roberta is a transformer-based model developed by Facebook AI to enhance BERT. It improves upon BERT by removing the Next Sentence Prediction (NSP) objective, training on a larger corpus with more data, and using a dynamic masking pattern during training to better understand language context. The specific version of Roberta used in this work (Roberta-base – cardiffnlp/twitter-roberta-base-sentiment) is trained on ~58M tweets and finetuned for sentiment analysis with the TweetEval benchmark (Barbieri et al., 2020). Roberta employs the same architecture as BERT, with multiple layers of self-attention and feed-forward neural networks, and is bidirectional, considering context from both sides of the text. These enhancements result in superior performance across various NLP tasks (Liu et al., 2019; RoBERTa). The model first processes the input text and outputs raw scores for each sentiment category (Negative, Neutral, Positive). These scores are then converted into probabilities using the softmax function, which normalizes them so that the sum of all probabilities equals 1, indicating the likelihood of each sentiment.

The decision to use multiple sentiment analysis models, such as Vader, Roberta, and Bert, stems from the unique characteristics and methodologies each model employs. Due to these differences, the output values from these models are not directly comparable. Instead of focusing on the individual sentiment scores, the goal is to identify overarching patterns across the models. This approach allows for a more comprehensive and nuanced understanding of the emotional tone, leveraging the strengths of each model to provide a balanced analysis. Even though the direct comparison is not straightforward this approach seems to be the latest trend and can provide a more holistic overview of the emotional tone (Qi & Shabrina, 2023).

The decision to use sentiment analysis models that focus on the English language instead of Italian was driven by several factors. Firstly, English is the most widely studied language in the field of NLP, resulting in a greater availability of pre-trained models and annotated datasets (Zhu et al., 2024). The extensive research and resource availability enhance the reliability and performance of

English-centric models (Wankhade et al., 2022). Additionally, while there are emerging models for other languages, such as Italian, they often lag behind in terms of accuracy and robustness due to limited training data and resources (Catelli et al., 2022). Therefore, leveraging well-established English models like Vader, Robertaa, and Bert ensures more accurate and consistent analysis of sentiments. A recent study in Nature shows that leveraging sentiment analysis in languages such as Arabic, Chinese, French and Italian through translation to English first achieves high accuracy and better results than independent pre-trained models (Miah et al., 2024). Model comparisons reveal superior capabilities of cross-lingual sentiment analysis across a variety of pre-trained language models (Zhu et al., 2024).

Analyses

After the data collection the responses were anonymised and analysed. The comparisons of the measured parameters were made with the focus on aforementioned research questions (Objectives and hypotheses):

RQ1: General Perception: What are the perceptions of clinicians regarding the use of human-AI collaboration tools in a medical setting?

General perceptions were directly related to proportions of answers to the initial questionnaire estimating familiarity vs unfamiliarity with AI tools and skepticism vs positive attitude towards AI in healthcare.

RQ2: Compliance: What are the compliance rates and how do they compare to established frameworks?

In this study, compliance refers to the extent to which healthcare professionals adhere to the recommendations provided by AI DSS. Compliance was measured by the frequency of following AI recommendations. Several factors influencing compliance are known to influence compliance

(Choudhury, 2022; Dlugatch et al., 2024). In this study, years of experience, attitudes towards AI in healthcare and time taken to complete the survey were compared across all clinicians.

RQ3: Method Comparison: How do clinicians' perceptions of comprehensibility, helpfulness and cognitive load differ among the three XAI methods (Shap, Araucana tree and Bayesian network)? Do they align with directly expressed method preferences?

Multiple comparisons of comprehensibility and helpfulness were made taking into account individual answers and prediction classes (TP, TN, FP, FN). Relevant comparisons were assessed statistically with Wilcoxon signed-rank test. To evaluate cognitive load, mean time spent on each explanation for two clinicians taking part in the think-aloud was compared with pairwise Wilcoxon signed-rank test.

RQ4: Sentiment Analysis (emotional tone): What is the emotional tone of clinicians' attitudes towards each XAI explanation assessed through sentiment analysis?

The exploratory comparison was made to detect patterns across chosen sentiment analysis models. The outputs of the models were visualized for each utterance form the think-aloud protocol as line plots to track the utterances and sentiments scores through time and as box-plots for aggregate comparisons. The models were introduced and described in the previous section (Think-aloud protocol). Statistical tests were not applied due to a small sample sizes and need for further fine-tuning of the analysis.

RQ5: Department Comparison: Are there any differences in terms of method perceptions and preferences between ER and ID departments on average?

Comparisons of comprehensibility and helpfulness were made taking into account prediction classes (TP, TN, FP, FN). The results were compared with Wilcoxon signed-rank test.

RQ6: Pair Comparison: Are there any similarities or consistent patterns of comprehensibility and helpfulness ratings within allocated ER - ID clinician pairs?

In order to compare the responses within the pairs, the helpfulness and comprehensibility ratings were compared with weighted Cohen's kappa (Landis & Koch, 1977). It is a measure of inter-rater agreement for categorical data, which includes ordinal scales such as the Likert scale used in this study.

RQ7 and **RQ8** focusing on more theoretical aspects were addressed by reflection of different steps of the analysis and study design (RQ7) and literature comparison (RQ8).

Results

Survey

Initial questionnaire

All clinicians completed the initial questionnaire with the following proportions of ‘yes’ and ‘no’ answers –Table 7.

Table 7. The answers of the clinicians to the initial questionnaire.

ID	YEARS OF EXPERIENCE	TEAM	SEX	1	2	3	4	5
1	2	ID	F	No	No	Yes	Yes	Yes
2	10	ER	M	No	No	Yes	Yes	Yes
3	5	ID	F	No	Yes	Yes	Yes	Yes
4	30	ID	M	No	No	Yes	Yes	Yes
5	1	ID	M	No	No	Yes	Yes	Yes
6	15	ID	F	No	No	Yes	Yes	No
7	24	ER	M	Yes	No	Yes	Yes	Yes
8	2	ER	M	No	No	No	Yes	No
9	6	ER	M	Yes	No	Yes	Yes	Yes
10	30	ER	M	No	No	Yes	Yes	Yes

Based on the questions and answers, the general attitudes of clinicians towards AI in healthcare were identified: being familiar or unfamiliar with AI and being skeptic or positive towards AI.

Familiarity was determined by answering ‘yes’ to either question 1 or 2. Three clinicians turned out to have some familiarity with AI answering yes to at least one of those questions (Table 8). Perceived usefulness or skepticism was determined by ‘no’ answers to questions 3 or 5. Two clinicians out of ten turned out to be skeptic towards AI in healthcare. One answering ‘no’ to two and one to one question (Table 8). Answering ‘yes’ to these questions was referred to having positive attitude towards AI in healthcare.

The rest of the clinicians (5) were neither familiar nor skeptic towards AI and are classified as unfamiliar with AI but having positive attitude towards it.

Table 8. Proportions of answers to the initial questionnaire measuring general attitudes towards AI in healthcare.

NO.	QUESTION	YES	NO
1	I have good knowledge in the field of Artificial Intelligence.	2	8
2	I have been in contact with and/or used Artificial Intelligence systems in my work.	1	9
3	I am convinced that Artificial Intelligence can help me answer questions more correctly and quickly when I am uncertain.	9	1
4	I am convinced that being assisted by Artificial Intelligence (e.g., a virtual assistant) in my work or study can increase my productivity.	10	0
5	I am convinced that being assisted by Artificial Intelligence can improve the effectiveness of what I do.	8	2

Compliance

The survey was completed by all the clinicians in the course of two months.

In general, the compliance (agreement with the prediction) was ranging from 100% to 70% (while half of the cases were true positive (TP) and true negatives (TN) and half false positives (FP) and false negatives (FN) – 50:50 ratio. One clinician showed compliance of 70%, four of 80%, three 90% and two 100% (Figure 9). The average compliance was equal to 86%.

Comparing how years of experience influence compliance, there is a general trend of increased compliance with the years of experience (Figure 10). This trend is the same for male and female clinicians (Figure 11 (left)). The relationship of increased compliance with greater years of experience is strong for the ID department and has the slope close to 0 for the ER department (Figure 11 (right)).

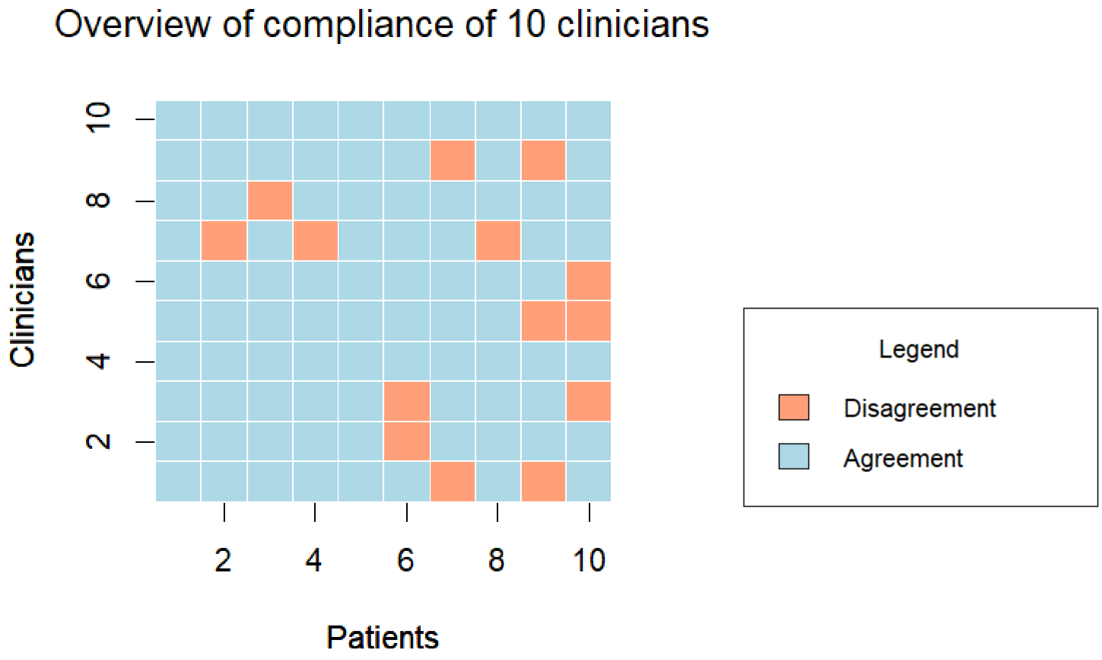


Figure 9. Overview of compliance, agreement with the prediction, of all clinicians. Average compliance was equal to 86%. Starting from patient 6 the cases transition from TP and TN to FP and FN.

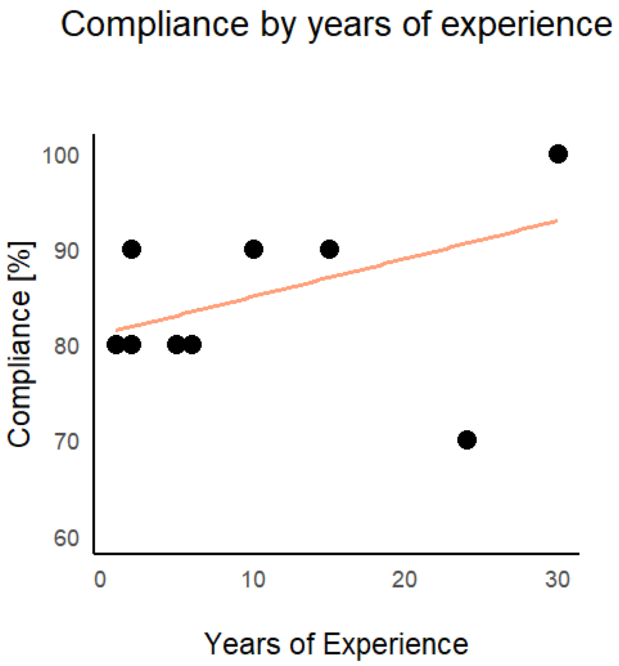


Figure 10 Compliance compared to years of experience for all clinicians.

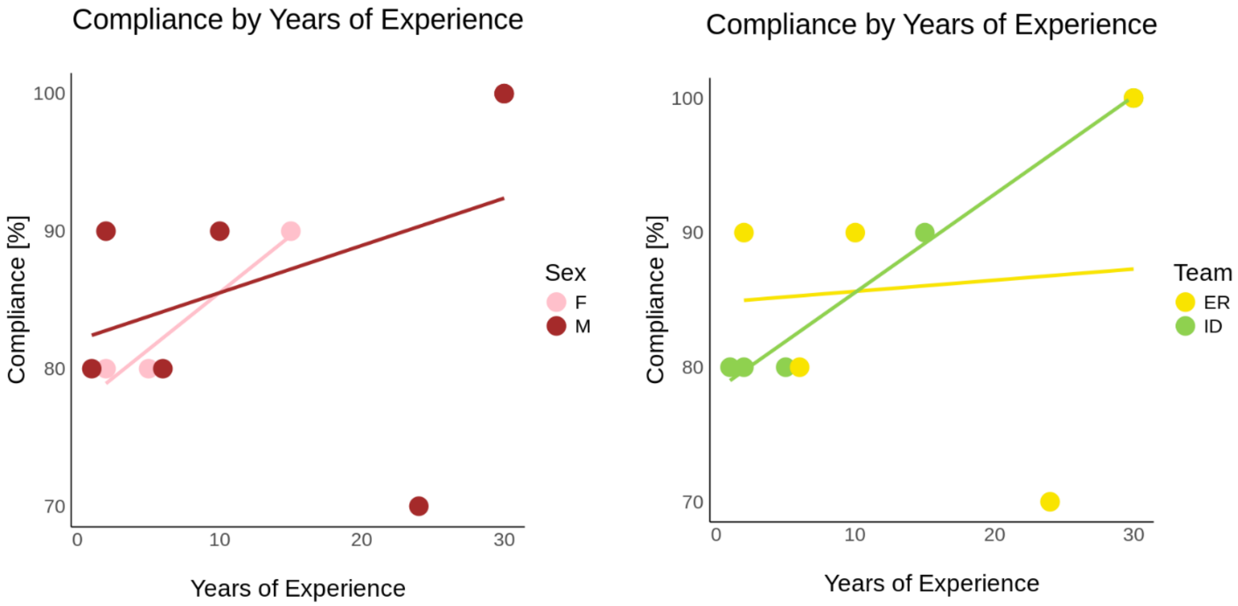


Figure 11 Comparison of compliance for male and female clinicians (left) and comparison of compliance for two departments – ER – Emergency Room and ID – Infectious Diseases (right).

Time taken to complete the survey was in the range from 6 to 85 minutes with the average time of 28.8 minutes. There is a negative relationship between years of experience and time take to complete the survey and slight positive relationship between compliance and time taken to complete the survey (Figure 12).

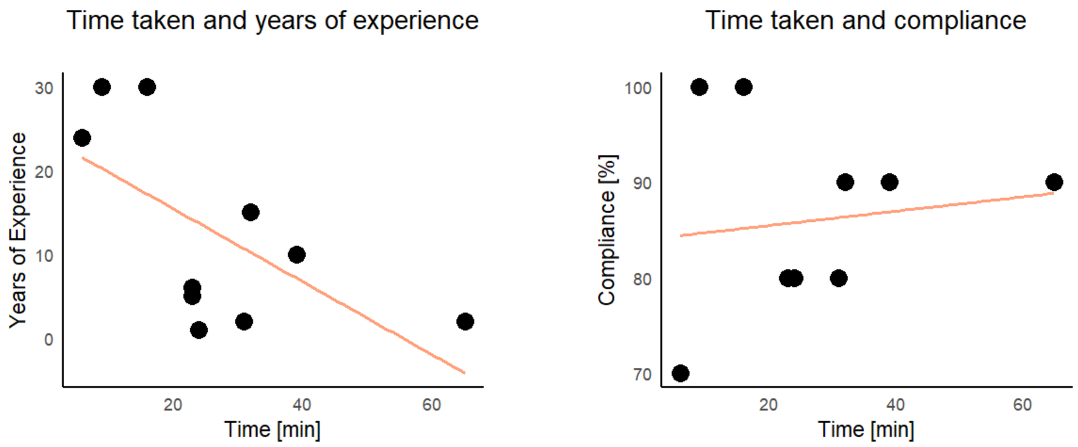


Figure 12 Scatterplot showing tested relationships between time taken to complete the survey and years of experience (left) or compliance (right).

The comparisons of attitudes were made with years of experience and compliance of the clinicians. In general, clinicians familiar with AI showed less compliance and had fewer years of experience (two out of three) (green samples in Figure 13). Being skeptic was not associated with different compliance rates or with years of experience (blue samples in Figure 14). Having positive attitude towards AI and years of experience together were associated with increased compliance (Figure 14).

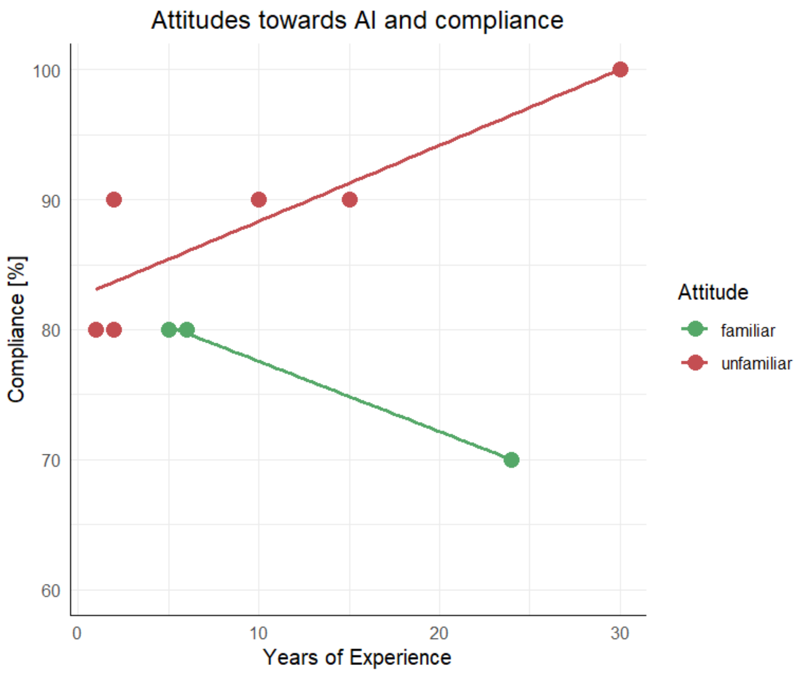


Figure 13. Comparing how general familiarity with AI of surveyed clinicians, measured through the initial questionnaire, relates to compliance and years of experience.

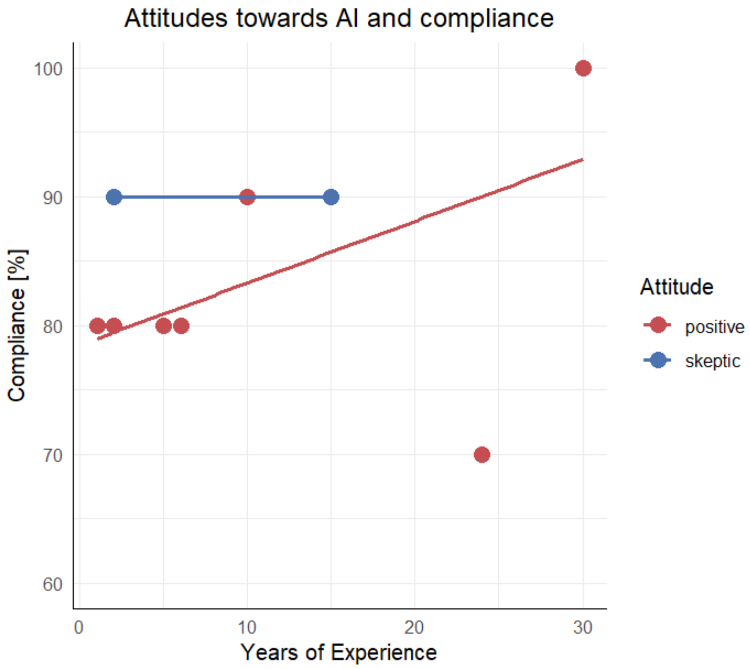


Figure 14. Comparing how general attitude towards AI of surveyed clinicians, measured through the initial questionnaire, relates to compliance and years of experience.

Comprehensibility and helpfulness

In order to identify the strategy for the analysis of Likert scales of helpfulness and comprehensibility, individual ratings over the course of the whole survey were visualised (Figure 15; Figure 16; Figure 17; Figure 18). It allowed for an overview of the answers within pairs and between departments. The overall helpfulness and comprehensibility scores showed approximately normal distributions (Figure 19).

