A User Study Evaluating Argumentative Explanations in Diagnostic Decision Support

Felix Liedeker[a,*], Olivia Sanchez-Graillet[a], Moana Seidler[c], Christian Brandt[b], Jörg Wellmer[c] and Philipp Cimiano[a]

[a]Semantic Computing Group, CITEC, Bielefeld University, Bielefeld, Germany
[b]Bielefeld University, Medical School and University Medical Center OWL, Mara Hospital, Department of Epileptology, Bielefeld, Germany
[c]Ruhr-Epileptology, University Hospital Knappschaftskrankenhaus Bochum, Ruhr-University, Bochum, Germany
ORCID (Felix Liedeker): https://orcid.org/0009-0006-2556-9430, ORCID (Olivia Sanchez-Graillet): https://orcid.org/0000-0003-3483-265X, ORCID (Christian Brandt): https://orcid.org/0000-0001-8666-1640, ORCID (Jörg Wellmer): https://orcid.org/0000-0003-2919-0496, ORCID (Philipp Cimiano): https://orcid.org/0000-0002-4771-441X

Abstract. As the field of healthcare increasingly adopts artificial intelligence, it becomes important to understand which types of explanations increase transparency and empower users to develop confidence and trust in the predictions made by machine learning (ML) systems. In shared decision-making scenarios where doctors cooperate with ML systems to reach an appropriate decision, establishing mutual trust is crucial. In this paper, we explore different approaches to generating explanations in eXplainable AI (XAI) and make their underlying arguments explicit so that they can be evaluated by medical experts. In particular, we present the findings of a user study conducted with physicians to investigate their perceptions of various types of AI-generated explanations in the context of diagnostic decision support. The study aims to identify the most effective and useful explanations that enhance the diagnostic process. In the study, medical doctors filled out a survey to assess different types of explanations. Further, an interview was carried out post-survey to gain qualitative insights on the requirements of explanations incorporated in diagnostic decision support. Overall, the insights gained from this study contribute to understanding the types of explanations that are most effective.

1 Introduction

The use of ML models in medical decision support systems has been increasing recently [23]. Since false predictions of such systems can potentially have detrimental outcomes for humans, the need for explainability in ML-based decision support in medical systems has been widely recognised [44, 3]. Explainability is a crucial element of trust [18]. Furthermore, explainability can play a key role in reducing biases and promoting a collaborative environment where AI and human decision-makers can work together seamlessly to enhance decision-making [6].

However, while research on eXplainable AI (XAI) has gained momentum, a precise definition of explainability and the requirements that make an explanation fit for purpose are lacking [46, 4]. This

poses a significant challenge for developing AI systems because there is very limited guidance on how explanations should be designed to empower a human user to gain confidence in a prediction.

Despite the acknowledgement of the diverse needs of stakeholders regarding XAI [28], empirical evaluations of these approaches are relatively rare [3]. Addressing this lack of empirical evidence is essential though to validate and refine XAI techniques.

It is precisely this gap that we are addressing with our work by conducting a user study with medical experts to evaluate the assessment of different XAI approaches. In particular, we rely on different methods to generate arguments that make the reasons explicit why a model made a certain prediction. For this, we define templates and instantiate these templates using standard XAI approaches, feature importance, and counterfactual explanations in particular.

The research questions we would like to answer are:

- *RQ1* How do users evaluate explanations generated by state-of-the-art XAI approaches?
- *RQ2* Are there differences in the way the different XAI approaches are evaluated?
- *RQ3* Would doctors rely on the XAI-generated explanations to explain their decision to a medical colleague?
- *RQ4* Which type of explanations would the experts use and how do they compare to XAI-generated explanations?

The above questions are designed to evaluate the level of confidence that medical users place into XAI-generated explanations and how they differ from the explanations they would give.

We present the results of a user study involving eight medical experts in the field of neurology and epileptology from two hospitals. The user study was administered by way of a questionnaire relying on real patient cases. After receiving the questionnaire results, a debriefing was conducted to generate qualitative findings on how the explanations were perceived.

The remainder of this paper is organized as follows: Section 2 provides some background on XAI methods. Section 3 discusses related work on argumentation and XAI. Section 4 presents a real-world use

---

* Corresponding Author. Email: fliedeker@techfak.uni-bielefeld.de.

case demonstrating the creation of argument-based explanations in natural language using the output of different XAI methods. This section also includes the evaluation of the generated explanations through a survey applied to expert users. In Section 5, we discuss the results of our methods and the user survey. Section 6 presents the conclusions and future perspectives of our work.

2 Background

The current state of XAI largely revolves around the explanation of black-box models by auxiliary methods - so-called post-hoc explanations [4]. Beyond that, inherently interpretable white-box models that do not require auxiliary explanation methods are frequently used. Well-known examples of white-box models used (in the medical domain) are decision trees [40] as well as Bayesian Networks (BN) [38].

In the area of post-hoc explainability, feature attribution methods and counterfactual explanations (CFs) are among the most frequently used methods. Feature attribution methods like SHAP (SHapley Additive exPlanations) [32] or LIME (Local Interpretable Model-agnostic Explanations) [41] provide insights into the contributions of individual features to the predictions of ML models, enhancing interpretability by explaining model behaviour in a comprehensible manner. CFs provide indirect explanations through hypothetical, but similar counterexamples in the form of *If X had been different, the prediction would have changed from P to Q* [21]. CFs are especially popular because they resemble the way humans tend to explain decisions [49].

So far, the XAI community has not devoted too much focus to the fact that explanations are supposed to increase confidence and empower humans. Studies evaluating the impact of explanations on users in terms of enhancing understanding, increasing confidence, or empowering users to make better decisions are rare, see e.g. [34, 2] for some exceptions. Meske and Bunde [34] showed that XAI is able to increase perceived usefulness, ease of use, and intention to use of users. While results were not conclusive, the general potential to improve human accuracy in decision-making by leveraging XAI was shown by Alufaisan et al. [2]. A recent study by Hamm et al. [16] found that XAI, if resulting in a high degree of perceived explainability, can lead to positive effects on perceived usefulness and trust.