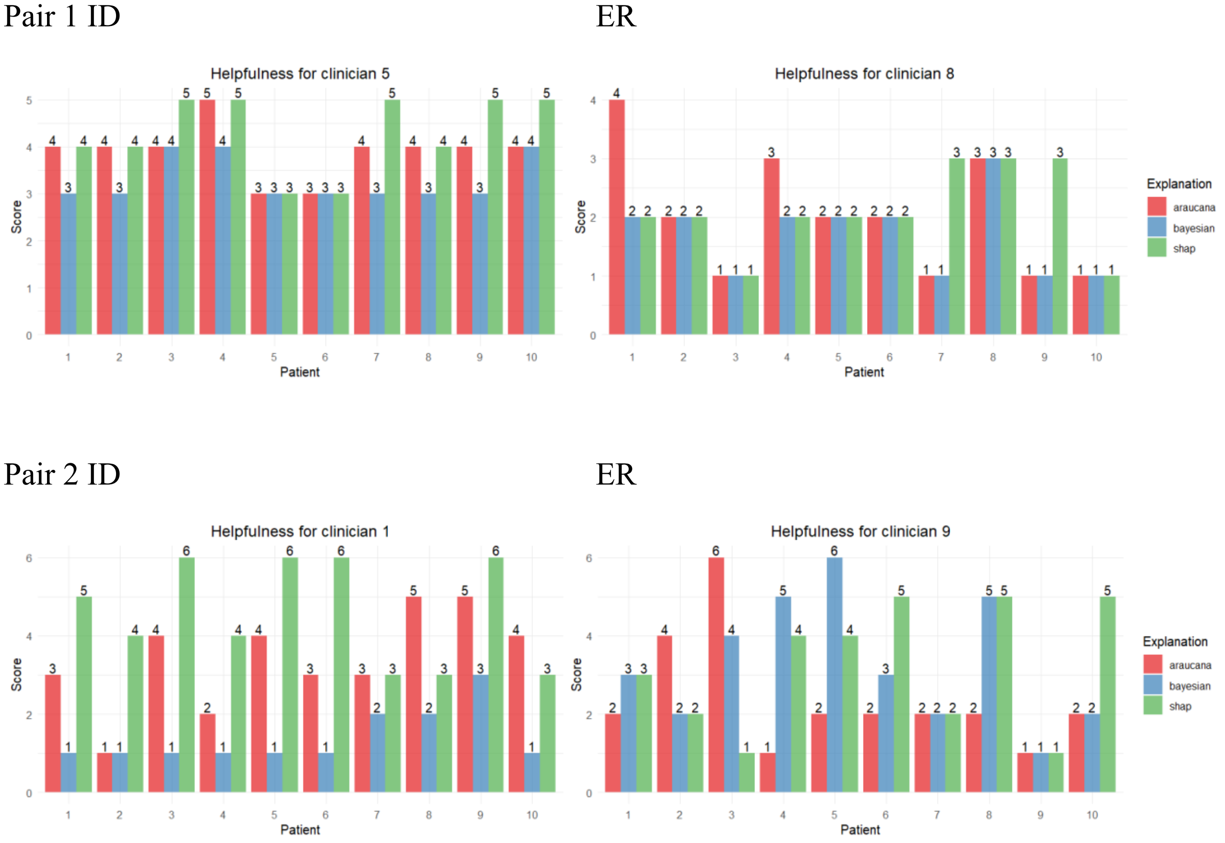
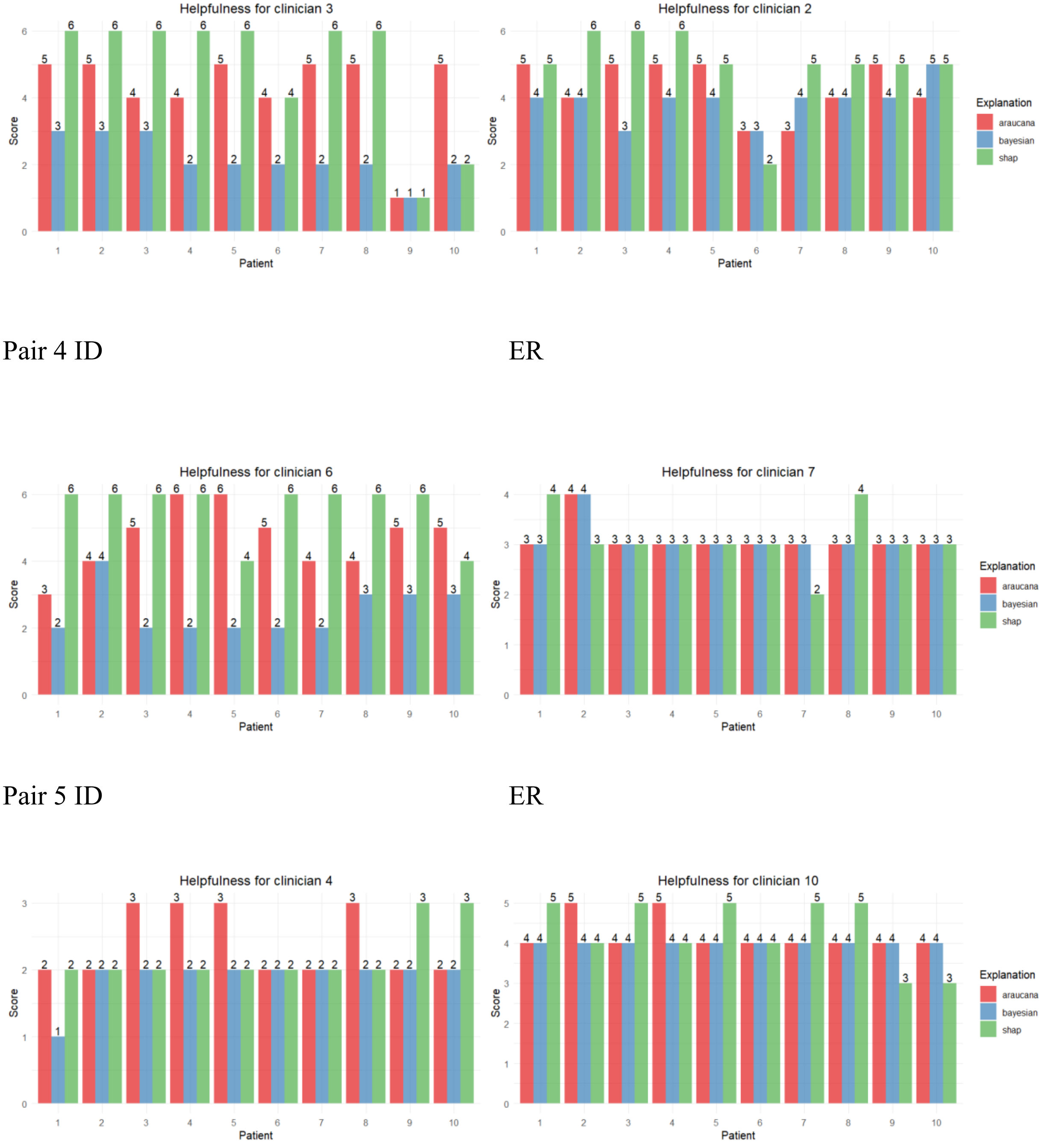


Figure 15. The comparison of helpfulness scores of the explanations (Araucana, Bayesian network and Shap) for all patients for the three first pairs of clinicians – 1 and 2.

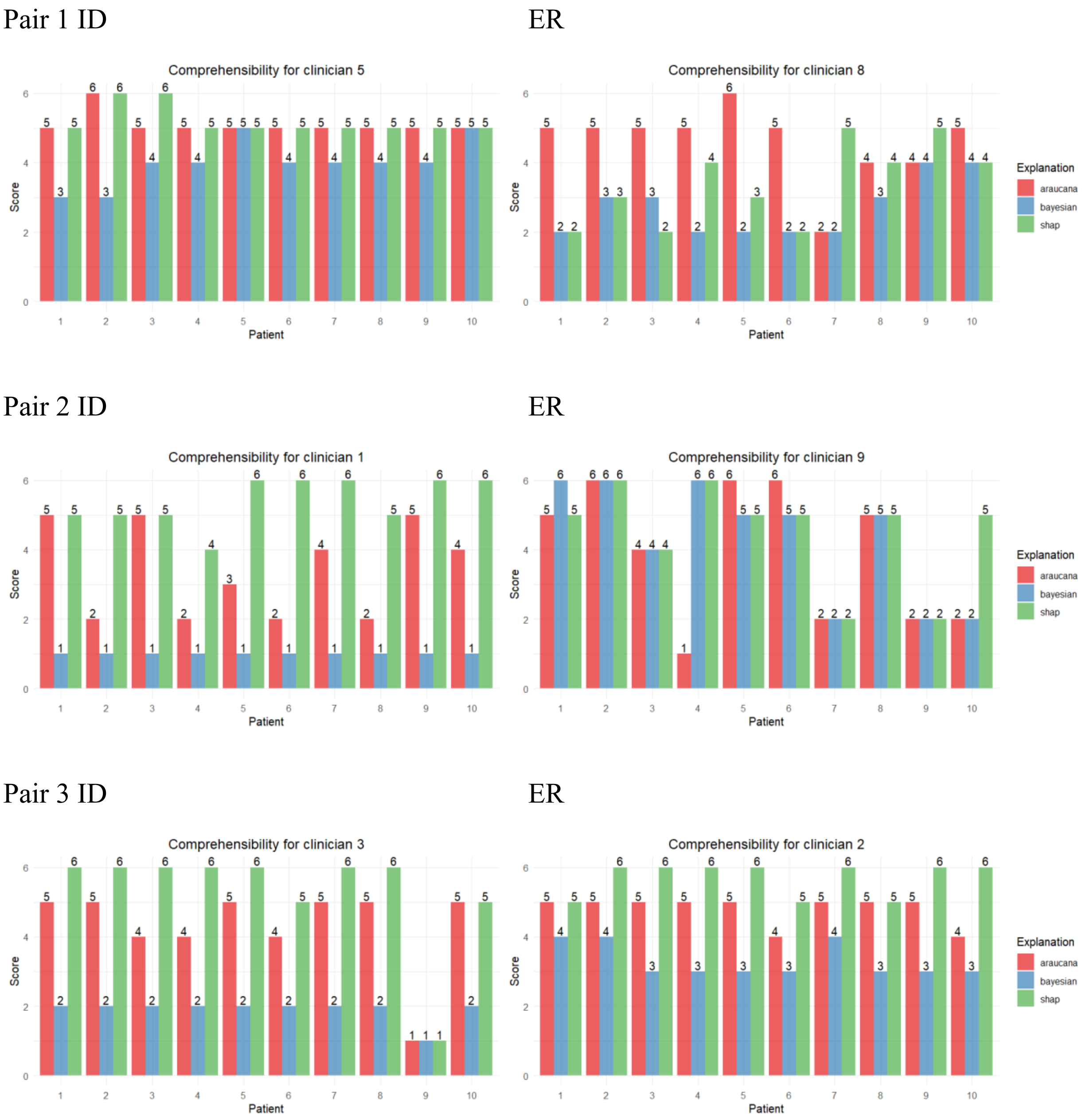
Pair 3 ID ER



Pair 4 ID ER

Pair 5 ID ER

Figure 16. The comparison of helpfulness scores of the explanations (Araucana, Bayesian network and Shap) for all patients for the clinicians pairs 3, 4 and 5.



Pair 1 ID

ER

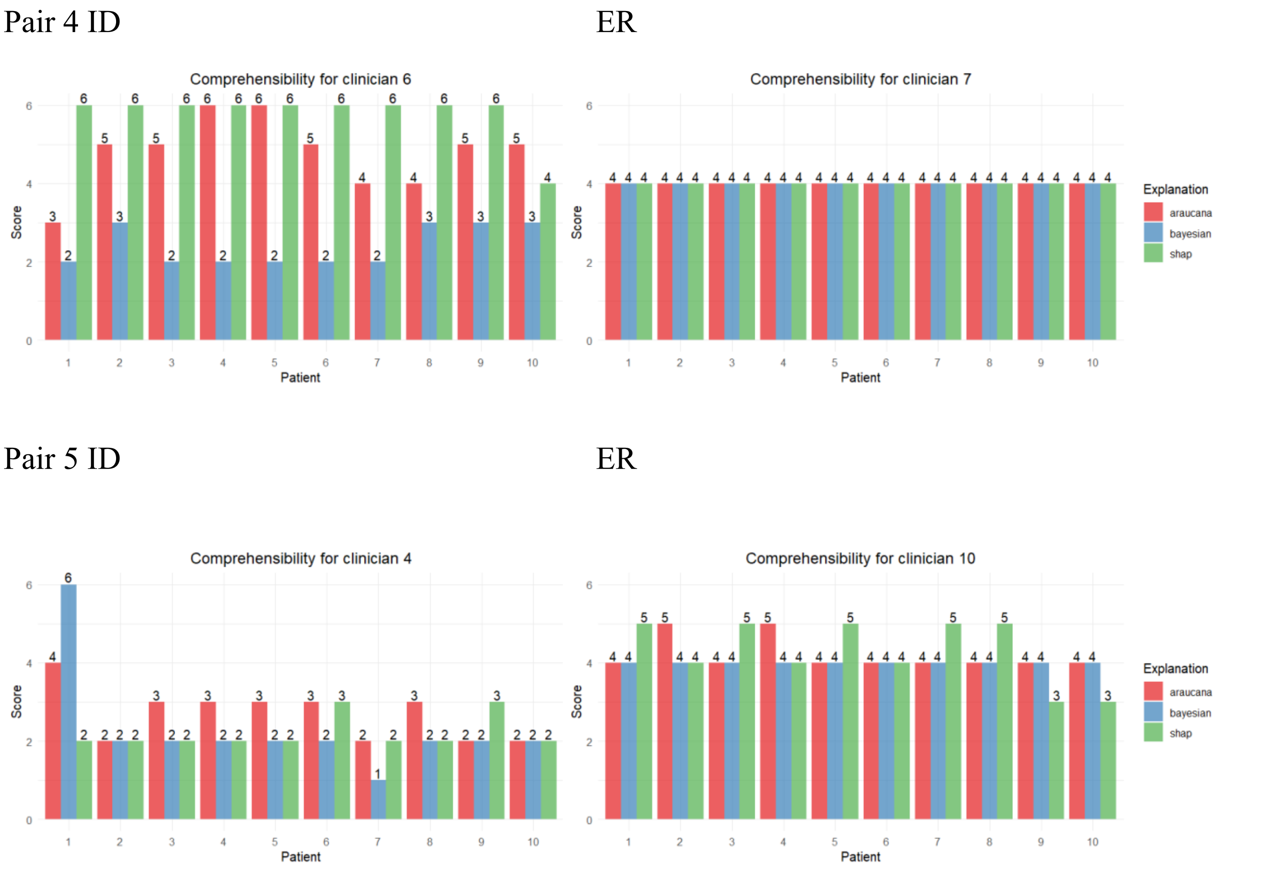
Pair 2 ID

ER

Pair 3 ID

ER

Figure 17. The comparison of comprehensibility scores of the explanations (Araucana, Bayesian network and Shap) for all patients for the three first pairs of clinicians – 1, 2 and 3.



Pair 4 ID

ER

Pair 5 ID

ER

Figure 18. The comparison of comprehensibility scores of the explanations (Araucana, Bayesian network and Shap) for all patients for the clinicians pairs 4 and 5.

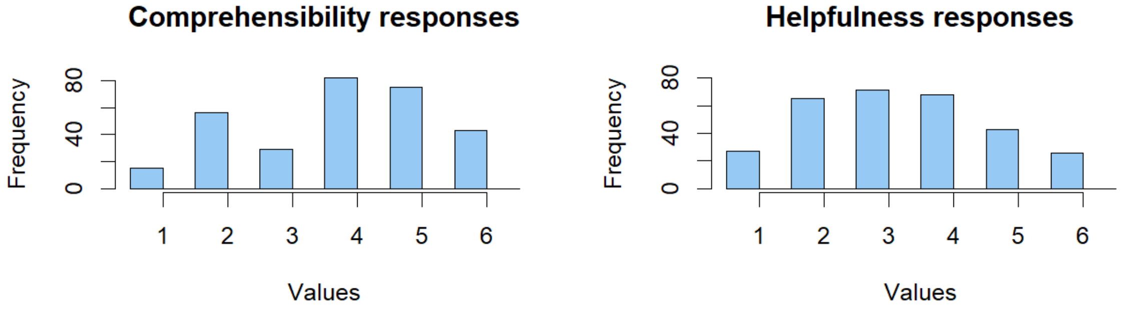


Figure 19. Distributions of all the scores for all the explanations for comprehensibility (left) and helpfulness (right).

Densities of the scores compared individually for each method show that Shap and Araucana are shifted towards more positive scores compared to the Bayesian network (Figure 20). Indeed, the comparisons of the distributions show that both for helpfulness and comprehensibility Shap achieved the highest scores, Bayesian network the lowest and Araucana with the scores in the middle (Figure 21). The Wilcoxon signed-rank test of the scores shows that all the comparisons are statistically significant (Figure 21).

Splitting helpfulness and comprehensibility scores by department shows that for the ID department (Figure 22) the differences between the three explanations are large for both usability measures; with Shap achieving the highest scores, followed by Araucana, and Bayesian network. For the ER department (Figure 23), the differences are small for comprehensibility, with no differences between Araucana and Shap, and Bayesian network achieving lower scores. While for helpfulness there are no differences between the three explanations.

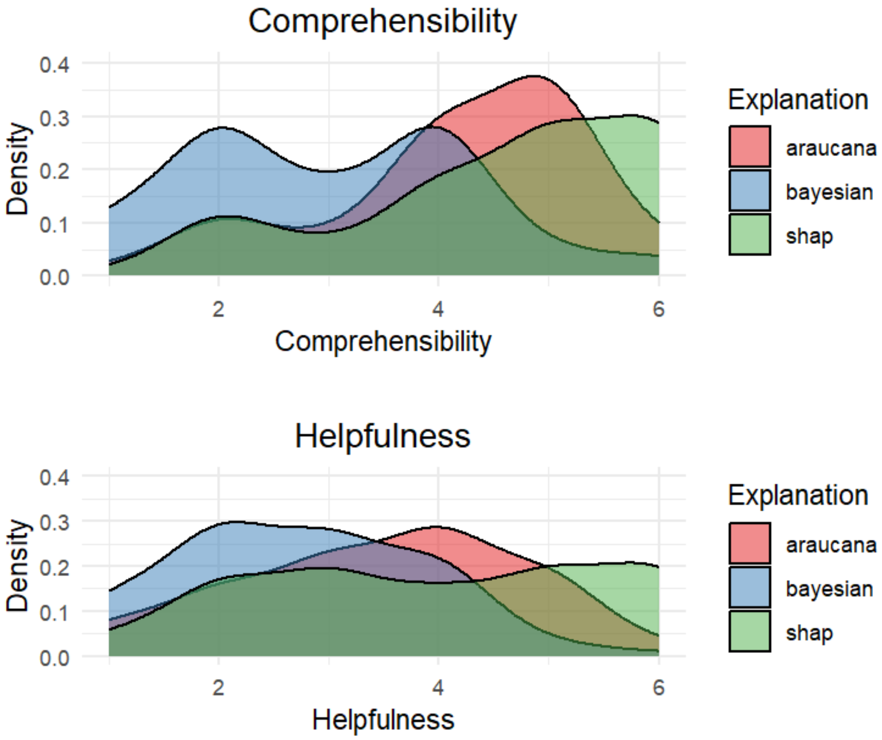


Figure 20. Densities of the scores of 10 clinicians for comprehensibility and helpfulness for three explanations: Araucana, Bayesian network and Shap.

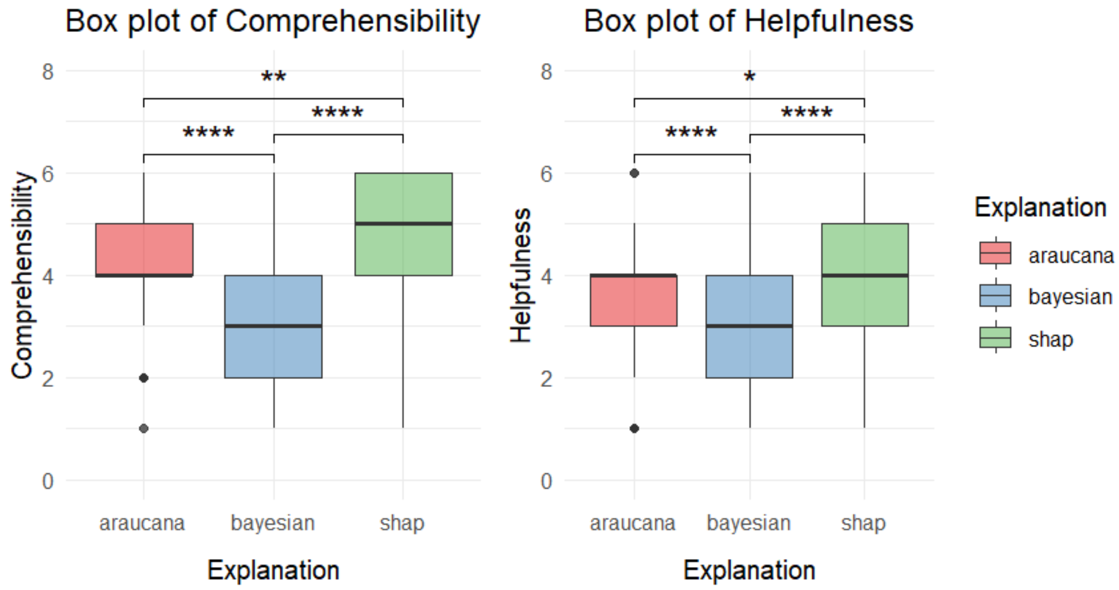


*Figure 21. Box plots of the scores of 10 clinicians for comprehensibility (left) and helpfulness (right) for three explanations: Araucana, Bayesian network and Shap. Significance calculated with pairwise Wilcoxon test. * - p < 0.5, ** - p < 0.01, *** - p < 0.001, **** - p < 0.0001.*

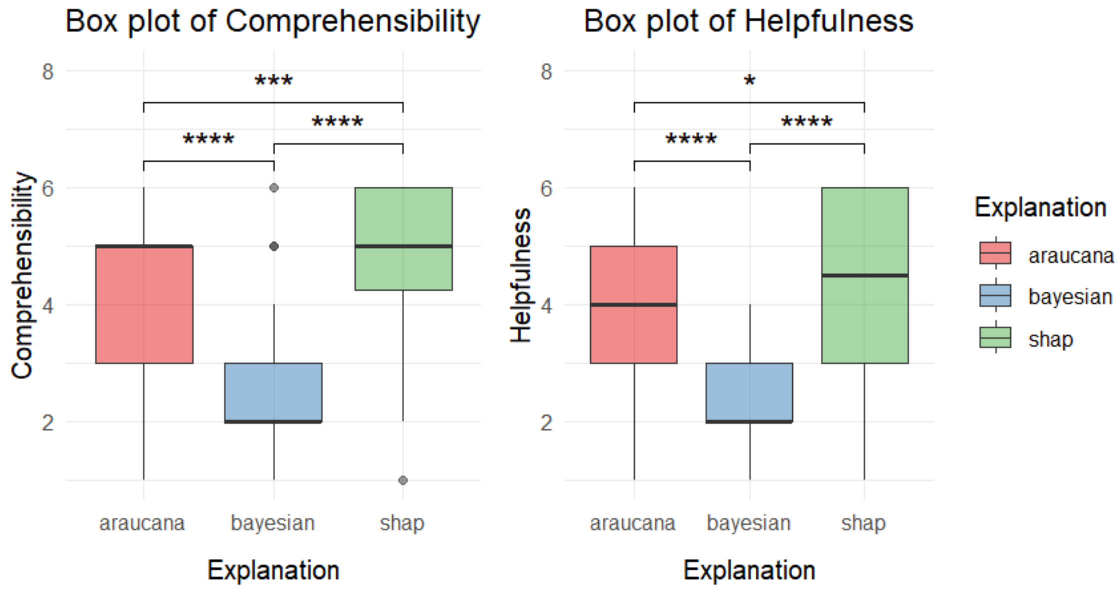


*Figure 22. Box plots of the scores of 5 clinicians from ID department for comprehensibility and helpfulness for three explanations: Araucana, Bayesian network and Shap. Significance calculated with pairwise Wilcoxon test. * - p < 0.5, ** - p < 0.01, *** - p < 0.001, **** - p < 0.0001.*

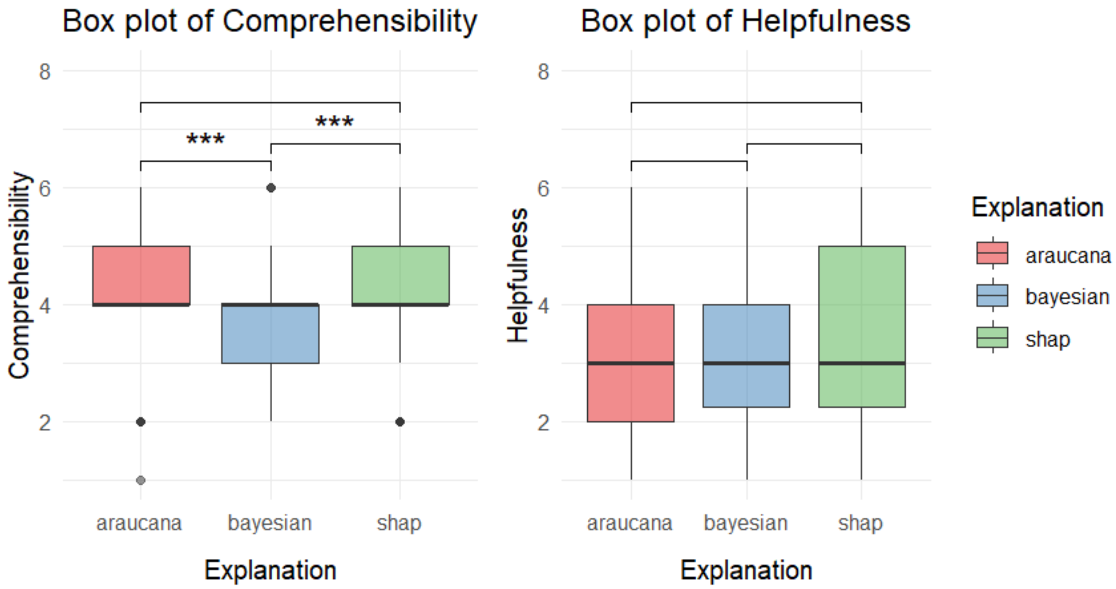


*Figure 23. Box plots of the scores of 5 clinicians from ER department for comprehensibility and helpfulness for three explanations: Araucana, Bayesian network and Shap. Significance calculated with pairwise Wilcoxon test. * - p < 0.5, ** - p < 0.01, *** - p < 0.001, **** - p < 0.0001.*

The choice of favourite explanation (Figure 24) for each patient was also different for the clinicians of the two departments. The clinicians from the ER department chose Shap and Araucana with almost the same frequency counts and Bayesian network around one in five cases. Clinicians from the ID department largely chose Shap as their favourite method followed by Araucana chosen around 25% of the time and Bayesian network with negligible count score.

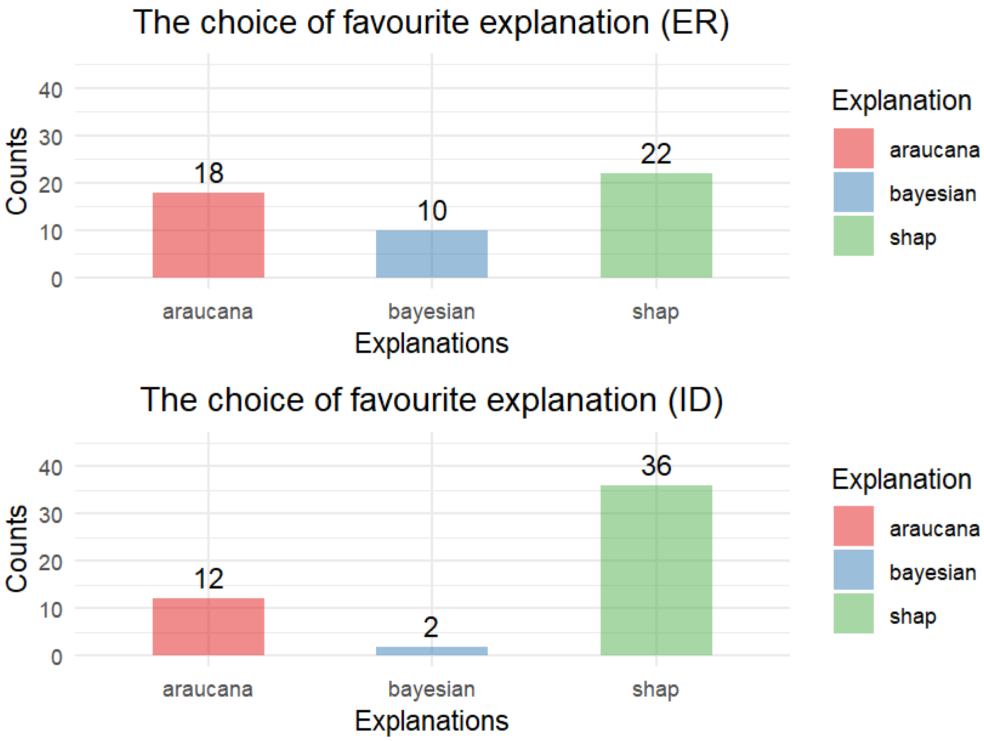


Figure 24. Expressed explanation preferences according to the department of each clinician – 5 from Emergency room (ER) (top) and 5 from Infectious diseases (ID) (bottom). The clinicians were asked directly to express explanation preference for each case out of 10 and corresponding explanations. It resulted in 50 expressed preferences for each department in total.

The comparisons of the scores dividing them into four classes – TP, TN, FP and FN were also visualized for three explanations. The comprehensibility scores for the four classes (Figure 25) are statistically significant comparing Shap values and Araucana to the Bayesian network. On the other hand, the comparisons between Araucana and Shap are less clear and significant only for FP cases. For the helpfulness scores (Figure 26) the comparisons are generally significant for all the classes comparing Bayesian network with Araucana and Shap. The least significant differences between the methods are found in the FN case. The comparisons between Araucana and Shap are significant only for FP cases. For other classes the medians and distributions are very similar for two methods.

Multiple other comparisons were made such as the comparison of few cases in which the clinicians disagreed with the predictions (Figure 27). Out of 14 cases in which the clinicians disagreed: 3 were TN, 3 were FP and 8 were FN. The distribution of scores of correct disagreements with FN predictions was visualized in order to show a different pattern than the overall score distributions (Figure 25; Figure 26).

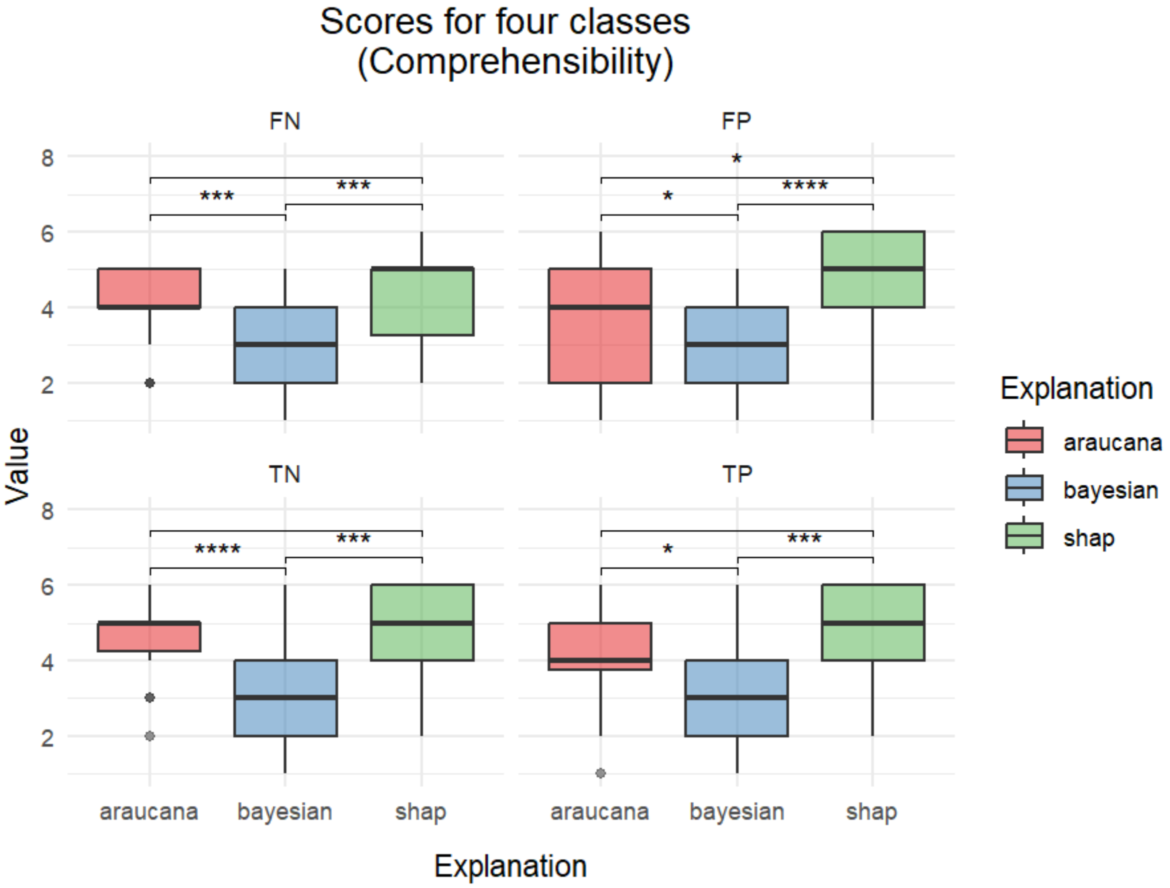


*Figure 25. The comparison of the comprehensibility scores for each explanation split by the class of the patient: FN (false negative), FP (false positive), TN (true negative), TP (true positive). Significance calculated with pairwise Wilcoxon test. * - p < 0.5, ** - p < 0.01, *** - p < 0.001, **** - p < 0.0001.*

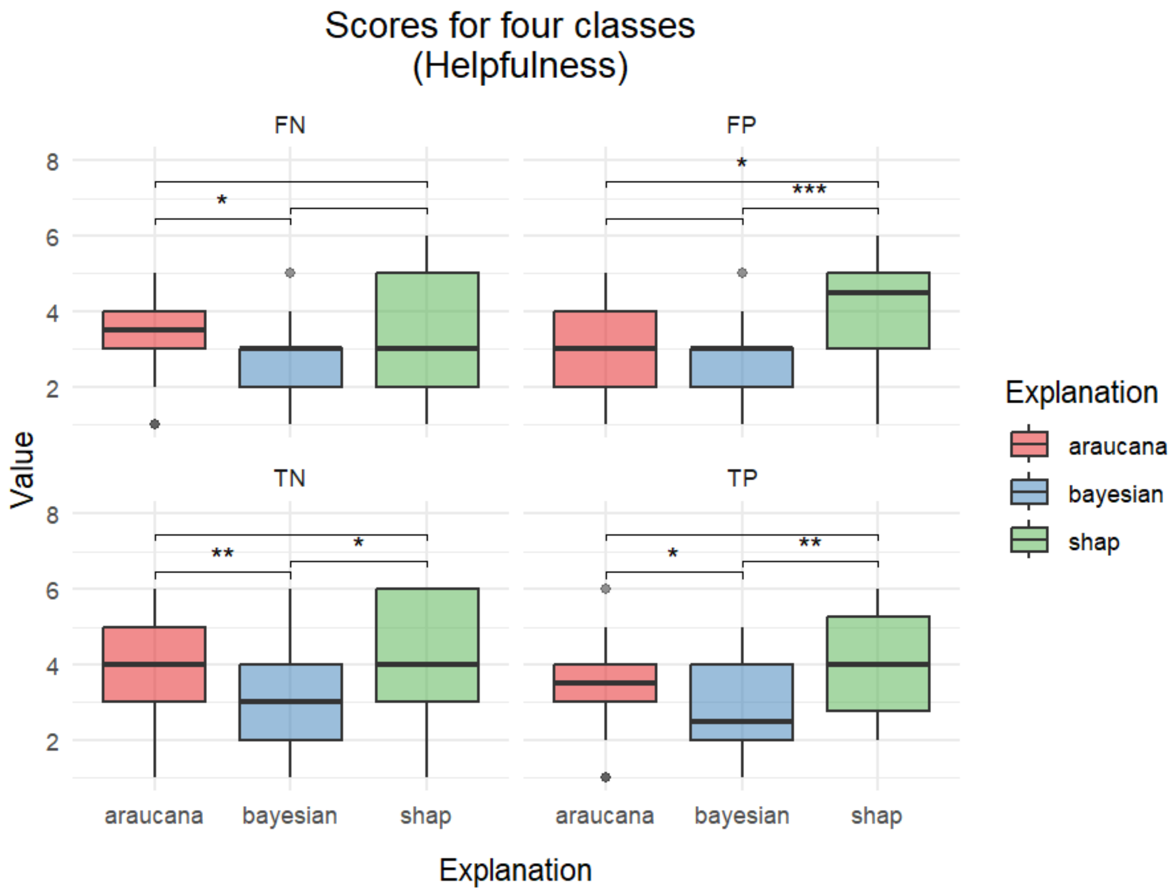


*Figure 26. The comparison of the helpfulness scores for each explanation split by the class of the patient: FN (false negative), FP (false positive), TN (true negative), TP (true positive). Significance calculated with pairwise Wilcoxon test. * - p < 0.5, ** - p < 0.01, *** - p < 0.001, **** - p < 0.0001.*

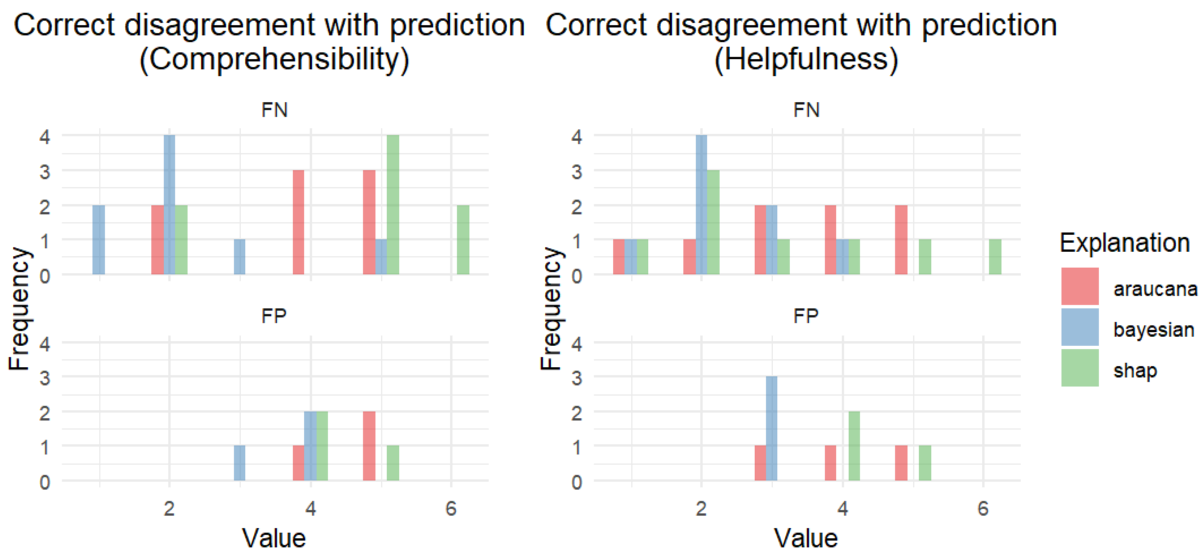


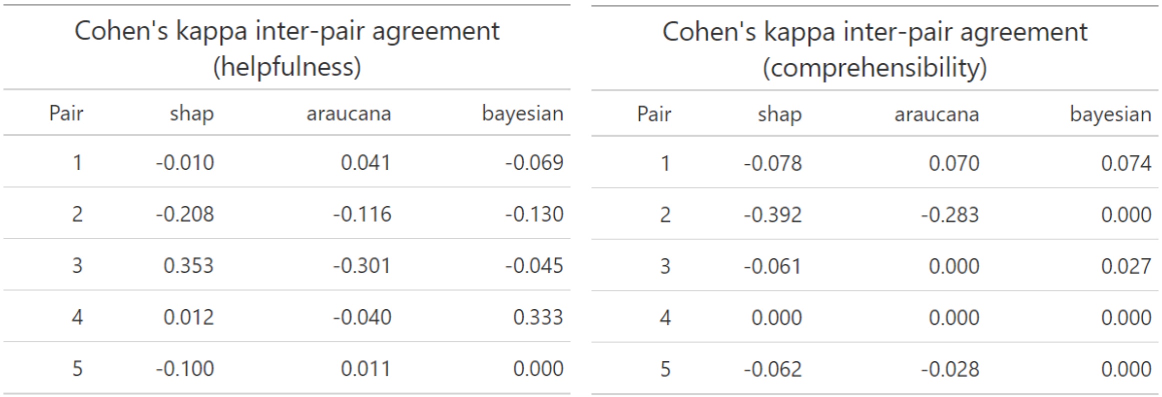
Figure 27. Histogram of scores for few cases that clinicians disagreed with and correctly classified as wrong prediction. FN (false negative) – 8 patients and FP (false positive) – 3 patients.

Within-pair comparison

In order to systematically test the similarity of helpfulness and comprehensibility scores, weighted Cohen’s kappa was calculated. The table of pairs with clinician indices is listed in the methods section (Table 4). First, the scores for each explanation were calculated pairwise within each pair (Table 9). The scores were compared with the reference from the literature (Table 10). None of the scores reached the 0.4 threshold which is considered a moderately good agreement. The highest obtained score was 0.35 for helpfulness of Shap for pair 3 (Table 9).

In order to get an overview of the similarity of the scores and check if any clinicians show particular similarity to any other clinician the weighted Cohen’s kappa was also calculated for all the clinicians and all three explanations (Figure 28). Few significantly high and low scores were detected but they were not consistent, for example the difference was present only for one explanation within a particular pair. For example, clinicians 5 and 10 had high similarity score for Araucana – 0.44, but score 0.00 for Bayesian network and -0.31 for Shap Figure 28).

Table 9. Weighted cohen's kappa scores for each of 5 pairs of clinicians.



Cohen's kappa inter-pair agreement (helpfulness)			
Pair	shap	araucana	bayesian
1	-0.010	0.041	-0.069
2	-0.208	-0.116	-0.130
3	0.353	-0.301	-0.045
4	0.012	-0.040	0.333
5	-0.100	0.011	0.000

Cohen's kappa inter-pair agreement (comprehensibility)			
Pair	shap	araucana	bayesian
1	-0.078	0.070	0.074
2	-0.392	-0.283	0.000
3	-0.061	0.000	0.027
4	0.000	0.000	0.000
5	-0.062	-0.028	0.000

Table 10. The standard ranges for interpreting cohen's kappa values (Landis & Koch, 1977).

KAPPA (K) VALUE	INTERPRETATION
K = 1	Perfect agreement
K > 0.75	Excellent agreement
0.40 < K ≤ 0.75	Moderate to good agreement
K ≤ 0.40	Poor agreement
K < 0	Agreement worse than chance

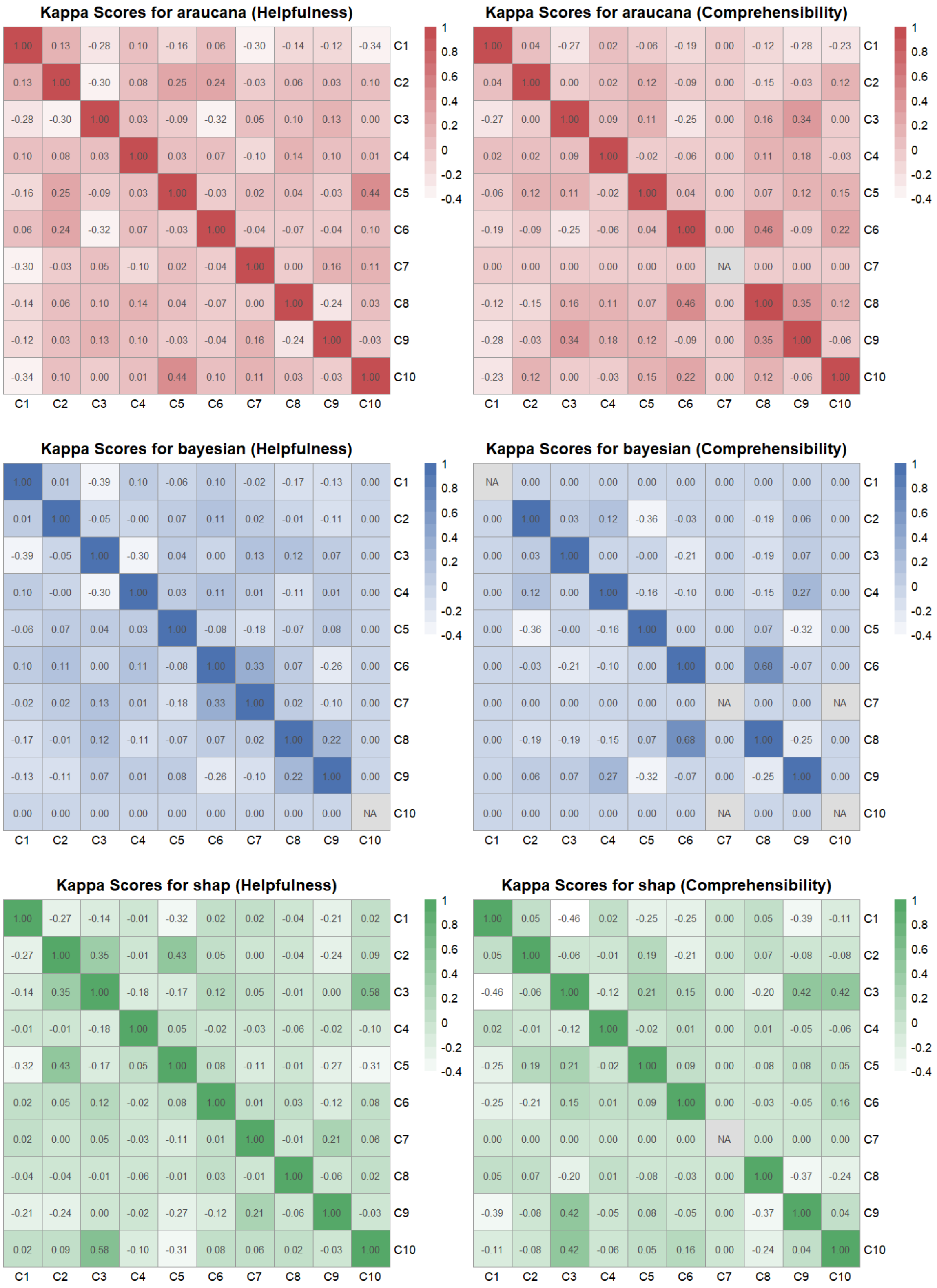


Figure 28. Weighted Cohen's kappa comparison to evaluate within-pair agreement. Pairs: 5:8, 1:9, 3:2, 6:7, 4:10. (Value of 0.00 is the result of no variability in rating for all 10 patients by one clinician, NA stands for variability equal to 0 during rating of two compared clinicians)

Think-aloud

Analysis of time

The data was obtained and the time was split according to the methods section. For the general overview and meaningful analysis, the individual time sections spent on each explanation were visualized, as well as the mean time spent on each explanation (Figure 29).