Although XAI has become an important and active field of research, only limited attention has been paid to the question of what kind of explanations are actually most useful and under what circumstances. Approaches to overcome this shortcoming consist, for example, of investigating which general desiderata for explanations can be found [35] or in examining various characteristics of explanations, such as the representation form of an explanation [48]. Furthermore, it has been shown that the generation process, i.e. the method used to generate an explanation, affects the user perception of such explanations [51]. This realm also includes investigations of explanation complexity, e.g. focusing on complexity imposed by constraints of different generation methods [1], the presentation complexity of explanations [20], or the overall explanation complexity due to explanation type and feature count within explanations [29]. Results generally suggest the existence of a 'sweet spot' between the complexity of an explanation and its value, quantified in terms of enhanced understandability or usefulness.

While a vast amount of different metrics have been proposed for automatically evaluating explanations, the gold standard for explanation evaluation is still the evaluation by humans [8], which underlines the importance of empirical user studies.

In the recent past, there has been a movement from the 'classical', static, one-shot post-hoc explanations towards more human-centred and interactive XAI approaches. Rohlfing et al. [43] proposed a "co-constructive" approach to decision-making and explanation giving, aiming for a high level of interactivity and thus explainable decisions that are not dictated by an AI system but reached in close interaction between human and machine. Miller [36] argued for a paradigm shift towards "evaluative AI", proposing the superiority of machine-in-the-loop decision support which not only provides recommendations to a human user, but rather provides evidence supporting and challenging decisions, allowing the user to evaluate the evidence and taking a decision on this basis.

We believe that this offers a great opportunity, particularly in the area of decision support in medicine, to overcome users' reservations such as perceiving AI systems as a potential threat to their own occupation [27] and raising awareness for the often underestimated risk of making errors in diagnostic decisions [14] by not only keeping the "doctor-in-the-loop" [17], but moving towards interactive, co-constructive decision support.

3 Related Work

Explainability of AI systems aims to trace the steps that lead to decisions or conclusions. Explainability was relatively straightforward as early AI systems were logic-based, allowing transparency into how decisions were reached. However, as AI methods advanced, they became less transparent. XAI was then developed to provide transparency through explainability to these complex systems. In tasks that require explanations, such as decision-making, justification, and dialogues across various fields of knowledge, argumentation can be very useful [47].

Examples of argumentation frameworks that have been used to generate explanations include the Abstract Argumentation Framework (AAF) and the Assumption-Based Argumentation Framework (ABA) [12, 11] which can be applied across different areas.

In recent years, XAI has integrated argumentation due to its strong capability to provide explanations. Argumentation can enable explainability by translating an AI system's decision into a step-by-step argumentation procedure that demonstrates the rationale behind the final decision [47]. For example, a given set of possible decisions can be mapped to a graphical representation with predefined attack/support elements that subsequently lead to the accepted decision, and show the steps followed.

In the field of medicine and healthcare, dealing with incomplete and contradictory information is necessary, and previous conclusions must be questioned when new evidence appears. Thus, argumentation is an important approach in developing applications that support decision-making by generating rationalized recommendations or explanations. There is a wide range of applications using argumentation in this field, including prognosis, diagnosis, planning and implementation of (pre)clinical trials, preparation of medical guidelines, and patient counselling. Argumentation-based techniques have been applied in the medical context since the mid-1980s, utilizing different representations and formalisms such as logic-based argumentation, abstract argumentation, argumentation schemes, and defeasible logic programming.

For example, Krause, Fox, and Judson [25] used the logic of argumentation (LA) [26], which is a variant of first-order logic, in an application for qualitative reasoning in toxicology (StAR). The system provides arguments both for and against the compound being carcinogenic for each toxic alert. StAR showed that by using

an argumentation approach, it was possible to offer valuable advice even when statistical information was not available. Hunter and Williams [19] established an argument-based framework based on Dung's model [9] for aggregating clinical trial evidence. This approach involves generating inductive arguments from evidence found in medical guidelines, and the superiority of interventions is determined based on specific preference criteria applied to these arguments. The method was assessed based on how well the results of the use cases aligned with those in published clinical guidelines.

Grando et al. [15] developed rules based on argument schemes and their critical questions to explain anomalous patient responses to treatments. They utilized the ASPIC platform [39], an implementation of Dung's model, to formalize and assess the acceptance status of the arguments. To evaluate their method, the authors compared the results provided by their method with a previous version that offered fewer explanations or justifications. Then, clinicians were asked to indicate their preferred presentation. Interestingly, in certain situations, clinicians preferred fewer explanations and textual over graphical ones. More recently, defeasible reasoning [31] and dialogue frameworks [33, 45] have been used in applications for the medical and healthcare field [13, 54].

Nevertheless, there may be limitations when evaluating argumentation-based methods. For instance, some methods are exploratory and lack comparison with similar approaches. Other methods focus on very specific applications, making it difficult to find comparable methods or results for comparison. Abstract representations are not always verbalized, which makes it difficult for humans to evaluate them. Most existing argumentation-based approaches do not offer a quantitative evaluation against a given gold standard of explanations due to the lack of gold standards for specific medical areas or goals. Finally, as Longo and Dondio [30] pointed out, some approaches based on refutable reasoning and argumentation theory in the medical field rely on anecdotal evidence, which characterizes several use cases.

In spite of the fact that XAI and computational argumentation are conceptually close, it is only recently that work has emerged in a field denoted as *argumentative XAI*, leveraging argumentative mechanisms to enhance explainability of ML systems (see [7, 47] for recent overviews). Following this new line of research, in this paper we combine computational argumentation and XAI, relying on XAI methods to instantiate Walton-like argumentative schemes [50]. We create such argumentative explanations by the verbalization in natural language of explanations generated by state-of-the-art XAI methods. As a first step, we evaluate the usefulness of such explanations in the context of a medical decision-making task.

4 Methods

To evaluate our AI-generated explanations in a real-world setting, we consider the use case of diagnosing transient loss of consciousness (TLOC) [5]. With TLOC, three main differential diagnoses need to be considered: i) epileptic seizures, ii) syncopes, and iii) psychogenic non-epileptic seizures (PNES) [52].

The differential diagnosis task is challenging since symptoms of TLOC mostly occurs spontaneously outside of diagnostic procedures. At the time of presentation to medical staff, patients might be asymptomatic again, making an accurate diagnosis based on the type of seizure challenging. In this situation, the key challenge for doctors is to inquire about the exact clinical course (the semiology) and framing conditions of the event including potentially facilitating or provoking factors. In particular, a differential diagnosis cannot be

reliably determined solely based on the isolated presence or absence of a specific symptom [37]. In case of a lack of distinct arguments for one particular type of TLOC (a gold standard), combinations of weaker arguments for one type, and exclusion criteria against other TLOC types, become the most important source of decision-making.

This state of affairs makes the diagnosis of epileptic seizures an ideal use case for AI-assisted decision support and therefore also for our user study. The three types of explanations utilised in our study are attribution-based explanations, counterfactual explanations (CFs), and explanations based on the exclusion principle, i.e. specifying why a certain diagnosis can be excluded from the set of possible hypotheses. Explanations and their generation process will be further detailed in the following.

4.1 Data and Modelling

The Ruhr-Epileptology at the University Hospital Knappschaftskrankenhaus Bochum, Germany, provided 32 anonymized and annotated transcripts of outpatient letters as the basis for our study. Annotations consist of the ground-truth diagnosis and the annotation of characteristics which, by presence or absence, support or oppose one of the three possible diagnoses (epileptic seizure, syncope, PNES). The dataset consists of ten patients with epileptic seizures, eleven patients with syncopes, and eleven patients with PNES.

Since this small amount of data is insufficient to train a useful model, the data set was combined with data collected by Wardrope et al. [53], who conducted a retrospective questionnaire study, collecting self- and witness reports from 300 patients (100 each with epilepsy, syncopes, and PNES). They initially collected 117 features and used a random forest approach to reduce the number of features to the 36 most important ones. With this data, we constructed a three-layer Bayesian Network (BN) disease model [42]. We manually grouped variables into risk factors, diseases, and symptoms, encoded by binary nodes, indicating their presence or absence. The BN also allows modelling causal relationships between variables, i.e. risk factors causing diseases, which in turn are the cause of symptoms. The resulting structure of our BN is shown in Figure 1. After defining the structure, parameters for the BN are learned from the aforementioned data. The BN serves as our ML model for predicting the diagnosis and determining the relevant feature used to construct the arguments for our study as it is explained in Section 4.3.
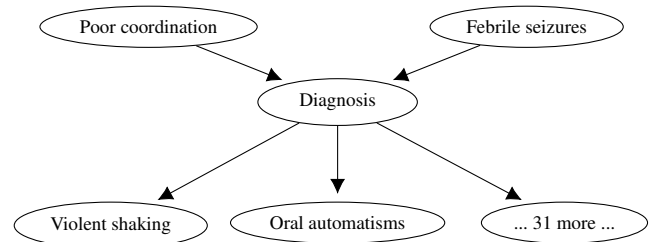
Figure 1. Three-layer Bayesian Network with the 36 most relevant variables determined by Wardrope et al. [53].

4.2 XAI Methods

The basis for generating our explanations are two XAI standard methods: LIME [41] and CFs. LIME identifies the features that have the most significant impact on the model's output, thus highlighting the most important features for that particular prediction. The most

influential features are determined by perturbing the input data, and observing the resulting changes in the prediction to highlight which features have the most influence on the prediction.

CFs offer insights by describing how altering specific input features could impact the model's prediction. These explanations provide similar but hypothetical counterexamples that help users understand how different variables can influence the outcome. We calculate CFs using the open-source Python library Alibi Explain [24].

However, the output of these methods is only feature-based. In the case of LIME, we get the most important features for a given prediction, i.e. diagnosis, and with the calculated CFs, we end up with features and their values that would change the prediction. These feature value pairs are used in conjunction with the annotations from the outpatient letters to construct argumentation-based explanations by instantiating different templates with these feature value pairs. In the following, we will describe this process and the underlying templates in more detail.

4.3 Transforming the Output of XAI Methods into Arguments

In this section, we describe the templates that are used to create argumentation-based explanations in natural language given the feature-based output of the used XAI methods.

Attribution-based explanation This type of explanation is based on the most important features that support the given diagnosis. The argument to support the prediction of *diagnosis X* is created by instantiating the template:

```
Because feature1 is A and feature2 is B,
the patient is most likely to have diagnosis
X.
```

We use this template and fill it with the two most important features identified by LIME. If these features are not among the relevant variables identified a priori for a certain diagnosis by our medical experts, we replace them with other less important features that fulfil this condition.

Counterfactual explanation While more complex than attribution-based arguments, CFs are still found to be easily understood by humans [49]. The argument to support the prediction of *diagnosis X* is created by instantiating the template:

```
If feature1 was A and feature2 was B, the
patient would most likely have diagnosis Y.
If feature3 was C and feature4 was D, the
patient would most likely have diagnosis Z.
```

Where *Y* and *Z* are the two potential alternative diagnoses other than *X*. We then generate a counterfactual for each of the possible alternative diagnostic outcomes.

Exclusion principle explanation This type of explanation is based on the exclusion principle. The idea is to build an argument for why other possible diagnoses can be ruled out as possible hypotheses. The rationale behind this is to resemble the way doctors approach differential diagnosis tasks, i.e. by coming up with possible hypotheses based on the information available, refining those hypotheses, and ultimately eliminating hypotheses that are not sufficiently supported by the available evidence [10]. The argument to support the prediction of *diagnosis X* is created by instantiating the template:

```
Because feature1 is A and feature2 is B,
diagnosis Y can most likely be ruled out.
```

```
Because feature3 is C and feature4 is D,
diagnosis Z can most likely be ruled out.
```

It is noteworthy that in the case of CFs and exclusion principle explanations, both explanations within the argument (targeting *diagnosis Y* and *Z*) could, in principle, share one common feature.

Individual explanations always contain two features, since this is a reasonable amount making explanations still easily comprehensible while not being overly simplistic. Different studies have indeed shown that having two features in an explanation is a reasonable choice [22, 29].

Examples of template-generated explanations for a real case as shown in the user study are given in Table 3 of the Appendix.