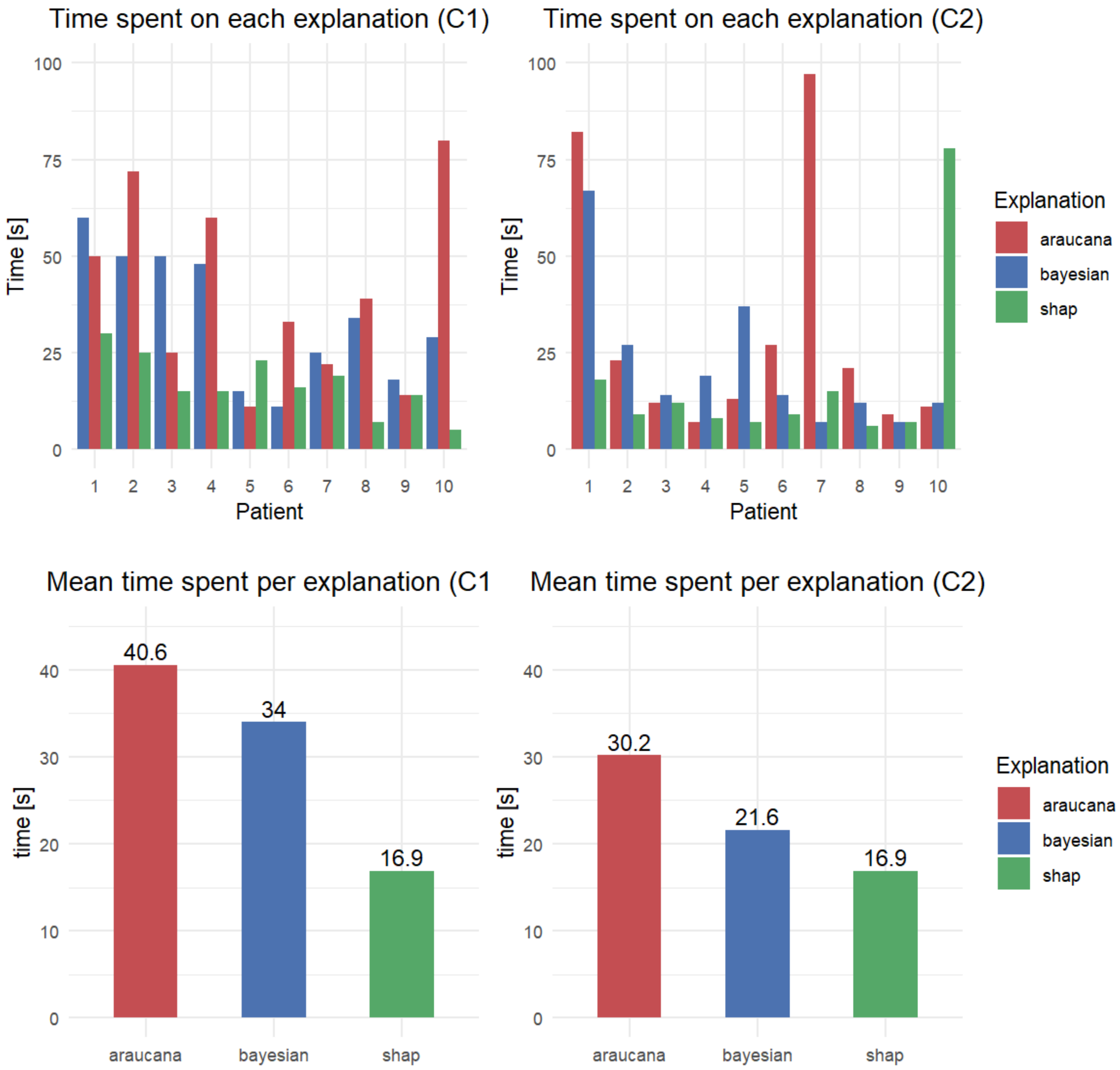


Figure 29. Time spent on each explanation over the course of the survey (top) and mean time spent on each explanation for C1 (left) and C2 (right).

Both clinicians (C1 and C2) spent least amount of time on the Shap explanation and the greatest amount of time looking at Araucana. Bayesian network achieved the middle position. C1 spent more time looking at the Bayesian network and Araucana explanation than C2. On the other hand, both clinicians spent the same mean time on Shap explanation (Figure 29).

Additionally, the time spent initially looking at each patient was also visualized for the general interpretation of the results (Figure 30). C1 spent more time initially looking at the patients, starting from patient 6. The distribution of time spent on initially looking at each patient is uniform for C2.

The distributions of time spent on each explanation show that both clinicians spent the least amount of time on Shap explanations with the lowest variability (Figure 29, Figure 31)The results are statistically significant according to the pairwise Wilcoxon signed-rank test for C1 for Shap explanations compared to Bayesian network and Araucana (Table 11).

Additional comparisons were made to compare the times spent on each patient based on the class it belongs (TP, TN, FP, FN). For C1 there is a visible increase of time during the presentation of the first patient for which the prediction is incorrect – patient 6 FN (Figure 32). A similar pattern is not observed for C2. For C2 a similar peak for the first FN case is also observed but in the sum of time spent on all three explanations (Figure 33).

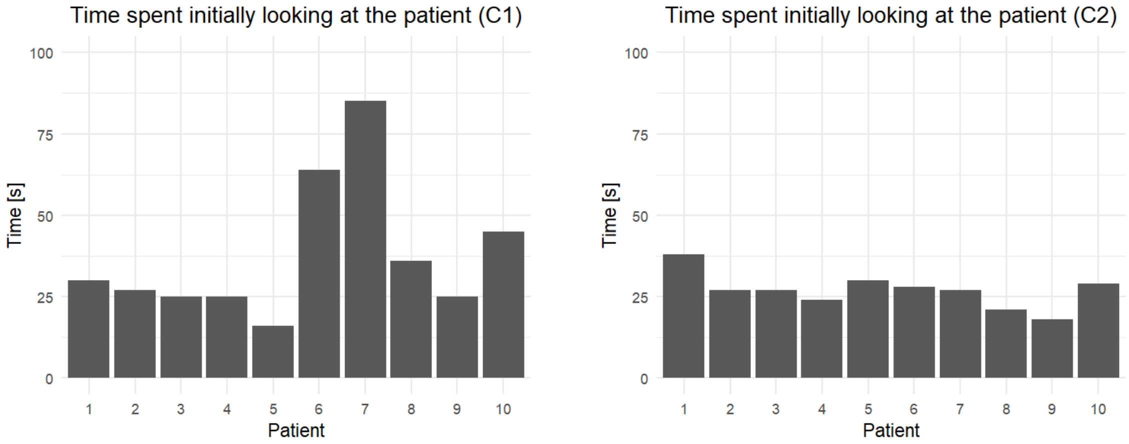


Figure 30. Time spent on each patient over the course of the survey for C1 (left) and C2 (right).

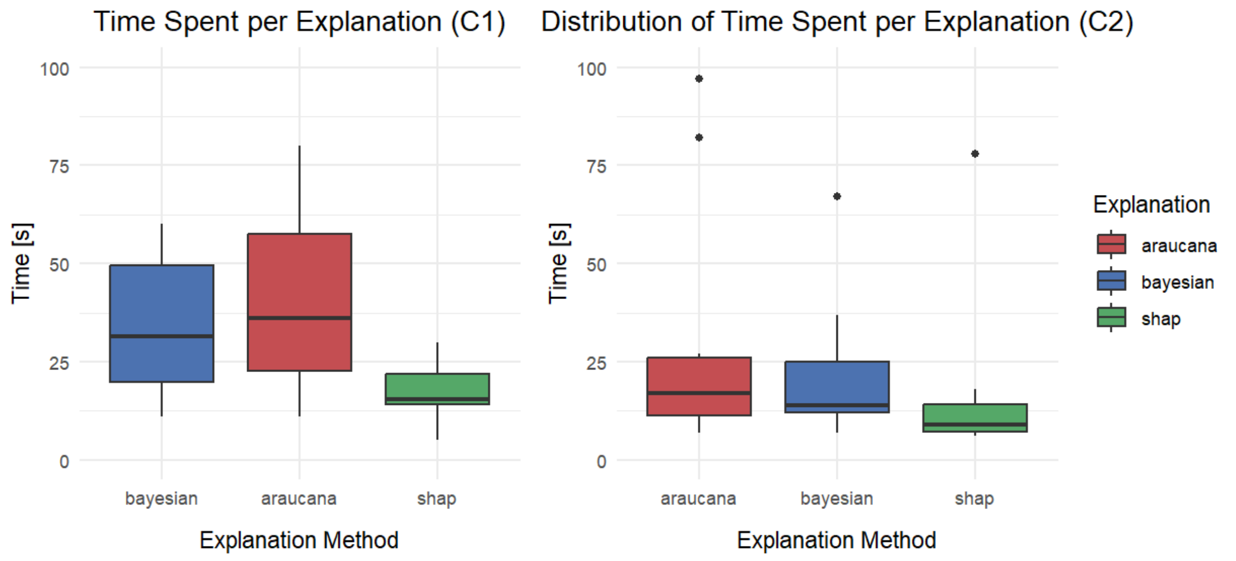


Figure 31. Box plots of the time spent for each explanation over the course of the survey for C1 (left) and C2 (right).

*Table 11. Pairwise Wilcoxon signed-rank comparisons of time spent on three explanations. * indicates statistically significant differences.*

PAIRWISE WILCOXON SIGNED-RANK TEST	C1	C2
COMPARISON	**P Value**	**P Value**
Araucana vs Bayesian	0.475	0.721
Araucana vs Shap	0.024*	0.123
Bayesian vs Shap	0.027*	0.193

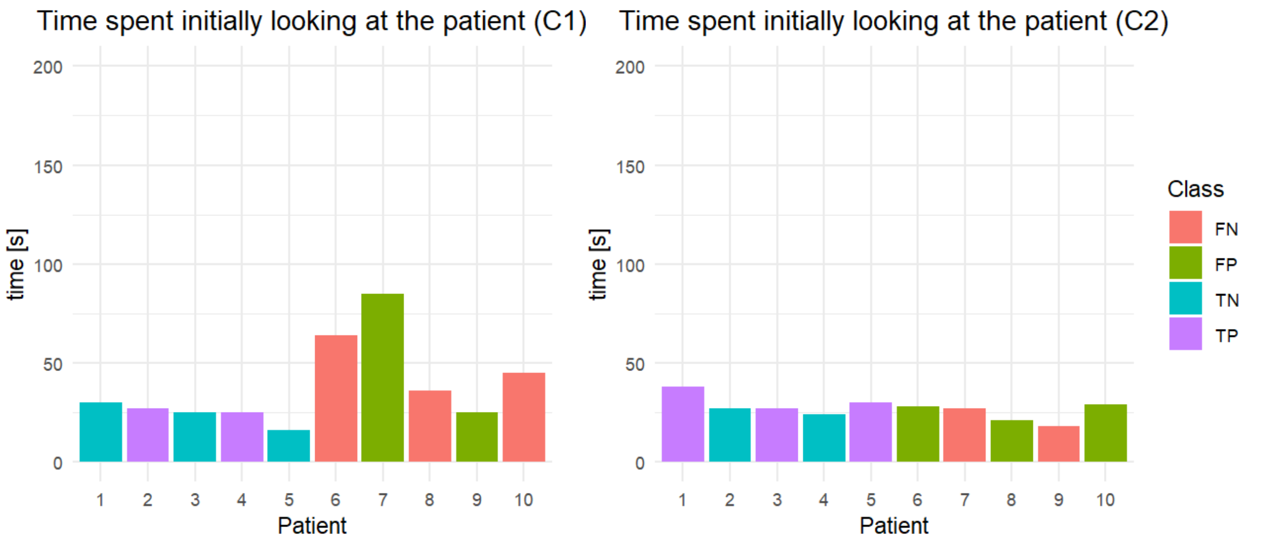


Figure 32. Time spent initially looking at each patient by C1 (left) and C2 (right).

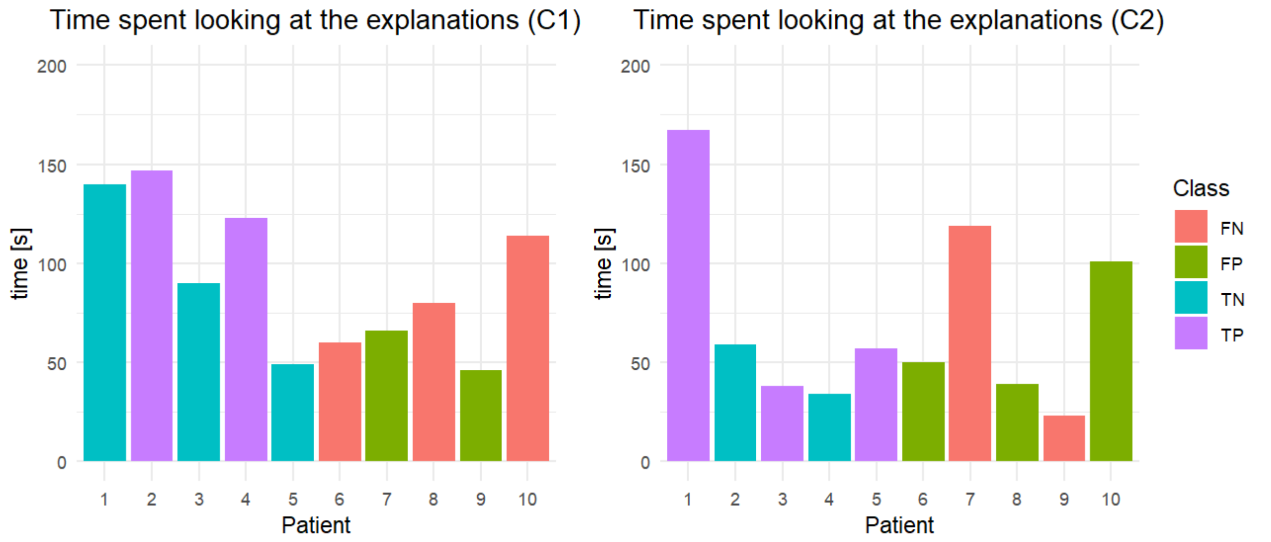


Figure 33. Sum of time spent looking at all the explanations for each patient by C1 (left) and C2 (right).

Analysis of sentiments

On the aggregate level there are no visible difference between three explanations for C1 for all three models (Figure 34 (left)), although Bert has lower average compound values than the other two models. On the other hand, the sentiments scores show the same pattern for all three models for C2 (Figure 34 (right)): the median and spread are shifted with Shap having most positive sentiments, Araucana slightly lower and Bayesian network with the lowest sentiment scores.

The time courses allow for the identification of individual utterances and outliers for both clinicians (Figure 35).

Only selected key utterances were directly displayed (Table 12). They represent the examples of the cognitive reasoning while interacting with the three explanations (Table 12 id 1-3) and ideas for improvement of the methods (Table 12 id 4-6). Clinicians indicated some strong opinions about the explanations such as very negative feelings towards the Bayesian network or highlighting the completeness of Shap (Table 12 id 10 & 9). Also, direct comparisons between the three explanations were made (Table 12 id 9). Additionally, the possible differences between studied departments ER and ID were mentioned (Table 12 id 11) and the differences in the level of difficulty of classification between presented patients were highlighted (Table 12 id 8). Additionally, the clinician from the ID department highlighted the strong need for seeing the patient for correct classification (Table 12 id 12).

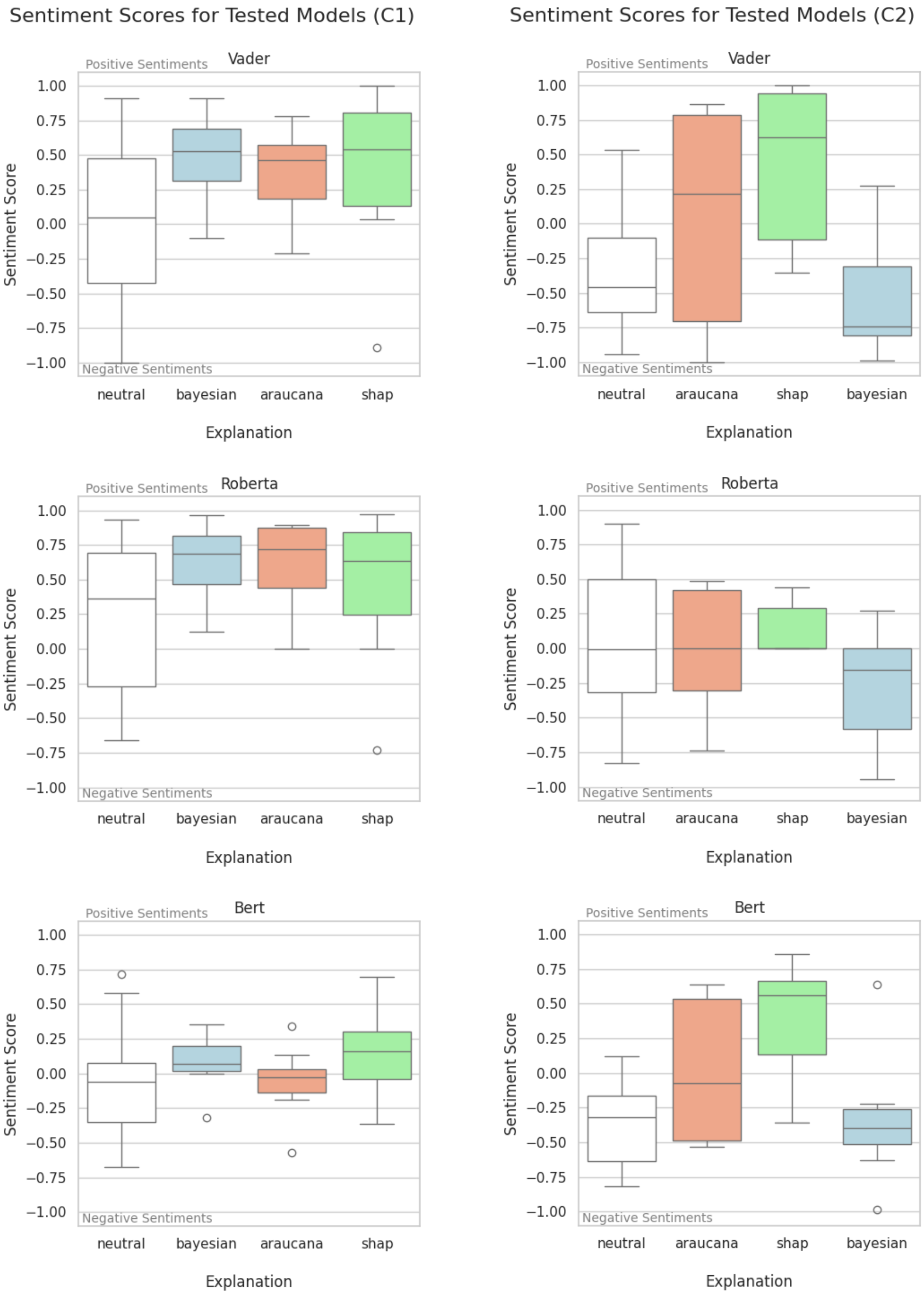


Figure 34. The sentiment comparison of three models (Vader, Roberta and Bert) for three explanations (neutral – control, Araucana, Shap, Bayesian), Clinician 1 (left), Clinician 2 (right).

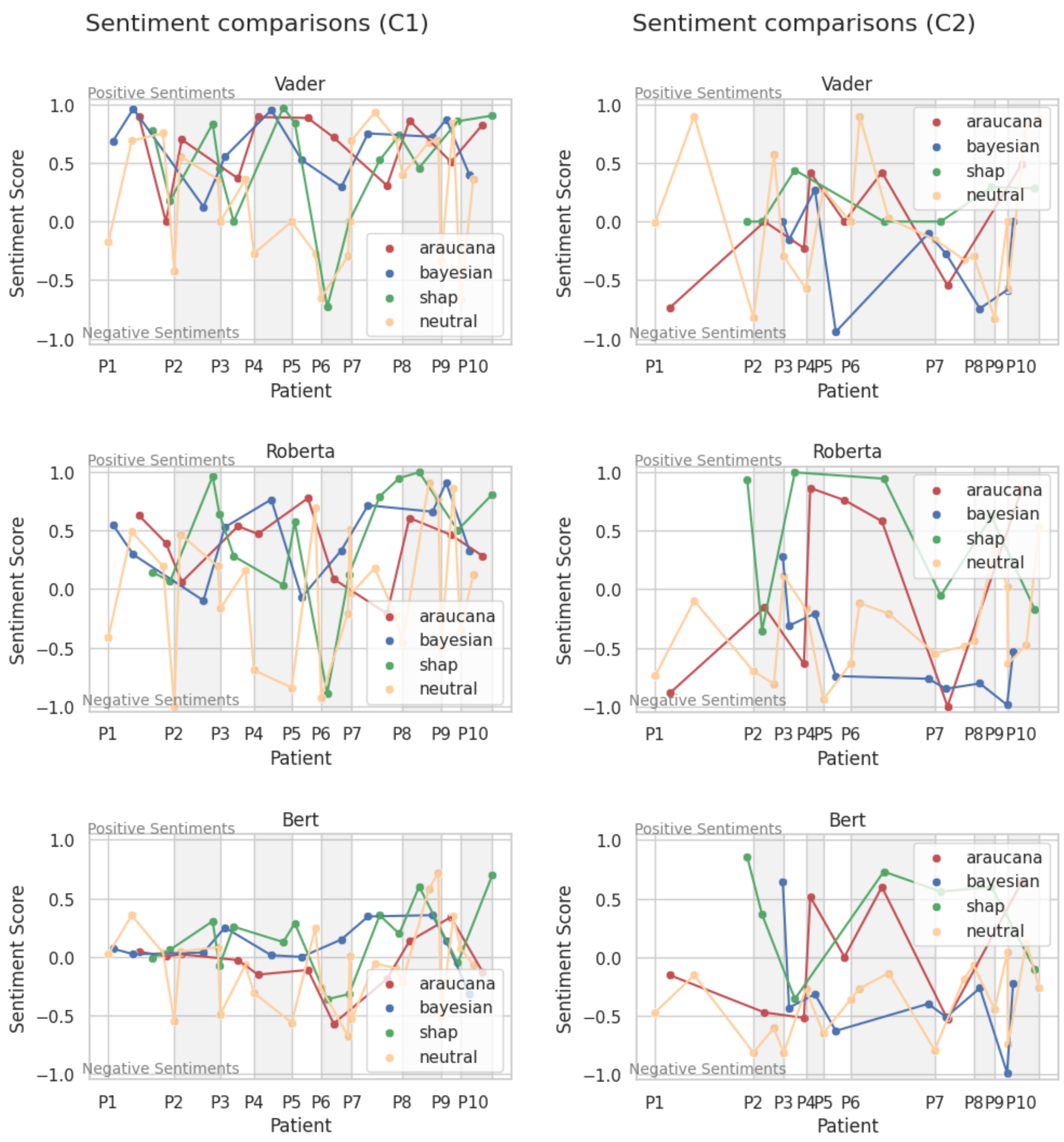


Figure 35. The sentiment comparison of C1 - Clinician 1 (left) and C2 - Clinician 2 (right) for three models (Vader, Roberta and Bert) used for sentiment analysis over the course of the survey for Shap, Araucana and Bayesian network. The y-axis represents the sentiment score with -1 indicating strongly negative, 1 strongly positive, and 0 neutral sentiment score. The x-axis represents the time course of the survey while assessing 10 patients (P1-10).