4.4 Online User Study

From the initial 32 outpatient letters, 10 cases (4 epileptic seizures, 3 syncopes, 3 PNES) were randomly selected for inclusion in our user study. Based on these letters, slightly shortened and anonymized case reports have been compiled. The user study was structured as follows:

Participant's Background Using a first set of four questions, we captured information about the participants, including their age, sex and experience in diagnosing epileptic seizures as well as experience of dealing with AI systems. Questions about users' experience were answered on a 5-point Likert scale from 1 (no experience) to 5 (expert level).

Case Introduction In the user study, ten cases of real patients were described using a short case report[1]. These cases were constructed based on the anonymized transcripts of outpatient letters provided by our partners.

After participants were presented with a case report, they were asked about the following specific aspects:

Own diagnosis Participants were asked to provide their own estimate of the diagnosis of the patient given the textual description. They had the possibility to select one option from the three possible diagnoses: epileptic seizure, syncope or PNES. Further, they were asked to rate the statement *'I am sure about the diagnosis.'* on a 5-point Likert scale from 1 (does not apply at all) to 5 (fully applies) to assess the user's confidence in the diagnosis. In addition, participants were prompted to select the most likely sub-classification of the diagnosis (if applicable). All possible sub-classifications are summarised in Table 1.

Table 1. Possible sub-classifications of diagnoses. Note: The diagnosis of PNES has no sub-classifications.

Epileptic: focal
Epileptic: multifocal
Epileptic: generalized
Epileptic: focal to bilateral tonic-clonic
Epileptic: other
Syncope: cardiogenic
Syncope: vaso-vagal
Unclear

Before assessing AI-generated explanations, users were asked to provide their own explanations about their proposed diagnosis. This was done to compare their explanations with the ones provided by

---

[1] An example case report is provided in the Appendix

the XAI system. The goal was to determine if there was sufficient objectivity in the task and cross-expert agreement.

Assessment of explanations For each case, our three different argumentative explanations are shown one after one to the user. Experts were asked to rate the following statements on a Likert scale from 1 (does not apply at all) to 5 (fully applies), in order to assess the perceived value of each explanation.

- *The explanation is comprehensible.*
- *The explanation is plausible.*
- *The explanation is complete and contains the most important features to justify the diagnosis.*
- *I would use a similar explanation to explain the facts to a colleague.*

The study was conducted online in an anonymous fashion via a link that was sent to all participants. Participation in the study took around one and a half hours.

Eight physicians were recruited from the Department of Epileptology at Mara Hospital, Bielefeld University and the Ruhr-Epileptology, University Hospital Knappschaftskrankenhaus Bochum, as participants for our user study. Four participants (50%) were male and four participants (50%) were female. The age of the participants ranged from 27 to 54 years (mean = 42.29, standard deviation (SD) = 9.53). Participants reported a mean experience with the differential diagnosis of epileptic seizures of 4 (SD = 1) on a scale from 1 (No experience) to 5 (Expert level), whereas the reported experience in dealing with artificial intelligence was 1.71 (SD = 0.48) on the same scale.

4.5 Follow-up Interviews

After the online study was completed and we have analysed the results, we conducted semi-structured interviews with our participants to gather qualitative data on explanation preferences or requirements in order to resolve questions arising from the online study. Interviews were done online via Zoom, taking around 45 minutes per participant. In particular, we went through the following topics with the participants:

Discussion of results from the user study At the beginning of each interview, the results of the study, cf. Section 5, were presented and discussed.

AI-generated explanations vs. user-provided explanations To further elaborate on *RQ3* and *RQ4*, i.e. the question of the reliability of AI-generated explanations to explain a case to a colleague and to gain quantitative data on what, according to medical experts, constitutes a good explanation, the experts were specifically asked why they think the discrepancy between the reliability rating and user-provided explanations with similar patterns exists, as described in Section 5. Also, the question of how the existing, AI-generated explanations should be improved to meet the users' expectations was discussed in the follow-up interviews.

Related topics In addition to the previously described topics directly aimed at answering our research questions, the following topics have been addressed in the follow-up interviews:

- Importance of interactive or co-constructive explanations.
- Other explanation formats beyond textual representations, e.g. diagrams.
- Major challenges arising from implementing AI decision support in clinical practice.
- General remarks or questions from the users.

5 Results

In the following, we present the results of our study and the results obtained from the conducted follow-up interviews.

In order to check the inter-expert agreement on the diagnoses as well as the self-assessment of the experts, we first looked at the results of the user-provided diagnoses and the reported confidence to compare them among each other and with the ground truth diagnosis provided by our cooperation partners.

The overall certainty of diagnoses across all experts and cases is very high with a mean score of 4.13 (SD = 1.04). This is also reflected in the fact that, in 8 out of 10 cases, all participants agreed on the correct diagnosis. Only the sub-classification of diagnoses revealed some deviations. For example, in one case participants selected 'focal', 'multi-focal', and 'focal to bilateral tonic-clonic' as their sub-classifications for an epileptic seizure. Note that this could also be due to the fact that case reports were a bit shortened and the physicians may not have been able to ask questions about the patients, as it would typically occur in a real diagnostic scenario.

In terms of diagnostic certainty, only two cases showed substantially worse results than the other cases: Case 6 (certainty 2.5 (SD = 0.76) and Case 9 with a certainty of 2.8 (SD = 1.07). However, in both cases, experts agreed on the diagnosis ('Epileptic: focal') except for two participants who gave no answer on the correct diagnosis.
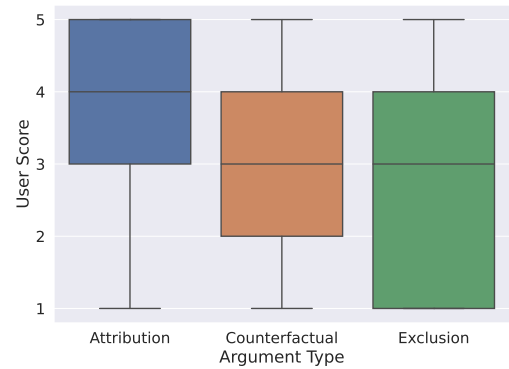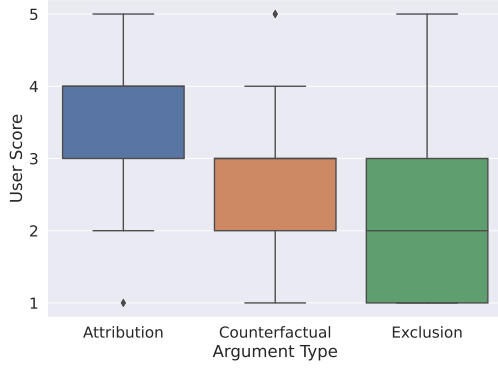

Figure 2. User ratings on the **comprehensibility** of the three generated argument types.

In the following, results are visualized with boxplots and it must be noted, that in some cases (e.g. attribution and counterfactual arguments in Figure 3), the median overlaps with one of the quartiles and is therefore not recognizable in the plot.

Figure 2 shows the results on user comprehensibility of the different argument types. It can be seen that the comprehensibility is always at least moderate to good (median > 3). Users perceived the attribution-based arguments by far as the most comprehensible argument - this applies to both the median value and the distribution of responses. For both CFs and attribution-based explanations, the median value is significantly lower and answers are more distributed on the rating scale.

Concerning the plausibility of explanations, which is depicted in Figure 3, a similar trend can be observed: Attribution-based arguments again achieved the best rankings on plausibility. Compared to the comprehensibility of the arguments, CFs and exclusion-based arguments fall further behind the attribution-based arguments and reported user scores are more concentrated (on the lower spectrum of the scale).

Figure 3. User ratings on the **plausibility** of the three generated argument types.

The completeness of all argument types is perceived as rather low, or moderate in the case of attribution-based arguments as shown in Figure 4. The differences observed in this property are quite surprising, considering that all explanations contain the same number of features. Additionally, both the counterfactual and exclusion-based arguments include alternate diagnoses. Therefore, one would expect these to be perceived as more complete arguments. However, as shown in Figure 4, this is not the case.

To summarise the results of how users evaluated explanations generated by state-of-the-art XAI approaches (*RQ1*), it can be said that arguments generated by state-of-the-art methods are perceived moderately good or good, depending on the evaluated dimension under consideration.
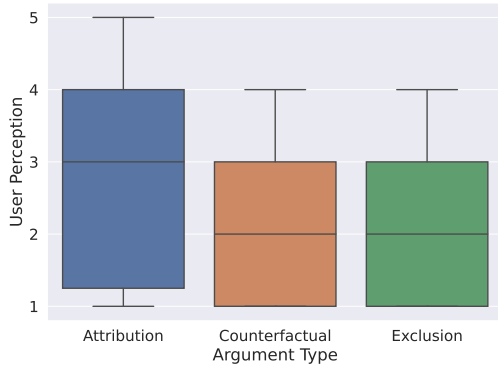
**Figure 4.** User ratings on **completeness** of the three generated argument types.

The differences between the individual argument types can be very marked sometimes, e.g. when comparing the estimated plausibility of an explanation to the question of whether experts would use similar explanations in a discussion with a colleague. On the other hand, differences in comprehensibility and completeness are less significant, as all argument types receive similar evaluations in these aspects. These insights also answer *RQ2* and the question about differences in the evaluation of the different types of arguments. Figure 5 shows the combined results for all evaluated dimensions per argument type. It can be seen that counterfactual-based arguments and arguments based on the exclusion principle are very similar with respect to the median score in the combined analysis, while the simpler arguments based on feature attribution achieved better results.
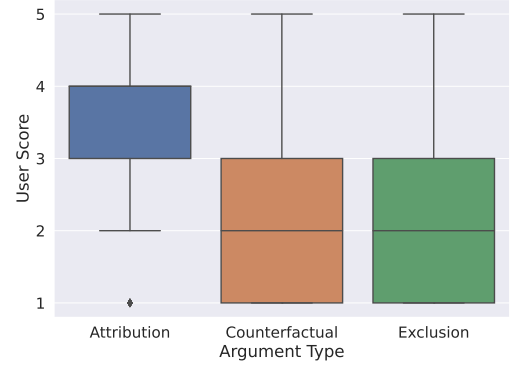
Figure 5. User ratings along **all dimensions combined** for each argument type.

When users were asked whether they would use a similar explanation to the AI-generated one to explain a case to a colleague, most experts assigned rather low scores for all argument types. Although some experts (see outliers in Figure 6) would use all three types, including CFs and exclusion-based explanations to explain a case to a colleague, in most cases they would not use XAI-generated explanations to explain their decision to a colleague. This finding seems to stand in opposition to the assessment of explanations along the dimensions of plausibility, comprehensibility, and completeness.

We thus decided to investigate the reasons for this discrepancy in qualitative follow-up interviews (see below).
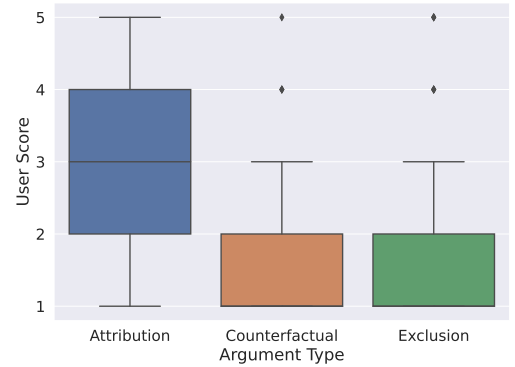
Figure 6. Ratings for the statement *"I would use a similar explanation to explain the facts to a colleague"* for our three argument types.

To address *RQ4*, which is the question of what types of explanations experts would use and how they compare to XAI-generated explanations, we analysed the responses provided by participants when they were asked to give a concise explanation of the diagnosis. We identified distinct types or patterns in the provided explanations and summarised them in Table 2, ordering them according to the number of occurrences.

Interestingly, the exclusion of other diagnoses is a pattern that was frequently encountered in the user-provided explanations. However, experts gave low scores to this kind of explanation. This also applies to the listing of symptoms which was the third most frequent pattern found in the users' explanations. This pattern is similar to our attribution-based arguments. Although this type of explanation was evaluated better, based on the analysis of user explanations, one might have expected a higher ranking.

Table 2. Explanation patterns found in user-provided explanations.