Table 12. Key utterances extracted from the think-aloud. They show the examples of the decision-making process for three explanations, key insights indicated by the clinicians about improving the explanations and few additional comments important for holistic understanding of clinical reasoning. Some utterances were slightly altered or paraphrased to make the utterance more clear and concise.

TABLE ID (CLINICIAN)	**UTTERANCE**	**SUMMARY**
1 **(C1)**	Okay, there is the explanation of Bayesian network, in which the various parameters are fixed and then it decides to send her home, on the basis basically of the parameters highlighted in red, just as an explanation of it, so it weighs the age towards hospitalization and the others instead all in sum are less than 0.3, 0.3, 0.3, 0.1 and I agree, both clinically and then as examinations and radiology, it is a lady that I would have sent home too. The explanation comes almost easier on the table than on the Bayesian net so designed in the sense that it is a bit more linear - my thinking, than buttons, lines, arrows, so the interpretation for goodness sake, I grasp it. In the sense of the coefficients in red, so it is clear to me, but not too much.	**Example of decision-making process for Bayesian network**
2 **(C1)**	And this is the Shap, so everything goes towards the discharge at home. Practically PCR, white blood cell oedema, langopassitis, so the characteristics, we say, main radiological results are good enough to go home also this is quite clear, certainly more than the Bayesian net in the sense that it immediately gives me an overview, but also as stuff, so I would say that like the other, it is always difficult to give the maximum in these things because you expect maybe there is something that goes beyond, but surely it is stands out from the first.	**Example of decision-making process for Shap**
3 **(C2)**	[Looking at Araucana explanation] Then, infiltration, less than 06, so go home. If you have difficulty breathing, less than 05, go home. If you have oedema, 05, you are hospitalized. So, intuitively understandable explanation. If you have difficulty breathing, less than 05, you go home. I do not understand this,	**Example of decision-making process for Araucana**

	also because here says difficulty breathing less than 05, but here tells you difficulty breathing false.	
4 **(C1)**	Here, it is more complete as reported data but remains a bit my confusion in having these lines without an arrow or a direction so the flow does not follow it really well but at least I understand that the radiological part conditions the choice and the final decision for the hospitalization even if with a coefficient not so high. Intuitive I would say no, 4 or less, the explanation helps me? being not so intuitive, not so much.	**Highlighting possible improvements for Bayesian network**
5 **(C1)**	At home, and instead here it pushes everything to send it home, I don't agree too much, in the sense that frankly intuitive absolutely yes, but I don't agree, i.e. it doesn't help me because it drags everything to one side in a way too striking, so there it is.	**Highlighting possible improvements for Shap**
6 **(C1)**	I agree with the hospitalization but logically I miss a passage here, while [referring to the previous explanation of Shap] before all were negatives, she goes home. Here there is one positive and the other is negative but she would have gone anyway seeing the various results. The decision is pulled a lot from the first parameter of PCR. It is definitely intuitive, but in my clinical reasoning - to be clearer and closer to the reasoning that I did - I would have put radiological data in it. Seeing that the white blood cells are not pathological, instead it reports the value of white blood cells.	**Highlighting possible improvements for Araucana**
7 **(C1)**	This, however, here is preferable because there are more parameters that are the clinically relevant ones that allow me to have a greater glance in this case than for Araucana so with this I would say that intuitively it helps me. This is really clear, more than the previous case.	**Expressing the completeness of Shap**
8 **(C1)**	Support of this kind for these borderline cases, in the sense that at least the support can help more than in clearly striking cases where you immediately decide, ‘this person needs to be hospitalized’.	**Indication of borderline cases**
9 **(C1)**	In this case I would say that absolutely [Shap] helps me the most, I would say yes, we agree, on the whole I agree. I believe Shap won by a wide margin, then Araucana, then the Bayesian Network in confidence and overlap. Sometimes the impression surpasses even the strength of the clinical decision. Shap	**Comparison between explanations**

	pushes it more strongly than my evaluation, but both graphically and in terms of the parameters it considers, it is the one that I like the most as an explanation.	
10 **(C2)**	I have to take a course on this [looking at Bayesian network]. *But actually, we can reveal that almost everyone has had this problem* [Experimenter]. Now I feel better, otherwise it would depress me. I consider myself very ignorant in this field. I really struggle with this. Maybe for a young doctor, yes. For me, a system like this is not intuitive at all.	**Expressing lack intuitiveness of the Bayesian network**
11 **(C2)**	[Paraphrased and summarised answer] I find Shap and Araucana more intuitive, especially for emergency doctors who think in algorithms. In the ward [ID], I make decisions through clinical reasoning and seeing the patient rather than algorithms. While such models might not replace a doctor's decision, I think they could be helpful in consultations, especially in emergency settings. This is similar to how we made decisions during Covid by examining data and tests.	**Explaining differences between two departments ID and ER.**
12 **(C2)**	There's no way to see them - the patients. Yes, this is a piece of paper. This is the classic patient you should see, you can't skip seeing the patients. This is an old man 92 years old who has pneumonia, you should actually see what this pneumonia is like.	**Highlighting the need for seeing the patient**

Discussion

The analyses provided qualitative and quantitative results. Mixed approach leads to more holistic overview of the decision-making process and multiple comparisons between the three explanations – Araucana, Shap and Bayesian network. The analysis of compliance and the positioning of the results in the existing frameworks provided a wide overview of the practical implementation of human-AI and human-XAI collaboration.

Survey

Compliance

In the study, the average compliance obtained equal to 86% was relatively high, especially taking into consideration that half of the presented cases were incorrect predictions (Figure 9).

In the context of the dual-stream theory in medical settings, general compliance with AI DSSs can influence both cognitive streams. Clinicians often rely on System 1 processes due to their extensive experience and the need for quick decisions. It occurs especially after people's initial positive assumptions of AI's consistent performance. A more popular term in the literature related to compliance is reliance which is defined as the degree of trust and dependence that clinicians place on the AI-DSS (Dlugatch et al., 2024). Reliance is gauged by the extent to which clinicians depend on the AI-DSS for their decisions. In this context, compliance is more relevant because clinicians were asked to evaluate the case only after seeing the patients' data and AI explanations. Nevertheless, high compliance can indicate reliance and trust in the AI-DSS so both were be considered.

High compliance might occur when clinicians overly rely on the highly intuitive recommendations of AI (System 1), potentially leading to over-trust and reduced critical evaluation (System 2). It is usually referred to as automation bias - the tendency of individuals to over-rely on automated systems, often leading to errors when the automation fails (Kazim & Tomlinson, 2023). Automation bias can lead to significant errors, especially in high-stakes environments like healthcare. For instance, when physicians rely too heavily on AI-generated diagnoses without verifying with their own expertise, it can result in misdiagnoses. Additionally, the complexity of verifying automated outputs can further increase this bias, which is very relevant for AI-driven support, where there is usually no possible verification in case of doubts. In that case, physicians may find it challenging to engage System 2 processes to critically assess the AI's recommendations (Goddard et al., 2014).

An important factor that was relevant in the analysis of compliance were years of experience of the clinicians. Even though the sample size in the study was small (10 clinicians), the positive relationship between years of experience and compliance is clear (Figure 10). It aligns with research which generally suggests the over-reliance on intuitive thinking (System 1) of experienced clinicians (Caddick et al., 2023; Tsalatsanis et al., 2015). Dual-process theory highlights the interplay between heuristics from years of experience (System 1) and logical deduction (System 2). Over their careers, clinicians develop significant expertise through accumulated System 1 experiences, leading to quick and accurate pattern recognition. However, this same experience can also introduce gaps and biases due to individual clinical encounters. These biases and knowledge gaps underscore the importance of continuous training to maintain cognitive skills. Research indicates that physicians further from their training years often perform worse on medical knowledge tests (Caddick et al., 2023). This decline is influenced by factors including specialization and cognitive changes over time. Additionally, acquired skills may become less accessible due to lack of study, aging, and competing knowledge.

Another factor influencing compliance is task difficulty. The inclination to over-rely on decision aids is influenced by the complexity of the task. As tasks become more challenging and approach the user's cognitive limits, there is a growing tendency to depend on external resources, which can sometimes lead to erroneous reliance (Tahtali et al., 2024). Moreover, the higher the stake of the decision the higher the reliance on the decision aid (van Dongen & van Maanen, 2013). However, the clinicians were not asked about the level of difficulty and stakes they had in mind while performing the tasks. Their subjective perception of task difficulty and stakes could shed more light on the high compliance rates. The think-aloud analysis suggests that the FP and FN cases were much more difficult to classify compared to TP and TN in the study (Table 12 id 7), and, therefore, are likely to be labelled as borderline cases.

The difficulty of establishing the ground truth of the cases assessed makes the interpretation of the compliance rates challenging. As indicated by the authors of the ALFABETO project (Catalano et al., 2023; Nicora et al., 2021) one of the limitation of the study was testing and validation of classifier results. The data was collected from a single centre without external validation. Additionally, the authors point out that all patients were managed in the tertiary centres, which creates the potential for the survival bias, particularly for patients categorized as having mild outcomes.

To sum up, high compliance indicates strong alignment between AI suggestions and clinician decisions. Providing Explainable AI (XAI) explanations most likely enhances trust and compliance (Schemmer et al., 2022), though high compliance also indicate over-reliance on the DSS which has to be addressed and mitigated.

To further understand the compliance in case of XAI it is essential to adapt the study design to collect clinicians' compliance both before and after seeing each XAI method. Present study cannot fully assess the differences between the three methods because the clinicians were asked to indicate if they agree with the prediction only after seeing patient's characteristics and all three explanations.

Comprehensibility and helpfulness

The comparisons of the individual scores revealed that clinicians have very different strategies and patterns related to scoring (regardless of department or years of experience). Some showed a lot of variance between the three explanations, while others showed little to no variance in scoring (Figure 18 Pair 4). Similarly, most clinicians showed differences between comprehensibility and helpfulness scores, others showed almost no variance, e.g. Pair 5 clinician 10 (Figure 16; Figure 18).

The analysis of the scores comparing Shap, Araucana and Bayesian network, have several important implications for the use of different XAI methods in medical decision-making. Firstly, the pattern of highest scores for Shap in both comprehensibility and helpfulness suggest that this method may provide more reliable and understandable explanations for clinicians, potentially leading to better-informed decisions. The intermediate scores of Araucana indicate it is a viable alternative, offering a balance between interpretability and accuracy. The lower scores for the Bayesian network highlight potential limitations in question its ability to provide clear and useful explanations in the present context (Figure 21; Figure 22). Some clinicians gave consistently low scores to Bayesian network which suggests lack of understanding of the method throughout the course of the survey, e.g. Pair 2 clinician 1 (Figure 17).

The comparisons of scores divided by class (TP, TN, FP, FN) reveal more nuanced differences between the methods. For example, both Shap and Araucana have significantly higher comprehensibility scores compared to Bayesian network for all classes (Figure 25). The differences between Shap and Araucana are less significant; the significant difference is found only in FP cases. It suggests that for the FP cases, Shap is showing the most pronounced and statistically significant differences for comprehensibility. However, having in mind that only 3 out of 24 FP cases were correctly detected (Figure 27), it seems that Shap may inadvertently drive the compliance. Similarly, the differences in comprehensibility between FN cases considering Shap and Araucana are much smaller than between TN and TP cases (Figure 25). The same pattern is observed for helpfulness, but the differences are less pronounced than for comprehensibility (Figure 26).

The analysis of few cases in which the clinicians correctly disagreed with the FN predictions show that, for helpfulness, Araucana receives higher scores than Shap and Bayesian network (Figure 27). It suggests that lower level of comprehensibility can be the nudge to question AI in case of doubt and engage more system 2. On the other hand, high comprehensibility can be the driver of compliance, increasing reliance and trust, being too convincing. This anecdotal evidence is in line with literature. High comprehensibility and helpfulness scores of Shap reflect high usability (Hoffman et al., 2018), but can be also the drivers of over-reliance, especially in high-stakes and difficult decisions (Zytek et al., 2021). Conversely, less comprehensible methods such as Araucana require more effort leading to lower usability scores compared to Shap. But at the same time they can mitigate over-reliance engaging critical reasoning (Clement et al., 2023). Nevertheless, low comprehensibility scores may indicate a negative bias towards a method or a lack of understanding, as seen with Bayesian networks when they are overly complex and unintuitive.

These results underscore the importance of selecting the appropriate XAI method based on the specific needs of the clinical context: previous experience, the decision stakes, time constraints, cognitive resources and many more. The ability of Shap and Araucana to provide higher and statistically significant scores suggests they are best candidates to enhance the trust and usability of AI systems in healthcare, ultimately improving patient outcomes. Further research and continuous evaluation are necessary to refine these methods and ensure they remain effective as standards of care evolve.

Department comparison

The analysis of explanation preference showed that clinicians from the ID department have very strong preference for Shap, while clinicians from ER show more balanced preferences choosing Shap and Araucana in almost the same proportions (Figure 24). It is reflected in the comprehensibility and helpfulness scores (Figure 22; Figure 23). It is the opposite of what might be expected, since ER is characterised by mainly fast decision-making (seconds to minutes) and ID by mainly slow decision-making (hours to days), as indicated by clinicians themselves during the think-aloud session (Table 12 id 11). The higher preference for Araucana by the ER department maybe be due to its algorithmic nature expressed by the clinicians and previous exposure to similar methods (Table 12 id 11). It hints at importance of prior experience of the clinicians in DSS design and the complexity of choosing the most suitable method. It is a factor that should be taken into account in future studies.

Within-pair comparison

The results of the study provide valuable insights into the similarity of helpfulness and comprehensibility scores among clinicians. The use of weighted Cohen’s kappa allowed for a systematic comparison of the scores, revealing several key findings.

Firstly, the pairwise comparison of scores within each clinician pair did not reach the threshold of 0.4, which would be considered a moderately good agreement (Table 9). The overall calculation of weighted Cohen’s kappa for all clinicians and explanations provided a broader view of the similarity in scores (Figure 28). Although a few significantly high and low scores were detected, these were not consistent across different explanations or clinician pairs. This inconsistency indicates that the agreement on helpfulness and comprehensibility scores may vary depending on some parameters that were not measured in this study.

These findings have several implications. The lack of consistent high agreement suggests that there may be inherent differences in how clinicians perceive and evaluate explanations. This could be due to individual differences in experience, expertise, or personal views (Dinos et al., 2017). Understanding these differences is crucial for improving the design and presentation of explanations to ensure they are helpful and comprehensible to a wide range of clinicians. The differences in scoring could be also caused by the lack of uniform understanding of the concepts of comprehensibility and helpfulness, as well as the lack of reference (Hoffman et al., 2018). This source of variability could be removed by providing examples of scoring and clear explanations of the terms at the beginning.

Overall, no consistent patterns or correlations were found. Nonetheless, the within-pair comparison has some limitations. The sample size of clinician pairs may not be large enough to generalize the findings to a broader population. Additionally, the use of weighted Cohen's kappa, while useful, may not capture all nuances of agreement and disagreement. Future research should consider larger sample sizes. It would allow for alternative methods of assessing agreement such as clustering of clinicians based on rating strategies (Dopp et al., 2020).

Think-aloud

Two general important points were raised by the clinicians during the think-aloud session that might improve similar studies and provide a fuller picture of the results.

The presented FP and FN cases might be borderline cases in which the ground truth is difficult to assess as indicated by the clinicians. It might (at least partly) explain the high compliance rates. Further evaluation of the presented cases and their classification in terms of level of difficulty could provide more holistic evidence. For instance, agreement with a prediction in ambiguous cases (where classification could go either way) does not carry the same weight as agreement in cases where the AI algorithm clearly misclassifies. (Table 12 id 7).

The differences between every day decision-making at the two departments should be taken into account during the interpretation of the results and while developing XAI DSSs. According to the think-aloud session, the ER practitioners use more algorithmic approach: they make the decisions more quickly due to time constraints and emergency cases. On the other hand, at the ID department the decision-making is more extended in time and involves many more parameters, i.e. radiological images and seeing and talking with the patient, as indicated by the clinicians (Table 12 id 11).

Analysis of time

The analysis of the time spent on each explanation and the initial patient observation provides valuable insights into the decision-making process. However, they cannot be discussed separately as they are meaningful only analysed together with the rest of the results.

The analysis of the mean time spent on each explanation reveals distinct patterns in the clinicians' approaches. Both clinicians, C1 and C2, allocated the least amount of time to Shap explanations, while dedicating the most amount of time to Araucana, with the Bayesian network falling in between (Figure 29). However, various factors can influence the amount of time a clinician dedicates to analyzing an explanation (Sagar & Saha, 2017). According to the utterances, which show a consistent pattern, low time spent on Shap indicates that it is highly intuitive and self-explanatory (Table 12 id 2). Araucana requires more time for following the conditional structure of the tree (Table 12 id 9 & 3). On the other hand, in many cases the explanation of the Bayesian network was skipped due to lack of comprehensibility and therefore lack of helpfulness (Table 12 id 1 & 10). In this context, the results suggest that the three explanations are on a complexity continuum. The more complex the method, the more time is needed to extract meaningful information from the explanation. While Araucana is still comprehensible but requires some time, the Bayesian network would require too much time and effort to extract meaningful information, according to the clinician's assessment (Table 12 id 1).

Interestingly, C1 consistently spent more time on both the Bayesian network and Araucana explanations compared to C2 (Figure 29), indicating a possible difference in their reasoning styles or depth of detail. The uniform time distribution for Shap across both clinicians highlights a shared perception of its relative simplicity and intuitiveness. C1 belongs to the ER department and C2 to the ID department and the time differences are directly mirroring the explanation preferences of both departments (Figure 24).

Additionally, the initial time spent at looking at the patient's tabular data varied significantly between the clinicians (Figure 30). C1 spent more time initially observing patients' data, particularly from patient 6 onwards, while C2 maintained a uniform distribution of initial observation times. This discrepancy may reflect differences in their diagnostic approaches or the accuracy of their initial assessments. The statistically significant results from the pairwise Wilcoxon test for C1's Shap explanations compared to the Bayesian network and Araucana further underscore its ease of use (Figure 31; Table 11). Comparing the time each clinician spent looking at the patients (at the beginning of the case) and looking at the three explanations, both clinicians generally spent more time on the FP and FN cases. It is reflected by time spent initially looking at the patient for C1 and longer time spent on the explanations for C2 (Figure 33). However, the doubts reflected by allocating more time to FP and FN cases did not result in prediction disagreement. Both clinicians showed high compliance of 90%.

However, it is important to note that the time measurements were taken during the think-aloud session (Sagar & Saha, 2017), i.e., while the clinicians were encouraged to spend more time on each explanation and to express all their thoughts. Sometimes the comments of different explanations overlap, making it difficult to set accurate and objective time boundaries. Additionally, general or neutral comments are scattered throughout the session. Time measurements during the survey for another group of clinicians without the think-aloud would account for those differences.

Analysis of sentiments

Analysis of sentiments is one of the approaches that are suitable for assessing another usability aspect not mentioned before and usually not discussed in the context of human-XAI collaboration – user experience (Imaduddin et al., 2023).

However, in the present case due to the small sample sizes and unresolved outlier instances, the results are only tentative and should serve mainly as a mean for generation of future hypotheses.

Taking into account the analysis of the aggregate sentiment scores, only the second clinician shows differences between the XAI methods. One possible reason might be the differences between the departments. The comparisons between the departments showed that ER department explanation preference is balanced with Shap and Araucana as a preferred method almost in equal proportions and Bayesian network being the preferred method at least in 1 in 5 of the assessed patients. On the other hand, clinicians from the ID department showed overwhelming preference for Shap. Taking into account that C1 belongs to the ER department and C2 to the ID department, the sentiment analysis may reflect this difference.

Alternatively, the lack of differences for C1 may be the unsuitability of the chosen models. Most of the models were trained on large social media datasets and on the English language. The lack of fine-tuned models for the specific medical jargon and languages other than English such as Italian makes the automatic classification of sentiments difficult. The full automation of the sentiment scoring might not be possible for highly-specialized settings such as healthcare (Imaduddin et al., 2023). On the other hand, it creates a unique opportunity for expressing and comparing sentiments quantitatively, provided the process is closely supervised, adjusted, and enough data is available. Preferably, manual scoring by multiple raters could serve as a meaningful comparison for automated sentiment scoring.

Another key point includes the difference in the structure and nature of the data that is used to usually train sentiment models and think-aloud transcripts. The models are trained on usually well-structured sentences on the other hand think-aloud represents a stream of thoughts and ideas including a lot of pauses, colloquial expressions and often lack of full and structured sentences. Multiple approaches should be tested including revision of the transcripts and adding more structure to the sentences and thoughts. However, this approach may lead to additional biases. Ultimately, the sentiment analysis is able to process only verbally expressed information discounting for other forms of communication and responses that are more difficult to capture such as tone of voice, body language, eye movements or physiological responses. Ideal assessment of the emotional tone would include the integration to at least a few techniques. Taking that into account, sentiment analysis seems to be a great option as one of them.

The time courses of the sentiments for different models allow for the identification of individual utterances and their sentiment scores through time. Nevertheless, the extraction of meaningful quantitative comparisons requires further investigation. It might involve further fine-tuning of the models, testing various criteria for division of the utterances and exploration of the accurate comparison level.