Prototype
Exclusion of other diagnoses
List of symptoms
Prototype & Exclusion of other diagnoses
Prototype & List of symptoms
List of symptoms + Exclusion of other diagnoses

Another interesting finding is that experts often construct an explanation by combining different patterns. For example, they may combine a list of symptoms that support a diagnosis with the mention that this case is a good prototype of the diagnosis:

```
Typical semiology for temporal lobe
seizures of the linguistic dominant
hemisphere and corresponding epilepsy-typical
potentials on the left temporal side.²
```

Another combination encountered multiple times is the exclusion of other diagnoses with mentioning the prototype character of the considered case:

```
Typical anamnesis and description of the
seizure, no indications for differential
diagnoses.²
```

What is missing in XAI-generated explanations? To address the discrepancy between expert's rating along dimensions of plausibility, comprehensibility, and completeness, and the fact that our users would rather not use an AI-generated explanation to explain a case to a colleague; in addition to further answering *RQ3*, this topic was specifically addressed in the follow-up interviews.

When asked about this matter, participants stated that they had not received any information about the quality and reliability of the generated explanations and the uncertainty of the prediction. This made them believe that they might not be able to trust the provided argumentative explanations and, as a result, they would not use them in practice. Participants indicated that information on both the quality of the explanation and the prediction of the AI system would be important.

Besides this regard, the role of personal preferences and the fact that it is difficult to develop a universal solution is also evident when it comes to additional information that could be included in the system, such as graphics, statistics on the patient population, etc.. While some participants would like to receive more information, others expressed concern that too much information could lead to cognitive overload or less intent to use a support tool.

Furthermore, participants found it problematic that the same types of argument were generated and displayed for all cases, which contradicted their intuition that explanations should vary depending on the complexity or ambiguity of a case. In addition to the suggestion to adapt explanations to the complexity of the case, the suggestion was also made to take into account the experience of the doctor and to provide less experienced doctors with more detailed explanations.

Interesting findings from the follow-up interviews not directly related to our research questions The following are some insights we gained from the interviews that are not directly related to our research questions, but are nonetheless worth mentioning:

- General criticism of the test setup. The constrained design of the study does not correspond to a realistic clinical setting in which the expert would rely on the help of an AI system.

---
² Translated from German.

- Participants have expressed that it might be difficult to convince doctors to use a support tool in general.
- The presented study design might also be useful in teaching scenarios or for conducting quality assurance measures within the hospital setting.

6 Conclusion and Outlook

In this paper, we have presented an empirical study to evaluate the usefulness of argumentative explanations generated by state-of-the-art XAI methods with medical users. The application context was in the field of diagnostic support and we considered explaining diagnostic decisions for transient loss of consciousness. Towards this goal, we have developed a simple template-based approach to transform the output of feature attribution methods such as LIME as well as counterfactual approaches into natural language arguments.

In a user study, the generated arguments were evaluated along four dimensions: plausibility, comprehensibility, completeness and applicability to expert explanations. While we found that all explanations score reasonably well along the first three dimensions, the attribution-based methods received the best overall or combined user evaluations. Depending on the considered dimensions, differences between explanation methods became very distinct: In terms of plausibility and comprehensibility, attribution-based arguments were evaluated significantly better than counterfactual or exclusion-based arguments. A surprising result was that, even though the overall assessment of the generated explanations was moderately good, the experts were reluctant to use them to explain their decisions to their colleagues. In a follow-up qualitative study, we could identify two main reasons for this discrepancy:

The lack of information regarding the quality and reliability of AI-generated explanations, as well as the uncertainty of the predictions, which led to a general lack of trust in the AI system. Furthermore, the lack of specificity in the explanations generated for a particular case was identified as a reason for experts' reluctance to use them to explain a case to a colleague. Since we were able to determine patterns in the user-supplied explanations that are very similar to the patterns we used to generate arguments, we believe that improving the specificity of explanations could lead to better AI-generated arguments.

In terms of limitations, it is important to mention that our population was very small. Yet, they are all experienced physicians who have scarce availability. Given their obvious domain expertise, we regard their answers as representative despite the smaller population.

As a next step, we intend to conduct a further study in which we leverage the insights from the current study and construct arguments that have a higher specificity for the given case, combining different XAI methods to generate more comprehensive argumentative explanations.

References

[1] J. Aechtner, L. Cabrera, D. Katwal, P. Onghena, D. P. Valenzuela, and A. Wilbik. Comparing User Perception of Explanations Developed with XAI Methods. In *IEEE International Conference on Fuzzy Systems*, pages 1–7, July 2022.

[2] Y. Alufaisan, L. R. Marusich, J. Z. Bakdash, Y. Zhou, and M. Kantarcioglu. Does Explainable Artificial Intelligence Improve Human Decision-Making?

[3] A. M. Antoniadi, Y. Du, Y. Guendouz, L. Wei, C. Mazo, B. A. Becker, and C. Mooney. Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Applied Sciences*, 11(11):5088, Jan. 2021. ISSN 2076-3417. doi: 10.3390/app11115088.

[4] V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, S. Mourad, P. Pedemonte, R. Raghavendra, J. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, and Y. Zhang. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques, Sept. 2019.

[5] T. Baumgartner and R. Surges. Synkope, epileptischer oder psychogener Anfall? Der Weg zur richtigen Diagnose. *DMW - Deutsche Medizinische Wochenschrift*, 144(12):835–841, June 2019. ISSN 0012-0472, 1439-4413. doi: 10.1055/a-0629-0362.

[6] A. Bertrand, R. Belloum, J. R. Eagan, and W. Maxwell. How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 78–91, Oxford United Kingdom, July 2022. ACM. ISBN 978-1-4503-9247-1. doi: 10.1145/3514094.3534164.

[7] K. Čyras, A. Rago, E. Albini, P. Baroni, and F. Toni. Argumentative XAI: A Survey, May 2021.

[8] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[9] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321 – 357, 1995. ISSN 0004-3702. doi: 10.1016/0004-3702(94)00041-X.

[10] A. S. Elstein and A. Schwarz. Clinical problem solving and diagnostic decision making: Selective review of the cognitive literature. *BMJ : British Medical Journal*, 324(7339):729–732, Mar. 2002. ISSN 0959-8138.

[11] X. Fan and F. Toni. On computing explanations in argumentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), Feb. 2015. doi: 10.1609/aaai.v29i1.9420.

[12] X. Fan, F. Toni, A. Mocanu, and M. Williams. Dialogical two-agent decision making with assumption-based argumentation. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS '14, page 533–540, Richland, SC, 2014. International Foundation for Autonomous Agents and Multiagent Systems.

[13] S. A. Gomez, A. Groza, and C. I. Chesñevar. An argumentative approach to assessing safety in medical device software using defeasible logic programming. In *International Conference on Advancements of Medicine and Health Care through Technology; 5th–7th June 2014, Cluj-Napoca, Romania*, pages 167–172. Springer, 2014. doi: 10.1007/978-3-319-07653-9_34.

[14] M. L. Graber. Reaching 95%: Decision support tools are the surest way to improve diagnosis now. *BMJ Quality & Safety*, 31(6):415–418, June 2022. ISSN 2044-5415, 2044-5423. doi: 10.1136/bmjqs-2021-014033.

[15] M. A. Grando, L. Moss, D. Sleeman, and J. Kinsella. Argumentation-logic for creating and explaining medical hypotheses. *Artificial Intelligence in Medicine*, 58(1):1–13, 5 2013. ISSN 0933-3657. doi: 10.1016/j.artmed.2013.02.003.

[16] P. Hamm, M. Klesel, P. Coberger, and H. F. Wittmann. Explanation matters: An experimental study on explainable AI. *Electronic Markets*, 33(1):17, May 2023. ISSN 1422-8890. doi: 10.1007/s12525-023-00640-9.

[17] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller. Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, 9(4):e1312, 2019. ISSN 1942-4795. doi: 10.1002/widm.1312.

[18] F. Horst, D. Slijepcevic, S. Lapuschkin, A.-M. Raberger, M. Zeppelzauer, W. Samek, C. Breiteneder, W. I. Schöllhorn, and B. Horsak. On the understanding and interpretation of machine learning predictions in clinical gait analysis using explainable artificial intelligence. *ArXiv*, abs/1912.07737, 2019. URL https://api.semanticscholar.org/CorpusID:209386808.

[19] A. Hunter and M. Williams. Aggregating evidence about the positive and negative effects of treatments. *Artificial Intelligence in Medicine*, 56(3):173–190, 11 2012. ISSN 0933-3657. doi: 10.1016/j.artmed.2012.09.004.

[20] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, and B. Baesens. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1):141–154, Apr. 2011.

[21] Y. Jia, J. McDermid, and I. Habli. Enhancing the value of counterfactual explanations for deep learning. In *International Conference on AI in Medicine*, pages 389–394. Springer, 2021.

[22] M. T. Keane and B. Smyth. Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai). In *International Conferenc of Case-Based Reasoning Research and Development*, pages 163–178. Springer, 2020.

[23] M. Khosravi, Z. Zare, S. M. Mojtabaeian, and R. Izadi. Artificial Intelligence and Decision-Making in Healthcare: A Thematic Analysis of a Systematic Review of Reviews. *Health Services Research and Managerial Epidemiology*, 11:23333928241234863, Mar. 2024. ISSN 2333-3928. doi: 10.1177/23333928241234863.

[24] J. Klaise, A. V. Looveren, G. Vacanti, and A. Coca. Alibi explain: Algorithms for explaining machine learning models. *Journal of Machine Learning Research*, 22(181):1–7, 2021.

[25] P. J. Krause, S. Ambler, M. Elvang-Gøransson, and J. Fox. A logic of argumentation for reasoning under uncertainty. *Computational Intelligence*, 11:113–131, 1995. doi: 10.1111/j.1467-8640.1995.tb00025.x.

[26] P. J. Krause, S. Ambler, M. Elvang-Gøransson, and J. Fox. A logic of argumentation for reasoning under uncertainty. *Computational Intelligence*, 11:113–131, 1995. doi: 10.1111/j.1467-8640.1995.tb00025.x.

[27] C. Krittanawong. The rise of artificial intelligence and the uncertain future for physicians. *European Journal of Internal Medicine*, 48:e13–e14, Feb. 2018. ISSN 1879-0828. doi: 10.1016/j.ejim.2017.06.017.

[28] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, and K. Baum. What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, 296:103473, July 2021. ISSN 0004-3702. doi: 10.1016/j.artint.2021.103473.

[29] F. Liedeker, C. Düsing, M. Nieveler, and P. Cimiano. An Empirical Investigation of Users' Assessment of XAI Explanations: Identifying the Sweet-Spot of Explanation Complexity. 2024.

[30] L. Longo and P. Dondio. Defeasible reasoning and argument-based systems in medical fields: An informal overview. In *2014 IEEE 27th International Symposium on Computer-Based Medical Systems*, pages 376–381, 2014. doi: 10.1109/CBMS.2014.126.

[31] L. Longo and L. Hederman. Argumentation theory for decision support in health-care: A comparison with machine learning. In K. Imamura, S. Usui, T. Shirao, T. Kasamatsu, L. Schwabe, and N. Zhong, editors, *Brain and Health Informatics*, pages 168–180, Cham, 2013. Springer International Publishing.

[32] S. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions, Nov. 2017.

[33] W. C. Mann. Dialogue games: Conventions of human interaction. *Argumentation*, 2(4):511–532, 11 1988. ISSN 1572-8374. doi: 10.1007/bf00128990.

[34] C. Meske and E. Bunde. Design Principles for User Interfaces in AI-Based Decision Support Systems: The Case of Explainable Hate Speech Detection. *Information Systems Frontiers*, 25(2):743–773, Apr. 2023. ISSN 1572-9419. doi: 10.1007/s10796-021-10234-5.

[35] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[36] T. Miller. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 333–342, New York, NY, USA, June 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594001.

[37] M. Oto. The misdiagnosis of epilepsy: Appraising risks and managing uncertainty | Elsevier Enhanced Reader. 2016. doi: 10.1016/j.seizure.2016.11.029.

[38] J. Pearl. *Causality*. Cambridge University Press, 2009.

[39] H. Prakken. An abstract framework for argumentation with structured arguments. *Argument and Computation*, 1(2):93–124, 6 2010. ISSN 1946-2174. doi: 10.1080/19462160903564592.

[40] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1): 81–106, Mar. 1986. ISSN 1573-0565. doi: 10.1007/BF00116251.

[41] M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages

1135–1144, 2016.