Additionally, carefully selected utterances provide key insights directly from the clinicians. It is an invaluable indication of their reasoning processes, general impressions and important comments referenced throughout the discussion section.

Integrated Discussion

According to (Zytek et al., 2021), the identification of the existing usability challenges is the first step to address them and appropriately guide the design of the solutions. It is important to remember that the choices are highly specific to a particular domain and context. Introduction of new DSSs would almost always require the observations, interviews and user studies in order to avoid usability issues. In that way DSSs with XAI do not differ extensively from other software tools and DSSs. The authors identify seven main usability challenges of XAI DSSs and corresponding solutions. According to the aforementioned framework clinicians surveyed here have trust in the system, know the consequences of actions, the prediction by the model is clear and relevant. The main challenge that can be identified is 'Difficulty Reconciling human-ML Disagreements'. For example in our study, during the think-aloud session, the clinicians expressed their doubts or disagreements for FP and FN cases but still proceeded with the direct answer that they agreed with the predicted class. The difficulty reconciling human-ML disagreements is suggested broadly to be mitigated by local explanations (Zytek et al., 2021). However, more precise measures targeting this issue should be further studied and developed. Evaluating compliance for each XAI technique separately would be invaluable.

An unexpected result of the study was the disproportionately low performance of the Bayesian network explanation compared to Araucana and Shap. It performed the worst out of three explanations almost on every dimension measured. The reasons of this disproportion should be further investigated. For example, determining the graphical structure of the Bayesian Network remains a major challenge, especially when modelling a problem under causal assumptions.

Solutions to this problem might include the automated discovery of Bayesian network graphs from data, constructing them based on expert knowledge, or a combination of the two (Kitson et al., 2023) at the same time taking into account the specific user needs. Moreover, the visualization could benefit from the studies of usability and interface design. By following the design principles such as Google's 'Material Design' (*Material Design*). The examples could include meaningful pre-selection of the displayed parameters, colour coding of the parameters, matching the size of the elements to mirror the strength of relationships or adjustment of fonts and text position.

Moreover, according to (Clement et al., 2023) Bayesian networks do not belong to the most explainable models representing more the class of medium complexity and medium explainability (Figure 36). It suggests that XAI could be as interpretable and understandable as the simplest models. It could surpass complex models even if they are explainable-by-design.

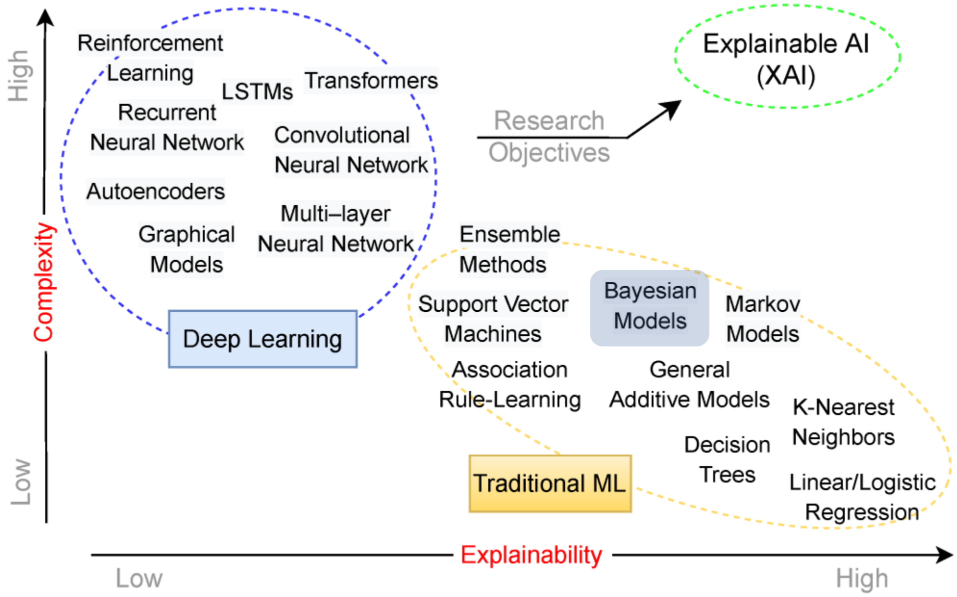


Figure 36. Classification of algorithms based on complexity and explainability (Clement et al., 2023).

Bayesian network explanation could be adjusted to better fit the context and cognitive needs of the users. The clinicians themselves indicated that both the design and lack of self-explainability of the Bayesian network don't make it usable. Appropriate design and education would increase its usability. However, the need for additional workshops to understand Bayesian network and higher cognitive demand compared to other explanations make it less competitive compared to Shap and Araucana which would not require additional courses because of their self-explanatory nature. Nevertheless, further research should be undertaken to make a final conclusion and identify which factor is responsible for lack of usability in this case.

Existing explainability frameworks such as (Combi et al., 2022) define usability as a separate dimension other than usefulness, interpretability and understandability (Figure 37).

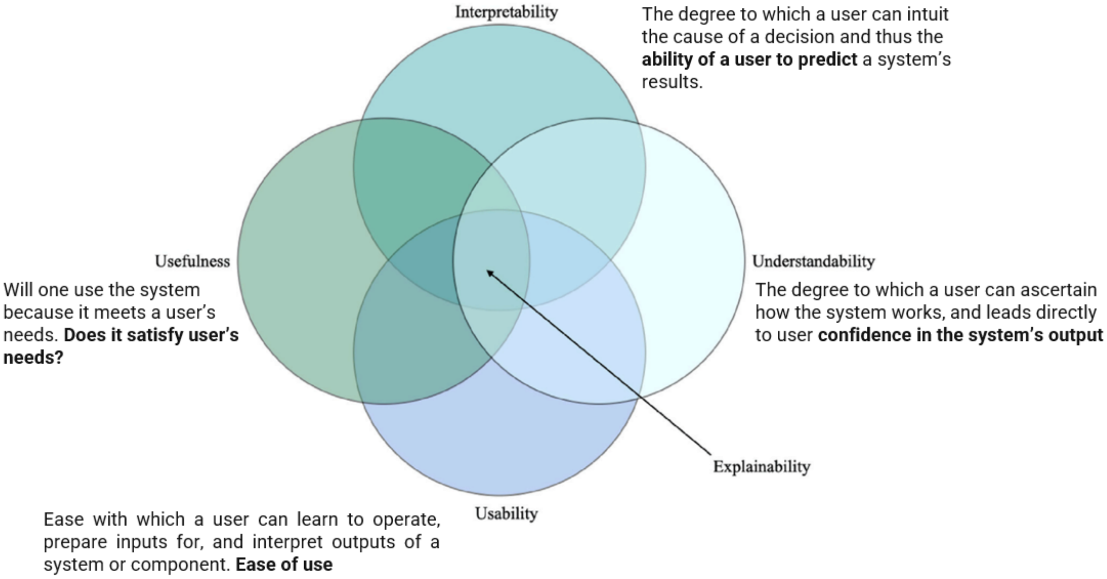

Interpretability
The degree to which a user can intuit the cause of a decision and thus the **ability of a user to predict** a system's results.
Usefulness
Will one use the system because it meets a user's needs. **Does it satisfy user's needs?**
Understandability
The degree to which a user can ascertain how the system works, and leads directly to user **confidence in the system's output**
Explainability
Usability
Ease with which a user can learn to operate, prepare inputs for, and interpret outputs of a system or component. **Ease of use**

Figure 37. The dimensions of explainability by the XAI Manifesto (Combi et al., 2022).

Here, usefulness and interpretability can be assessed through the measure of helpfulness and comprehensibility. The results of the ratings of all study measures show a similar pattern with Shap being the most helpful and comprehensible, Araucana being slightly less helpful and comprehensible and Bayesian network with the lowest scores for both dimensions. The dimensions of usability and understandability are much more difficult to interpret in this context. There is a great overlap of these two definitions with the comprehensibility measure (Press et al., 2015; Zhang & Adipat, 2005) because in order for a system to be easily learnable – sometimes used interchangeably with usable and understandable it should be also comprehensible by definition. Some literature sources (Arrieta et al., 2019) suggest that the main distinction between the two terms is that understandability focuses on the clarity of the model's function, while comprehensibility emphasizes the clarity of the knowledge the model has learned (Arrieta et al., 2019). In that case understandability is crucial for the developers of the AI systems while it is not relevant for the clinicians. The remining dimension is difficult to satisfy due to broad definition of usability. In the context of this study, the best indirect measure of usability could be the time spent on each explanation. It is due to the fact that time seems to be the best proxy for the cognitive load which is used in the usability studies and in evaluating XAI explanations (Herm, 2023). In that case again Shap would be the most usable, Araucana would be still usable but would require more cognitive resources. On the other hand, Bayesian network would require too much cognitive effort and was not analysed in detail by the clinicians. The evidence for it is revealed by the think-aloud session providing clear evidence that the underlying reasoning of Shap and Araucana are reflected in the decision-process (Table 12 id 2 & 3). On the other hand, the reasoning process for the Bayesian network is not reflected in clinicians' expressions (Table 12 id 1 & 10). They express rather the lack of understanding and lack of intuitiveness of the method. Therefore, both Shap and Araucana can be classified as understandable and Bayesian network in its present form as not understandable. It is in line with the cognitive load paradigms and usability research (Sagar & Saha, 2017; Zytek et al., 2021). The three methods are based on two different operations: Shap requires addition and subtraction, on the other hand Araucana and Bayesian network require evaluation of conditional statements. Even though direct cognitive load studies have not compared addition and evaluation of conditional statements, addition is intuitively less cognitively demanding. This is due to the multiple steps involved in each conditional step requiring remembering the thresholds and comparing the test result of the patient with the thresholds. Each of these steps increases the cognitive load, making the task more intricate (Sweller, 1988). Addition is generally more straightforward, particularly when working with smaller numbers or simple sums.

Reiterating on the definitions of usability and interpretability (Figure 37), they represent a difficult cognitive-load trade-off. Usability tries to minimize the cognitive-load while greater interpretability would often result in cognitive load increase. For example, in complex problems it might not be possible to make the explanation complete and at the same time easily readable for the user. In these cases, additional training might be necessary to equip the clinicians with necessary skills to interact with more complex explanations. However, it requires first the awareness and willingness of the physicians to adopt new techniques and making the explanations as usable and intuitive as possible while still retaining the adequate complexity of XAI.

The experiment design did not compare the agreement with the explanations separately. However, it could be easily speculated that higher usability would directly increase trust and agreement with the system (Cabitza et al., 2023). Unfortunately, it might be also illusory in the case of XAI. For example, the clinicians which participated in the think-aloud session expressed the understandability of Shap and Araucana but at the same time showed very high compliance rates (90%) on average. It suggests that simply increasing the usability of the system and explanations without additional measures mitigating over-reliance could inadvertently significantly increase the over-reliance automation bias.

The results clearly point out that Bayesian network and Araucana require more time to be understood but their explanations convey more complex information about relationships between the parameters. Therefore, if more complex causal links are important for a clinical decision, using simple summation explanations such as Shap could be very misleading and lead to clinical errors.

Another often overlooked issue is the timing and frequency of explanation requests from the system. Naïve and occasional users typically need frequent explanations at various stages of using the AI system, whereas experienced users, who integrate the system into their daily clinical routines, may need explanations less often, focusing on rare or unexpected situations. The level of detail and duration of these explanations can also vary, depending on the specific needs of different stakeholders in various contexts and with different objectives (Combi et al., 2022).

Additionally, lower levels of explainability can be acceptable in scenarios where they do not increase the risk of patient morbidity or mortality, such as the design of the disease screening tests with high false positive rates. Therefore, when the stakes are low, a lack of explainability can be tolerated, and the level of explainability is proportional to the stakes of a decision (Arbelaez Ossa et al., 2022; Zytek et al., 2021).

As previously discussed, XAI itself used in an inappropriate way can cause more harm than good for example increasing confidence in a faulty system (Cabitza et al., 2022, 2023). It is not a simple solution to AI transparency problems. XAI techniques can be faulty as they can be misled by adversarial attacks or noise, and discrepancies may exist between the explanation and the actual model behaviour. Users may struggle to understand good explanations and fall prey to biases like confirmation bias, but proper training and well-designed interfaces can address these issues, similar to managing biases in conventional medical procedures. The lack of objective measures for explainability makes it hard to assess XAI techniques, necessitating more research and interdisciplinary efforts to establish reliable evaluation methods. While XAI may work better on an aggregate level, individual-level explanations can be useful for certain data types like EHR, requiring continuous monitoring and oversight to detect and address unforeseen issues. XAI should not replace evidence-based evaluation methods but foster support in medical practice. There is no doubt that XAI can help to uncover biases and improve understanding of AI predictions when used correctly (Cinà et al., 2022).

Moreover, different stakeholders can benefit from XAI in different ways. XAI techniques can support various professionals using AI tools, such as developers, medical experts, and regulators, each with distinct explainability needs. While these techniques are beneficial for trained professionals like developers and clinicians, they should be fine-tuned. In a similar fashion a system tailored to a particular setting and user group cannot be easily transferable to another setting without re-evaluation (Cinà et al., 2022).

Conclusion

The combination of survey data and think-aloud protocols offers a robust understanding of clinicians' interactions with these tools, highlighting both their strengths and areas for improvement.

While the survey reflected the aware assessment of the methods, think-aloud captured the cognitive reasoning while interacting with XAI and cognitive load through time measurement. The cumulative evidence collected through the comparison of self-reported experience, time measurement and analysis of sentiments paints a clear picture of the assessed XAI methods.

Clinicians generally trusted the system as well as the explanations and found the model's predictions trustworthy and relevant (Table 13 H1). However, a significant remaining challenge is reconciling human-ML disagreements, reflected in high compliance (Table 13 H2). The specific experimental design of case presentation used provides further insight of compliance. The cases were always presented in blocks of few TP and TN patients first in order to prevent the inadvertent bias a clinician might have seeing FP and FN cases first. The goal was achieved but also the opposite effect was observed – the over-reliance or automation bias.

Comparisons of XAI explanations show consistent differences between the methods reflected in measured parameters (Table 13 H3). The results indicate that Shap is generally considered more comprehensible, helpful and usable than Araucana comparing self-reported comprehensiveness, helpfulness and cognitive load. Bayesian network received significantly lower scores for all three measures. Including the evidence from the think-aloud utterances it can be considered as not usable. This discrepancy needs further investigation, particularly in the graphical structure of Bayesian networks and their usability. The study suggests that better design and education could improve the usability of Bayesian networks, but they would require more cognitive effort and additional training in contrast to more self-explanatory methods like Shap and Araucana. Further research is needed to pinpoint the detailed factors affecting usability.

Table 13. Final hypotheses evaluation.

Hypotheses	Evaluation
H1: General Perception: Clinicians generally perceive human-AI collaboration tools as positive and trustworthy in the medical setting.	Accepted
H2: Compliance: Significant proportion of the compliant decisions are incorrect, indicating potential over-reliance on the XAI system.	Accepted
H3: Method Comparison: there are significant differences in comprehensibility, helpfulness and cognitive load measures among Shap, Araucana tree and Bayesian network.	Accepted
H4: Sentiment Analysis (emotional tone): Sentiment analysis reveal differences in the emotional tone between Shap, Araucana tree and Bayesian network mirroring the perceptions found with the use of the survey.	Partially accepted
H5: Department Comparison: There are significant differences in tool perceptions and preferences between ER and ID departments, with ER clinicians preferring tools that provide rapid, comprehensible explanations and ID clinicians preferring tools that offer more detailed, in-depth explanations.	Rejected
H6: Pair Comparison: Within allocated clinician pairs, there are consistent patterns in comprehensibility and helpfulness ratings, reflecting similar perceptions and preferences of clinicians from ER and ID departments while assessing same patients and corresponding XAI explanations.	Rejected
H7: Sentiment Analysis (suitability): General purpose sentiment analysis models are suitable for assessing emotional tone towards XAI tools in the medical setting.	Partially accepted
H8: Explainability Assessment: Studied tools can be fit into theoretical frameworks and highlight the need for theoretical and empirical studies being conducted together.	Accepted

The overall preferences is mostly driven by the clinicians from the ID department which show strong preference for Shap, moderate preference for Araucana and very low preference for Bayesian network. On the other hand, clinicians from the ER department show high preferences for both Shap and Araucana and moderate preference for Bayesian network. Shap can be considered as the least complex explanation out of three due to its additive nature compared to conditional reasoning necessary to extract the meaningful information from the Araucana tree or Bayesian network. It suggests that if a problem at hand is relatively simple and linear, Shap is a preferred XAI approach. On the other hand, Araucana would be more appropriate for explaining problems that require more complex non-linear operations. Unlike previously hypothesized ID department showed preference for fast and intuitive method – Shap and ER for more in-depth explanations Araucana and Bayesian network. (Table 13 H5).

Sentiment analysis median and spread results showed a similar pattern to self-reported measures reflecting the differences between the departments. However, it would require more participants and further fine-tuning to provide conclusive evidence (Table 13 H4). The limitation of the study was a relatively small number of participants which does not allow for generalization of the results and hindered statistical analyses with only two clinicians. Therefore, also the acceptance of the general purpose sentiment analysis models in medical setting cannot be fully confirmed without further research (Table 13 H7). However, presented preliminary results indicate a great potential for semi-automated sentiment analysis with general-purpose sentiment models.

Fitting the results into the 4 dimensional explainability framework (Table 13 H8), three dimensions seem to be relevant for XAI users: usefulness, interpretability and usability. Based on the results Shap achieved the highest explainablity, Araucana slightly lower and Bayesian network can be considered as not explainable. Two key dimensions – interpretability and usability – are in opposition to each other. suggesting a difficult cognitive load trade-off in making the explanations sufficiently interpretable and complete without hindering their usability . However, this trade-off cannot be definitely solved and has to be fine-tuned based on the specific needs such as time constraints, stakes of the decision and user expertise to name few. The frameworks for classification of task complexity and requirements gathering should be further investigated in the context of XAI DSSs.

This study provides a comprehensive evaluation of various XAI tools, highlighting significant differences in their usability and clinician preferences. The acceptance of hypotheses related to general perception, compliance and method comparison underscores the reliability and relevance of the findings. The combination of survey data and think-aloud protocols offers a nuanced understanding of how clinicians interact with these tools, revealing both strengths and areas for improvement. The findings emphasize the importance of tool design and education, particularly for more complex methods like Bayesian networks. Future research should focus on refining these tools to reduce cognitive load and improve usability, possibly through targeted training and adjusted graphical design.

Future directions

The future directions are multiple. Reconciling human-ML disagreement is an urgent issue. A valuable insight could be gained by performing a similar experiment but presenting the XAI explanations only for the difficult cases. This could prevent the clinicians from getting used to seeing the explanations being correct multiple times and it might be the easiest and the most effective way for decreasing over-reliance. Generally, large-scale studies with larger sample sizes, investigating long-term effects in XAI-collaboration have yet not been conducted. Additionally, recruitment of clinicians from varying institutions could provide greater generalizability in the future. As to XAI itself, many experts suggest that the ultimate explanation adapted for human cognition will be in the textual format leveraged by large language models (Mavrepis et al., 2024). They offer easily accessible and comprehensible help without the need for special training for the clinicians to use them. They require low cognitive effort which can both speed up decision making and make the usage more satisfactory (Michalowski et al., 2024). Clever integration of visual and textual explanations might be the ultimate solution allowing for more flexibility and completeness increasing interpretability and usability. It is potentially a solution to account for individual differences between the clinicians and varying levels and domains of expertise in dynamic healthcare settings. Additionally, sentiment analysis presents as an objective and quantitative way to evaluate XAI in terms of user experience which is the next step for XAI implementation in healthcare.

Bibliography

Abell, B., Naicker, S., Rodwell, D., Donovan, T., Tariq, A., Baysari, M., Blythe, R., Parsons, R., & McPhail, S. M. (2023). Identifying barriers and facilitators to successful implementation of computerized clinical decision support systems in hospitals: A NASSS framework-informed scoping review. *Implementation Science*, *18*(1), 32. https://doi.org/10.1186/s13012-023-01287-y

Anjara, S. G., Janik, A., Dunford-Stenger, A., Kenzie, K. M., Collazo-Lorduy, A., Torrente, M., Costabello, L., & Provencio, M. (2023). Examining explainable clinical decision support systems with think aloud protocols. *PLOS ONE*, *18*(9), e0291443. https://doi.org/10.1371/journal.pone.0291443

Anwar, S. S., & Khan, M. M. (2023). *Artificial Intelligence in Healthcare: An Overview*. https://www.semanticscholar.org/paper/Artificial-Intelligence-in-Healthcare%3A-An-Overview-Anwar-Khan/a79c2c3981d026f7ea2189eafd3ec838613e7773

Arbelaez Ossa, L., Starke, G., Lorenzini, G., Vogt, J. E., Shaw, D. M., & Elger, B. S. (2022). Re-focusing explainability in medicine. *DIGITAL HEALTH*, *8*, 20552076221074488. https://doi.org/10.1177/20552076221074488

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2019). *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI* (arXiv:1910.10045). arXiv. http://arxiv.org/abs/1910.10045

Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., & Neves, L. (2020). TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In T. Cohn, Y. He, &

Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1644–1650). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.148

Binder, A., Montavon, G., Bach, S., Müller, K.-R., & Samek, W. (2016). *Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers* (arXiv:1604.00825). arXiv. http://arxiv.org/abs/1604.00825

Boettcher, S. G., & Dethlefsen, C. (2003). deal: A Package for Learning Bayesian Networks. *Journal of Statistical Software*, *8*, 1–40. https://doi.org/10.18637/jss.v008.i20

Cabitza, F., Cameli, M., Campagner, A., Natali, C., & Ronzio, L. (2022a). *Painting the black box white: Experimental findings from applying XAI to an ECG reading setting* (arXiv:2210.15236). arXiv. http://arxiv.org/abs/2210.15236

Cabitza, F., Cameli, M., Campagner, A., Natali, C., & Ronzio, L. (2022b). *Painting the black box white: Experimental findings from applying XAI to an ECG reading setting* (arXiv:2210.15236). arXiv. https://doi.org/10.48550/arXiv.2210.15236

Cabitza, F., Campagner, A., Ronzio, L., Cameli, M., Mandoli, G. E., Pastore, M. C., Sconfienza, L. M., Folgado, D., Barandas, M., & Gamboa, H. (2023). Rams, hounds and white boxes: Investigating human–AI collaboration protocols in medical diagnosis. *Artificial Intelligence in Medicine*, *138*, 102506. https://doi.org/10.1016/j.artmed.2023.102506

Caddick, Z. A., Fraundorf, S. H., Rottman, B. M., & Nokes-Malach, T. J. (2023). Cognitive perspectives on maintaining physicians' medical expertise: II. Acquiring, maintaining, and updating cognitive skills. *Cognitive Research: Principles and Implications*, *8*(1), 47. https://doi.org/10.1186/s41235-023-00497-8

Catalano, M., Bortolotto, C., Nicora, G., Achilli, M., Consonni, A., Ruongo, L., Callea, G., Tito, A., Biasibetti, C., Donatelli, A., Cutti, S., Comotto, F., Stella, G., Corsico, A., Perlini, S., Bellazzi, R., Bruno, R., Filippi, A., & Preda, L. (2023). Performance of an AI algorithm during the different phases of the COVID pandemics: What can we learn from the AI and vice versa. *European Journal of Radiology Open*, *11*, 100497. https://doi.org/10.1016/j.ejro.2023.100497

Catelli, R., Pelosi, S., & Esposito, M. (2022). Lexicon-Based vs. Bert-Based Sentiment Analysis: A Comparative Study in Italian. *Electronics*, *11*(3), Article 3. https://doi.org/10.3390/electronics11030374

Chen, Z., Liang, N., Zhang, H., Li, H., Yang, Y., Zong, X., Chen, Y., Wang, Y., & Shi, N. (2023). Harnessing the power of clinical decision support systems: Challenges and opportunities. *Open Heart*, *10*(2), e002432. https://doi.org/10.1136/openhrt-2023-002432

Choudhury, A. (2022). *Factors Influencing Clinicians' Willingness to Use an AI-Based Clinical Decision Support System* (SSRN Scholarly Paper 4311930). https://papers.ssrn.com/abstract=4311930

Cinà, G., Röber, T., Goedhart, R., & Birbil, I. (2022). *Why we do need Explainable AI for Healthcare* (arXiv:2206.15363). arXiv. https://doi.org/10.48550/arXiv.2206.15363

Clement, T., Kemmerzell, N., Abdelaal, M., & Amberg, M. (2023). XAIR: A Systematic Metareview of Explainable AI (XAI) Aligned to the Software Development Process. *Machine Learning and Knowledge Extraction*, *5*(1), Article 1. https://doi.org/10.3390/make5010006

Combi, C., Amico, B., Bellazzi, R., Holzinger, A., Moore, J. H., Zitnik, M., & Holmes, J. H. (2022). A manifesto on explainability for artificial intelligence in medicine. *Artificial Intelligence in Medicine*, *133*, 102423. https://doi.org/10.1016/j.artmed.2022.102423

Derks, I. P., & de Waal, A. (2020). A Taxonomy of Explainable Bayesian Networks. In A. Gerber (Ed.), *Artificial Intelligence Research* (pp. 220–235). Springer International Publishing. https://doi.org/10.1007/978-3-030-66151-9_14

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Dinos, S., Ascoli, M., Owiti, J. A., & Bhui, K. (2017). Assessing explanatory models and health beliefs: An essential but overlooked competency for clinicians. *BJPsych Advances*, *23*(2), 106–114. https://doi.org/10.1192/apt.bp.114.013680

Directed acyclic graph. (2024). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Directed_acyclic_graph&oldid=1224755243

Dlugatch, R., Georgieva, A., & Kerasidou, A. (2024). AI-driven decision support systems and epistemic reliance: A qualitative study on obstetricians' and midwives' perspectives on integrating AI-driven CTG into clinical decision making. *BMC Medical Ethics*, *25*(1), 6. https://doi.org/10.1186/s12910-023-00990-1

Dobber, J., Harmsen, J., & van Iersel, M. (2023). Background Knowledge in Clinical Reasoning. In J. Dobber, J. Harmsen, & M. van Iersel (Eds.), *Clinical Reasoning and Evidence-Based Practice: Deliberate Decision-Making by Nurses* (pp. 3–40). Springer International Publishing. https://doi.org/10.1007/978-3-031-27069-7_1

Dopp, A. R., Parisi, K. E., Munson, S. A., & Lyon, A. R. (2020). Aligning implementation and user-centered design strategies to enhance the impact of health services: Results from a concept mapping study. *Implementation Science Communications*, *1*(1), 17. https://doi.org/10.1186/s43058-020-00020-w

Eiras-Franco, C., Guijarro-Berdiñas, B., Alonso-Betanzos, A., & Bahamonde, A. (2019). A scalable decision-tree-based method to explain interactions in dyadic data. *Decision Support Systems*, *127*, 113141. https://doi.org/10.1016/j.dss.2019.113141

Elstein, A. S., & Schwarz, A. (2002). Clinical problem solving and diagnostic decision making: Selective review of the cognitive literature. *BMJ*, *324*(7339), 729–732. https://doi.org/10.1136/bmj.324.7339.729

European Comission. (2019). *Ethics guidelines for trustworthy AI | Shaping Europe's digital future*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Comission (ALTAI). (2020, July 17). *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment | Shaping Europe's digital future*. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

Famiglini, L., Campagner, A., Barandas, M., La Maida, G. A., Gallazzi, E., & Cabitza, F. (2024). Evidence-based XAI: An empirical approach to design more effective and explainable decision support systems. *Computers in Biology and Medicine*, *170*, 108042. https://doi.org/10.1016/j.compbiomed.2024.108042

Feretzakis, G., Sakagianni, A., Anastasiou, A., Kapogianni, I., Bazakidou, E., Koufopoulos, P., Koumpouros, Y., Koufopoulou, C., Kaldis, V., & Verykios, V. S. (2024). Integrating

Shapley Values into Machine Learning Techniques for Enhanced Predictions of Hospital Admissions. *Applied Sciences*, *14*(13), Article 13. https://doi.org/10.3390/app14135925

Ferguson, R. L., & Jones, C. H. (1969). A Computer Aided Decision System. *Management Science*, *15*(10), B550–B561.

Fernandez, A., Herrera, F., Cordon, O., Jose del Jesus, M., & Marcelloni, F. (2019). Evolutionary Fuzzy Systems for Explainable Artificial Intelligence: Why, When, What for, and Where to? *IEEE Computational Intelligence Magazine*, *14*(1), 69–81. IEEE Computational Intelligence Magazine. https://doi.org/10.1109/MCI.2018.2881645

Gamborg, M. L., Mehlsen, M., Paltved, C., Vetter, S. S., & Musaeus, P. (2023). Clinical decision-making and adaptive expertise in residency: A think-aloud study. *BMC Medical Education*, *23*(1), 22. https://doi.org/10.1186/s12909-022-03990-8

Gerber, E. (2019, November 28). *A new perspective on Shapley values, part I: Intro to Shapley and SHAP*. Cake or Math: A Data/Science Blog. https://edden-gerber.github.io/shapley-part-1/

Gillies, M., Fiebrink, R., Tanaka, A., Garcia, J., Bevilacqua, F., Heloir, A., Nunnari, F., Mackay, W., Amershi, S., Lee, B., d'Alessandro, N., Tilmanne, J., Kulesza, T., & Caramiaux, B. (2016). Human-Centred Machine Learning. *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 3558–3565. https://doi.org/10.1145/2851581.2856492

Goddard, K., Roudsari, A., & Wyatt, J. C. (2014). Automation bias: Empirical results assessing influencing factors. *International Journal of Medical Informatics*, *83*(5), 368–375. https://doi.org/10.1016/j.ijmedinf.2014.01.001

Gupta, J., & Seeja, K. R. (2024). A Comparative Study and Systematic Analysis of XAI Models and their Applications in Healthcare. *Archives of Computational Methods in Engineering*. https://doi.org/10.1007/s11831-024-10103-9

Herm, L.-V. (2023, April 18). *Impact Of Explainable AI On Cognitive Load: Insights From An Empirical Study*. arXiv.Org. https://arxiv.org/abs/2304.08861v1

Hoffman, R., Mueller, S., Klein, G., & Litman, J. (2018). *Metrics for Explainable AI: Challenges and Prospects*.

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

Hudda, S., Kumar, Dr. R., & Negi, Dr. N. (2024). The Changing Landscape of Healthcare with State of the Art AI Technology. *International Journal for Research in Applied Science and Engineering Technology*, *12*(5), 1700–1705. https://doi.org/10.22214/ijraset.2024.61723

Imaduddin, H., A'la, F. Y., & Nugroho, Y. S. (2023). Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *14*(8), Article 8. https://doi.org/10.14569/IJACSA.2023.0140813

Introducing HealthAI. (2023). I-DAIR. https://www.i-dair.org/news/introducing-healthai

Jones, C., Thornton, J., & Wyatt, J. C. (2021). Enhancing trust in clinical decision support systems: A framework for developers. *BMJ Health & Care Informatics*, *28*(1). https://doi.org/10.1136/bmjhci-2020-100247

Kahneman, D. (2011). *Thinking, fast and slow* (p. 499). Farrar, Straus and Giroux.

Kazim, T., & Tomlinson, J. (2023). Automation Bias and the Principles of Judicial Review. *Judicial Review*, *28*(1), 9–16. https://doi.org/10.1080/10854681.2023.2189405

Kim, A., Yang, M., & Zhang, J. (2020). *When Algorithms Err: Differential Impact of Early vs. Late Errors on Users' Reliance on Algorithms* (SSRN Scholarly Paper 3691575). https://doi.org/10.2139/ssrn.3691575

Kitson, N. K., Constantinou, A. C., Guo, Z., Liu, Y., & Chobtham, K. (2023). A survey of Bayesian Network structure learning. *Artificial Intelligence Review*, *56*(8), 8721–8814. https://doi.org/10.1007/s10462-022-10351-w

Kong, X., Liu, S., & Zhu, L. (2024). Toward Human-centered XAI in Practice: A survey. *Machine Intelligence Research*, *21*(4), 740–770. https://doi.org/10.1007/s11633-022-1407-3

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). 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*, 103473. https://doi.org/10.1016/j.artint.2021.103473

Leighton, J. P. (2017). *Using Think-Aloud Interviews and Cognitive Labs in Educational Research*. Oxford University Press.

Leong, T.-Y. (2003). Decision Support Systems in Healthcare: Emerging Trends and Success Factors. In X. Yu & J. Kacprzyk (Eds.), *Applied Decision Support with Soft Computing* (pp. 151–179). Springer. https://doi.org/10.1007/978-3-540-37008-6_6

Lesley, U., & Kuratomi Hernández, A. (2024). Improving XAI Explanations for Clinical Decision-Making – Physicians' Perspective on Local Explanations in Healthcare. In J. Finkelstein, R. Moskovitch, & E. Parimbelli (Eds.), *Artificial Intelligence in Medicine* (pp. 296–312). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-66535-6_32

Liao, Q. V., & Varshney, K. R. (2022). *Human-Centered Explainable AI (XAI): From Algorithms to User Experiences* (arXiv:2110.10790). arXiv. http://arxiv.org/abs/2110.10790

Lipton, Z. C. (2017). *The Mythos of Model Interpretability* (arXiv:1606.03490). arXiv. https://doi.org/10.48550/arXiv.1606.03490

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. http://arxiv.org/abs/1907.11692

Longo, L., Goebel, R., Lecue, F., Kieseberg, P., & Holzinger, A. (2020). Explainable Artificial Intelligence: Concepts, Applications, Research Challenges and Visions. In A. Holzinger, P. Kieseberg, A. M. Tjoa, & E. Weippl (Eds.), *Machine Learning and Knowledge Extraction* (pp. 1–16). Springer International Publishing. https://doi.org/10.1007/978-3-030-57321-8_1

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777.

Maleki Varnosfaderani, S., & Forouzanfar, M. (2024). The Role of AI in Hospitals and Clinics: Transforming Healthcare in the 21st Century. *Bioengineering*, *11*(4), 337. https://doi.org/10.3390/bioengineering11040337

Manresa-Yee, C., Roig-Maimó, M. F., Ramis, S., & Mas-Sansó, R. (2022). Advances in XAI: Explanation Interfaces in Healthcare. In C.-P. Lim, Y.-W. Chen, A. Vaidya, C. Mahorkar, & L. C. Jain (Eds.), *Handbook of Artificial Intelligence in Healthcare: Vol 2: Practicalities and Prospects* (pp. 357–369). Springer International Publishing. https://doi.org/10.1007/978-3-030-83620-7_15

Material Design. (n.d.). Material Design. Retrieved 16 August 2024, from https://m3.material.io

Mavrepis, P., Makridis, G., & Fatouros, G. (2024). *XAI for All: Can Large Language Models Simplify Explainable AI?* https://arxiv.org/html/2401.13110v1

Miah, M. S. U., Kabir, M. M., Sarwar, T. B., Safran, M., Alfarhood, S., & Mridha, M. F. (2024). A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. *Scientific Reports*, *14*(1), 9603. https://doi.org/10.1038/s41598-024-60210-7

Michalowski, M., Wilk, S., Bauer, J. M., Carrier, M., Delluc, A., Le Gal, G., Wang, T.-F., Siegal, D., & Michalowski, W. (2024). Manually-Curated Versus LLM-Generated Explanations for Complex Patient Cases: An Exploratory Study with Physicians. In J. Finkelstein, R. Moskovitch, & E. Parimbelli (Eds.), *Artificial Intelligence in Medicine* (pp. 313–323). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-66535-6_33

Nicora, G., Tito, A. L., Donatelli, A., Callea, G., Biasibetti, C., Galli, M., Comotto, F., Bortolotto, C., Perlini, S., Preda, L., & Bellazzi, R. (2021). *ALFABETO: Supporting COVID-19 Hospital Admissions with Bayesian Networks*. SMARTERCARE@AI*IA.

https://www.semanticscholar.org/paper/ALFABETO%3A-Supporting-COVID-19-Hospital-Admissions-Nicora-Tito/1d69c25387a21bb3ca5518d9891e156292eb84d0

Noushad, B., Van Gerven, P. W. M., & de Bruin, A. B. H. (2024). Twelve tips for applying the think-aloud method to capture cognitive processes. *Medical Teacher*, *46*(7), 892–897. https://doi.org/10.1080/0142159X.2023.2289847

Parimbelli, E., Nicora, G., Wilk, S., Michalowski, W., & Bellazzi, R. (2023). Tree-based local explanations of machine learning model predictions, AraucanaXAI. *Artificial Intelligence in Medicine*, *135*, 102471. https://doi.org/10.1016/j.artmed.2022.102471

Pearl, J. (2018). Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 3. https://doi.org/10.1145/3159652.3176182

Press, A., McCullagh, L., Khan, S., Schachter, A., Pardo, S., & McGinn, T. (2015). Usability Testing of a Complex Clinical Decision Support Tool in the Emergency Department: Lessons Learned. *JMIR Human Factors*, *2*(2), e4537. https://doi.org/10.2196/humanfactors.4537

Qi, Y., & Shabrina, Z. (2023). Sentiment analysis using Twitter data: A comparative application of lexicon- and machine-learning-based approach. *Social Network Analysis and Mining*, *13*(1), 31. https://doi.org/10.1007/s13278-023-01030-x

Reverberi, C., Rigon, T., Solari, A., Hassan, C., Cherubini, P., GI Genius CADx Study Group, Antonelli, G., Awadie, H., Bernhofer, S., Carballal, S., Dinis-Ribeiro, M., Fernández-Clotett, A., Esparrach, G. F., Gralnek, I., Higasa, Y., Hirabayashi, T., Hirai, T., Iwatate, M., Kawano, M., … Cherubini, A. (2022). Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific Reports*, *12*(1), 14952. https://doi.org/10.1038/s41598-022-18751-2

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *'Why Should I Trust You?': Explaining the Predictions of Any Classifier* (arXiv:1602.04938). arXiv. https://doi.org/10.48550/arXiv.1602.04938

RoBERTa: An optimized method for pretraining self-supervised NLP systems. (n.d.). Retrieved 28 August 2024, from https://ai.meta.com/blog/roberta-an-optimized-method-for-pretraining-self-supervised-nlp-systems/

Sagar, K., & Saha, A. (2017). A systematic review of software usability studies. *International Journal of Information Technology*. https://doi.org/10.1007/s41870-017-0048-1

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Schneeberger, D., Rottger, R., Cabitza, F., Campagner, A., Plass, M., Muller, H., & Holzinger, A. (2023). *The Tower of Babel in Explainable Artificial Intelligence (XAI)*. Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-3-031-40837-3_5

Schwalbe, G., & Finzel, B. (2023). A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*. https://doi.org/10.1007/s10618-022-00867-8

Setzu, M., Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., & Giannotti, F. (2021). GLocalX -- From Local to Global Explanations of Black Box AI Models. *Artificial Intelligence*, *294*, 103457. https://doi.org/10.1016/j.artint.2021.103457

Severes, B., Carreira, C., Vieira, A. B., Gomes, E., Aparício, J. T., & Pereira, I. (2023). The Human Side of XAI: Bridging the Gap between AI and Non-expert Audiences. *Proceedings of the 41st ACM International Conference on Design of Communication*, 126–132. https://doi.org/10.1145/3615335.3623062

Shah, N. H., Halamka, J. D., Saria, S., Pencina, M., Tazbaz, T., Tripathi, M., Callahan, A., Hildahl, H., & Anderson, B. (2024). A Nationwide Network of Health AI Assurance Laboratories. *JAMA*, *331*(3), 245–249. https://doi.org/10.1001/jama.2023.26930

Shin, H. S. (2019). Reasoning processes in clinical reasoning: From the perspective of cognitive psychology. *Korean Journal of Medical Education*, *31*(4), 299–308. https://doi.org/10.3946/kjme.2019.140

Silcox, C., Zimlichmann, E., Huber, K., Rowen, N., Saunders, R., McClellan, M., Kahn, C. N., Salzberg, C. A., & Bates, D. W. (2024). The potential for artificial intelligence to transform healthcare: Perspectives from international health leaders. *Npj Digital Medicine*, *7*(1), 1–3. https://doi.org/10.1038/s41746-024-01097-6

Speith, T. (2022). A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2239–2250. https://doi.org/10.1145/3531146.3534639

Sundararajan, M., Taly, A., & Yan, Q. (2017). *Axiomatic Attribution for Deep Networks* (arXiv:1703.01365). arXiv. https://doi.org/10.48550/arXiv.1703.01365

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1016/0364-0213(88)90023-7

Tahtali, M. A., Snijders, C., & Dirne, C. (2024). Trust in Algorithmic Advice Increases with Task Complexity. In J. Baratgin, B. Jacquet, & H. Yama (Eds.), *Human and Artificial*

Rationalities (pp. 86–106). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-55245-8_6

Talati, D. (2023). AI in healthcare domain. *Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (Online)*, *2*(3), Article 3. https://doi.org/10.60087/jklst.vol2.n3.p262

Ter-Minassian, L., Ghalebikesabi, S., Diaz-Ordaz, K., & Holmes, C. (2023). *Challenges and Opportunities of Shapley values in a Clinical Context* (arXiv:2306.14698). arXiv. https://doi.org/10.48550/arXiv.2306.14698

Thompson, S., Moorley, C., & Barratt, J. (2017). A comparative study on the clinical decision-making processes of nurse practitioners vs. Medical doctors using scenarios in a secondary care environment. *Journal of Advanced Nursing*, *73*(5), 1097–1110. https://doi.org/10.1111/jan.13206

Trevan, L., Parker, V., & Thoumi, A. (2022). *Preventing Bias and Inequities in AI-Enabled Health Tools*. https://healthpolicy.duke.edu/publications/preventing-bias-and-inequities-ai-enabled-health-tools

Troya, J., Fitting, D., Brand, M., Sudarevic, B., Kather, J. N., Meining, A., & Hann, A. (2022). The influence of computer-aided polyp detection systems on reaction time for polyp detection and eye gaze. *Endoscopy*, *54*, 1009–1014. https://doi.org/10.1055/a-1770-7353

Tsalatsanis, A., Hozo, I., Kumar, A., & Djulbegovic, B. (2015). Dual Processing Model for Medical Decision-Making: An Extension to Diagnostic Testing. *PLOS ONE*, *10*(8), e0134800. https://doi.org/10.1371/journal.pone.0134800

VADER-Sentiment-Analysis Introduction—VaderSentiment 3.3.1 documentation. (n.d.). Retrieved 13 August 2024, from https://vadersentiment.readthedocs.io/en/latest/pages/introduction.html

Vale, D., El-Sharif, A., & Ali, M. (2022). Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law. *AI and Ethics*, *2*(4), 815–826. https://doi.org/10.1007/s43681-022-00142-y

van Dongen, K., & van Maanen, P.-P. (2013). A framework for explaining reliance on decision aids. *International Journal of Human-Computer Studies*, *71*(4), 410–424. https://doi.org/10.1016/j.ijhcs.2012.10.018

van Lent, M., Fisher, W., & Mancuso, M. (2004). *An Explainable Artificial Intelligence System for Small-unit Tactical Behavior*.