[42] J. G. Richens, C. M. Lee, and S. Johri. Improving the accuracy of medical diagnosis with causal machine learning. *Nature Communications*, 11(1):3923, Aug. 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17419-7.

[43] K. J. Rohlfing, P. Cimiano, I. Scharlau, T. Matzner, H. M. Buhl, H. Buschmeier, E. Esposito, A. Grimminger, B. Hammer, R. Häb-Umbach, I. Horwath, E. Hüllermeier, F. Kern, S. Kopp, K. Thommes, A.-C. Ngonga Ngomo, C. Schulte, H. Wachsmuth, P. Wagner, and B. Wrede. Explanation as a Social Practice: Toward a Conceptual Framework for the Social Design of AI Systems. *IEEE Transactions on Cognitive and Developmental Systems*, 13(3):717–728, Sept. 2021. ISSN 2379-8939. doi: 10.1109/TCDS.2020.3044366.

[44] C. Rudin. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature machine intelligence*, 1(5):206–215, May 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x.

[45] Q.-u.-a. Shaheen, A. Toniolo, and J. K. F. Bowles. Dialogue games for explaining medication choices. In *Rules and Reasoning: 4th International Joint Conference, RuleML+RR 2020, Oslo, Norway, June 29 – July 1, 2020, Proceedings*, page 97–111, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-57976-0. doi: 10.1007/978-3-030-57977-7_7. URL https://doi.org/10.1007/978-3-030-57977-7_7.

[46] T. Speith. A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2239–2250, June 2022. doi: 10.1145/3531146.3534639.

[47] A. Vassiliades, N. Bassiliades, and T. Patkos. Argumentation and explainable artificial intelligence: A survey. *The Knowledge Engineering Review*, 36:e5, Jan. 2021. ISSN 0269-8889, 1469-8005. doi: 10.1017/S0269888921000011.

[48] G. Vilone and L. Longo. Classification of Explainable Artificial Intelligence Methods through Their Output Formats. *Machine Learning and Knowledge Extraction*, 3(3):615–661, 2021.

[49] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR, Mar. 2018.

[50] D. Walton, C. Reed, and F. Macagno. *Argumentation Schemes*. Cambridge University Press, 2008. doi: 10.1017/CBO9780511802034. URL http://dx.doi.org/10.1017/CBO9780511802034.

[51] X. Wang and M. Yin. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *International conference on intelligent user interfaces*, pages 318–328, 2021.

[52] A. Wardrope, E. Newberry, and M. Reuber. Diagnostic criteria to aid the differential diagnosis of patients presenting with transient loss of consciousness: A systematic review. *Seizure*, 61:139–148, Oct. 2018. ISSN 1059-1311. doi: 10.1016/j.seizure.2018.08.012.

[53] A. Wardrope, J. Jamnadas-Khoda, M. Broadhurst, R. A. Grünewald, T. J. Heaton, S. J. Howell, M. Koepp, S. W. Parry, S. Sisodiya, M. C. Walker, and M. Reuber. Machine learning as a diagnostic decision aid for patients with transient loss of consciousness. *Neurology: Clinical Practice*, 10(2):96–105, Apr. 2020. ISSN 2163-0402. doi: 10.1212/CPJ.0000000000000726.

[54] M. Zielke, S. Houston, M. E. Mancini, G. Hardee, L. Cole, D. Zakhidov, U. Fischer, and T. Lewis. A serious-game framework to improve physician/nurse communication. In *International Conference on Virtual, Augmented and Mixed Reality*, pages 337–348. Springer, 2015. doi: 10.1007/978-3-319-21067-4_35.

Appendix

In the following, we show the translation into English of a German case report as it was part of the study. Case reports consist of three sections: Anamnesis, semiology, and examination results:

Medical history: The patient had an event of the semiology specified below. A similar event had occurred approximately one year earlier. The patient's medical history is unremarkable for typical epilepsy-predisposing factors, the family history is also negative. Without previous internal diseases. In a previously performed EEG, potential sequences were classified as suspicious for sharp-slow-wave and thus as possibly typical for epilepsy.

Semiology: While sitting playing Monopoly, dizziness, then stood up, felt weak and dizzy on the way to the toilet, knees weak, then blacked out, loss of consciousness, and fall. According to anamnestic information from others, the duration of unconsciousness is significantly less than 20 seconds, with no motor manifestations. Quickly re-orientated. Earlier event: comparable onset, but in the further course the patient also felt a cramping of the fingers and hyperventilated. Beyond this, there is no indication of the presence of simple-partial, complex-partial, or secondary-generalised seizures or of myoclonus or absences.

EEG: Well-defined, very regular alpha EEG around 9/sec, no general changes. No foci, no epilepsy-typical potentials. Under repeated vigilance fluctuation, however, frequent groups of generalised, frontally accentuated high-tension alpha and theta waves, but in each case without evidence of clear epilepsy-typical potentials (spike or spike wave). Taking into account the temporal association with phases of vigilance reduction, these are hypnagogic theta groups. The EEG found again in retrospect can also be interpreted as an EEG with hypnagogic theta groups. The single detection of an isolated right-temporal delta wave is not sufficient for the diagnosis of a focal finding.

Table 3. Examples of AI-generated and two user-provided explanations for the same case as it was present in the user study. AB: Attribution-based explanation, CF: Counterfactual explanation, EB: Exclusion principle explanation. All text was translated from German.

AB	Because the episodes occur exclusively while standing or walking, and lying down can alleviate the symptoms, the patient very likely has syncope.
CF	If the episodes had clear provoking factors and the course of the episodes fluctuated significantly, then the most likely diagnosis would be a psychogenic seizure.
	If the EEG were to show epilepsy-specific potentials and the episodes did not exclusively start from standing or walking, then the most likely diagnosis would be an epileptic seizure.
EB	Because the episodes do not have clear provoking factors and the EEG does not show epilepsy-specific potentials, an epileptic seizure can be ruled out with high probability.
	Because no psychiatric preconditions are known and no uncoordinated movement patterns are present, a psychogenic seizure can be ruled out with high probability.
P2	Provoking factors, the possibility of interruption by lying down, only while standing and walking.
P3	Exclusively while standing and walking, can be mitigated by sitting down.