Wang, Y., Lang, J., Zuo, J. Z., Dong, Y., Hu, Z., Xu, X., Zhang, Y., Wang, Q., Yang, L., Wong, S. T. C., Wang, H., & Li, H. (2022). The radiomic-clinical model using the SHAP method for assessing the treatment response of whole-brain radiotherapy: A multicentric study. *European Radiology*, *32*(12), 8737–8747. https://doi.org/10.1007/s00330-022-08887-0

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731–5780. https://doi.org/10.1007/s10462-022-10144-1

What is a Decision Tree? | IBM. (2021, November 2). https://www.ibm.com/topics/decision-trees

What is Usability—The Ultimate Guide. (n.d.). The Interaction Design Foundation. Retrieved 15 August 2024, from https://www.interaction-design.org/literature/topics/usability

World Health Organization. (2023). *WHO outlines considerations for regulation of artificial intelligence for health*. https://www.who.int/news/item/19-10-2023-who-outlines-considerations-for-regulation-of-artificial-intelligence-for-health

Yang, W., Wei, Y., Wei, H., Chen, Y., Huang, G., Li, X., Li, R., Yao, N., Wang, X., Gu, X., Amin, M. B., & Kang, B. (2023). Survey on Explainable AI: From Approaches, Limitations and Applications Aspects. *Human-Centric Intelligent Systems*, *3*(3), 161–188. https://doi.org/10.1007/s44230-023-00038-y

Zhang, D., & Adipat, B. (2005). Challenges, Methodologies, and Issues in the Usability Testing of Mobile Applications. *International Journal of Human–Computer Interaction*, *18*(3), 293–308. https://doi.org/10.1207/s15327590ijhc1803_3

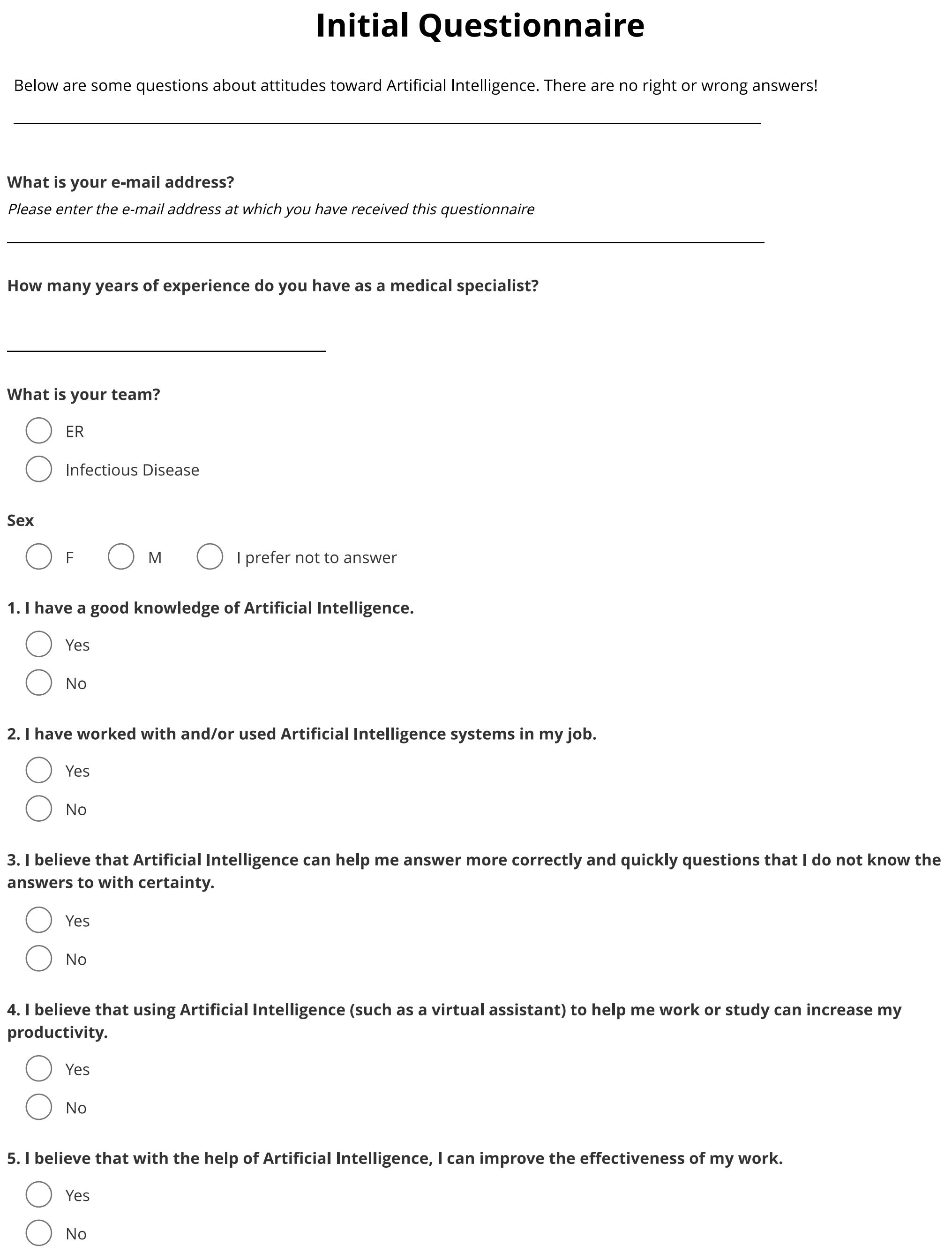
Zhu, X., Gardiner, S., Roldán, T., & Rossouw, D. (2024). *The Model Arena for Cross-lingual Sentiment Analysis: A Comparative Study in the Era of Large Language Models* (arXiv:2406.19358; Version 1). arXiv. https://doi.org/10.48550/arXiv.2406.19358

Zytek, A., Liu, D., Vaithianathan, R., & Veeramachaneni, K. (2021). *Sibyl: Understanding and Addressing the Usability Challenges of Machine Learning In High-Stakes Decision Making* (arXiv:2103.02071). arXiv. http://arxiv.org/abs/2103.02071

Appendix

1.1 Initial Questionnaire

The following figure represents the initial questionnaire structure and layout (Figure 38). It was utilized to collect demographical data and general attitudes towards AI in healthcare.



Initial Questionnaire

Below are some questions about attitudes toward Artificial Intelligence. There are no right or wrong answers!

____

What is your e-mail address?

Please enter the e-mail address at which you have received this questionnaire

____

How many years of experience do you have as a medical specialist?

____

What is your team?

○ ER

○ Infectious Disease

Sex

○ F ○ M ○ I prefer not to answer

1. I have a good knowledge of Artificial Intelligence.

○ Yes

○ No

2. I have worked with and/or used Artificial Intelligence systems in my job.

○ Yes

○ No

3. I believe that Artificial Intelligence can help me answer more correctly and quickly questions that I do not know the answers to with certainty.

○ Yes

○ No

4. I believe that using Artificial Intelligence (such as a virtual assistant) to help me work or study can increase my productivity.

○ Yes

○ No

5. I believe that with the help of Artificial Intelligence, I can improve the effectiveness of my work.

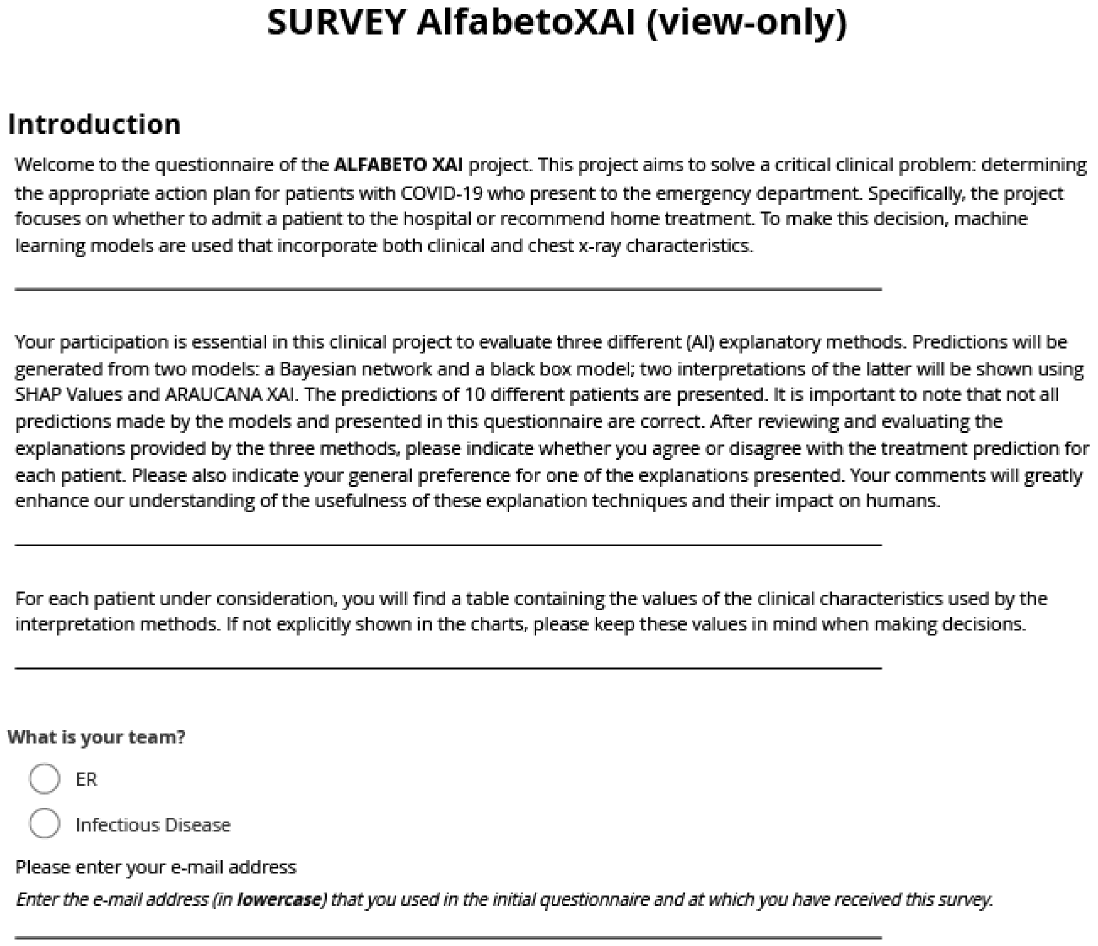
○ Yes

○ No

Figure 38. Initial questionnaire

1.2 Survey Questionnaire

The following series of figures represent the introduction and the example of the layout of one patient and corresponding explanations – Shap, Araucana and Bayesian network (Figure 39).



SURVEY AlfabetoXAI (view-only)

Introduction

Welcome to the questionnaire of the **ALFABETO XAI** project. This project aims to solve a critical clinical problem: determining the appropriate action plan for patients with COVID-19 who present to the emergency department. Specifically, the project focuses on whether to admit a patient to the hospital or recommend home treatment. To make this decision, machine learning models are used that incorporate both clinical and chest x-ray characteristics.

Your participation is essential in this clinical project to evaluate three different (AI) explanatory methods. Predictions will be generated from two models: a Bayesian network and a black box model; two interpretations of the latter will be shown using SHAP Values and ARAUCANA XAI. The predictions of 10 different patients are presented. It is important to note that not all predictions made by the models and presented in this questionnaire are correct. After reviewing and evaluating the explanations provided by the three methods, please indicate whether you agree or disagree with the treatment prediction for each patient. Please also indicate your general preference for one of the explanations presented. Your comments will greatly enhance our understanding of the usefulness of these explanation techniques and their impact on humans.

For each patient under consideration, you will find a table containing the values of the clinical characteristics used by the interpretation methods. If not explicitly shown in the charts, please keep these values in mind when making decisions.

What is your team?

○ ER

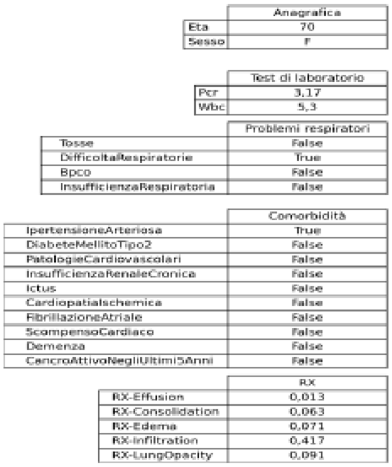
○ Infectious Disease

Please enter your e-mail address

*Enter the e-mail address (in **lowercase**) that you used in the initial questionnaire and at which you have received this survey.*

Patient 1

Patient's characteristics:



	Anagrafica
Eta	70
Sesso	F

	Test di laboratorio
Pcr	3,17
Wbc	5,3

	Problemi respiratori
Tosse	False
DifficoltaRespiratorie	True
Bpco	False
InsufficienzaRespiratoria	False

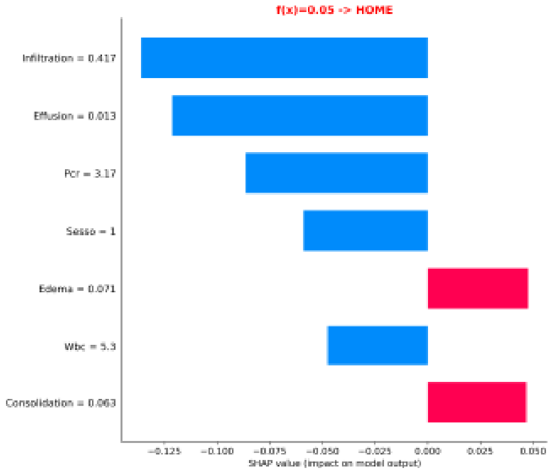
	Comorbidità
IpertensioneArteriosa	True
DiabeteMellitoTipo2	False
PatologieCardiovascolari	False
InsufficienzaRenaleCronica	False
Ictus	False
CardiopatiaIschemica	False
FibrillazioneAtriale	False
ScompensoCardiaco	False
Demenza	False
CancroAttivoNegliUltimi5Anni	False

	RX
RX-Effusion	0,013
RX-Consolidation	0,063
RX-Edema	0,071
RX-Infiltration	0,417
RX-LungOpacity	0,091

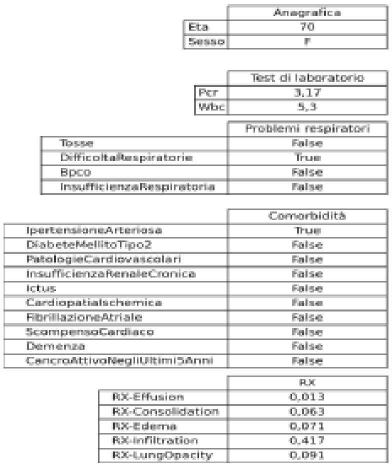
» Explanation 1

SHAP Values explanation

SHAP Values



Patient's characteristics:



	Anagrafica
Eta	70
Sesso	F

	Test di laboratorio
Pcr	3,17
Wbc	5,3

	Problemi respiratori
Tosse	False
DifficoltaRespiratorie	True
Bpco	False
InsufficienzaRespiratoria	False

	Comorbidità
IpertensioneArteriosa	True
DiabeteMellitoTipo2	False
PatologieCardiovascolari	False
InsufficienzaRenaleCronica	False
Ictus	False
CardiopatiaIschemica	False
FibrillazioneAtriale	False
ScompensoCardiaco	False
Demenza	False
CancroAttivoNegliUltimi5Anni	False

	RX
RX-Effusion	0,013
RX-Consolidation	0,063
RX-Edema	0,071
RX-Infiltration	0,417
RX-LungOpacity	0,091

1. I find the explanation intuitively understandable.
1-Strongly disagree, 6-Strongly agree

1 6

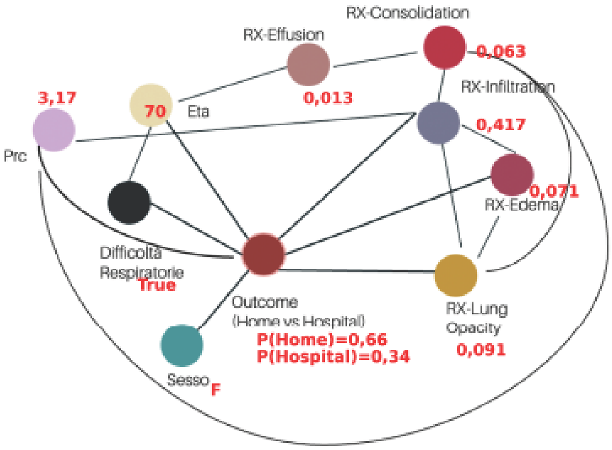
2. The explanation helps me to make a proper decision about the case at hand.
1-Strongly disagree, 6-Strongly agree

1 6

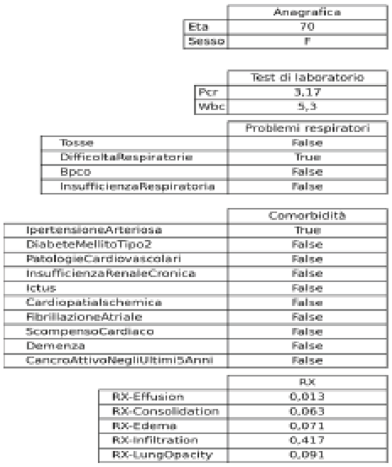
» Explanation 2

Bayesian Network explanation

Bayesian Network



Patient's characteristics:



	Anagrafica
Eta	70
Sesso	F

	Test di laboratorio
Pcr	3,17
Wbc	5,3

	Problemi respiratori
Tosse	False
DifficoltaRespiratorie	True
Bpco	False
InsufficienzaRespiratoria	False

	Comorbidità
IpertensioneArteriosa	True
DiabeteMellitoTipo2	False
PatologieCardiovascolari	False
InsufficienzaRenaleCronica	False
Ictus	False
CardiopatiaIschemica	False
FibrillazioneAtriale	False
ScompensoCardiaco	False
Demenza	False
CancroAttivoNegliUltimi5Anni	False

	RX
RX-Effusion	0,013
RX-Consolidation	0,063
RX-Edema	0,071
RX-Infiltration	0,417
RX-LungOpacity	0,091

1. I find the explanation intuitively understandable.

1-Strongly disagree, 6-Strongly agree

1					6

2. The explanation helps me to make a proper decision about the case at hand.

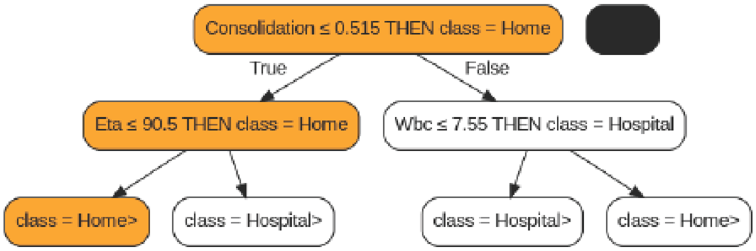
1-Strongly disagree, 6-Strongly agree

1 6

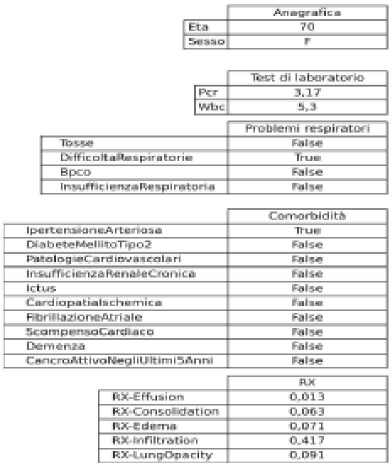
» Explanation 3

ARAUCANA XAI explanation

ARAUCANA XAI


Consolidation ≤ 0.515 THEN class = Home
True
False
Eta ≤ 90.5 THEN class = Home
Wbc ≤ 7.55 THEN class = Hospital
class = Home>
class = Hospital>
class = Hospital>
class = Home>

Patient's characteristics:



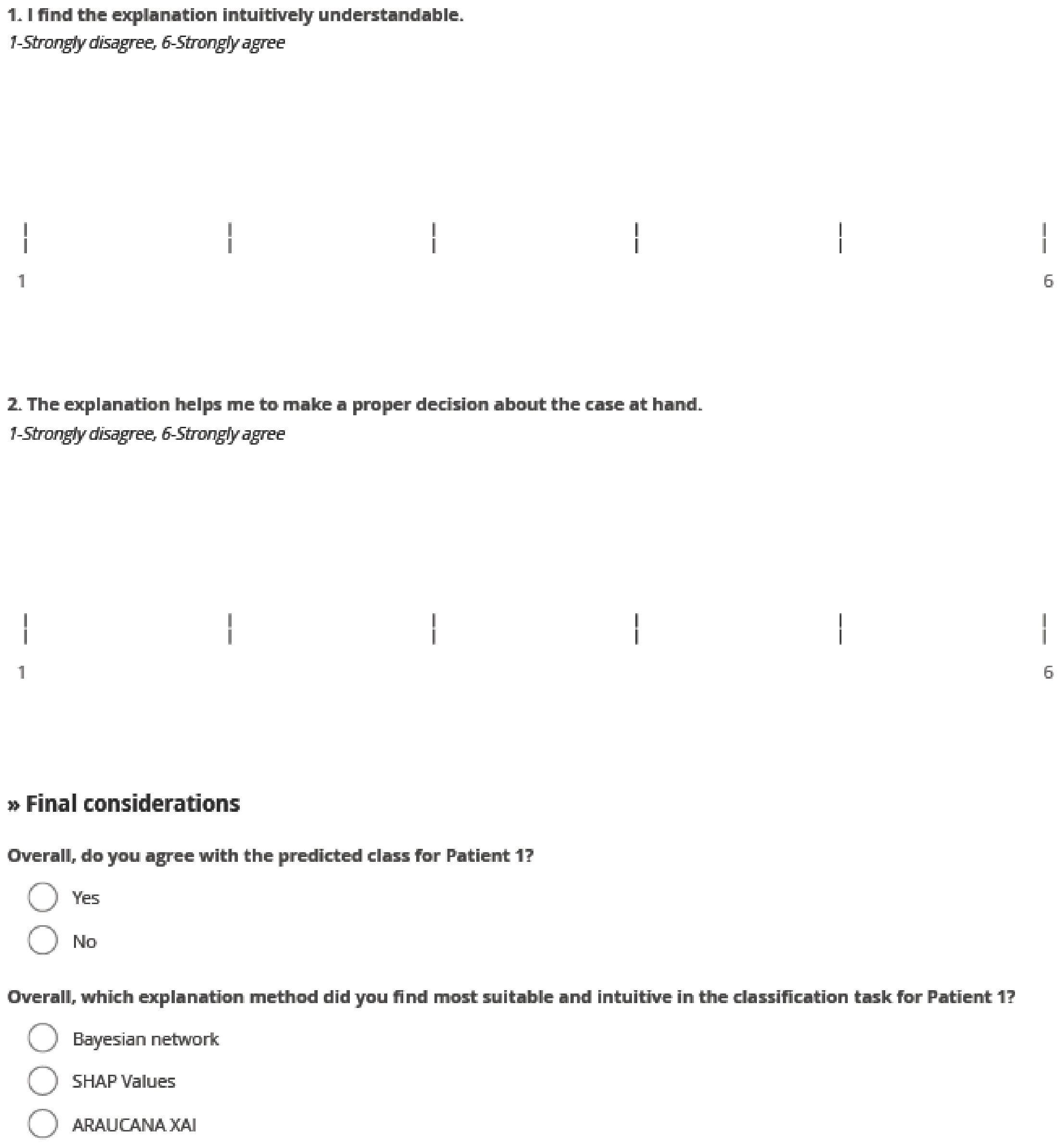
	Anagrafica
Eta	70
Sesso	F

	Test di laboratorio
Pcr	3,17
Wbc	5,3

	Problemi respiratori
Tosse	False
DifficoltaRespiratorie	True
Bpco	False
InsufficienzaRespiratoria	False

	Comorbidità
IpertensioneArteriosa	True
DiabeteMellitoTipo2	False
PatologieCardiovascolari	False
InsufficienzaRenaleCronica	False
Ictus	False
CardiopatiaIschemica	False
FibrillazioneAtriale	False
ScompensoCardiaco	False
Demenza	False
CancroAttivoNegliUltimi5Anni	False

	RX
RX-Effusion	0,013
RX-Consolidation	0,063
RX-Edema	0,071
RX-Infiltration	0,417
RX-LungOpacity	0,091



1. I find the explanation intuitively understandable.

1-Strongly disagree, 6-Strongly agree

2. The explanation helps me to make a proper decision about the case at hand.

1-Strongly disagree, 6-Strongly agree

» Final considerations

Overall, do you agree with the predicted class for Patient 1?

○ Yes

○ No

Overall, which explanation method did you find most suitable and intuitive in the classification task for Patient 1?

○ Bayesian network

○ SHAP Values

○ ARAUCANA XAI

Figure 39. Survey layout example