A Prototype of an Interactive Clinical Decision Support System with Counterfactual Explanations

Felix Liedeker[1,*], Philipp Cimiano[1]

[1]CITEC, Bielefeld University, Bielefeld, Germany

Abstract

We describe a prototype of a Clinical Decision Support System (CDSS) that provides (counterfactual) explanations to support accurate medical diagnosis. The prototype is based on an inherently interpretable Bayesian network (BN). Our research aims to investigate which explanations are most useful for medical experts and whether co-constructing explanations can foster trust and acceptance of CDSS.

Keywords

Explainable AI, Clinical decision support, Bayesian network, Counterfactual explanations,

1. Introduction

Diagnostic errors account for around 10 % of adverse events [1, 2]. Clinical Decision Support Systems have the potential to contribute to reducing errors in diagnosis and therapy selection [3]. Despite promising results, however, the adoption and utilisation of CDSS for diagnosis has been very limited so far [4, 3]. Important barriers to adaptation include users' reservations [5] and challenges related to the integration into clinical workflows [3]. A further important barrier for acceptance is the opaqueness of most state-of-the-art (black-box) AI systems [6]. A recent user study has indeed found that users of CDSS prefer to receive explanations instead of suggestions or recommendations only [7].

Towards developing CDSS that are more transparent, we start from a Bayesian network (BN) model, which is an inherently interpretable (white-box) model and allows to explicitly represent causal relationships between variables [8], to develop our CDSS. Our system uses the BN to provide (counterfactual) explanations to support the most probable diagnosis given evidence (such as the symptoms of a patient) to an end user. As a proof-of-concept, we have developed our system to support the prediction of the diagnosis of either epilepsy, syncope, or psychogenic non-epileptic seizures (PNES) in patients with transient loss of consciousness, relying on the data provided by Wardrope et al.[9].

The prototype will be instrumental to answer our main research question: What explanations are actually most useful for experts in the field? Does explainability or the co-construction of an explanation foster trust of the user in the system and can hence improve acceptance and

Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal

*Corresponding author.

fliedeker@techfak.uni-bielefeld.de (F. Liedeker); cimiano@techfak.uni-bielefeld.de (P. Cimiano)

0009-0006-2556-9430 (F. Liedeker); 0000-0002-4771-441X (P. Cimiano)

usage of CDSS? In order to answer this question, we are currently in the process of designing a user study based on the prototype described in this paper.

2. Related work

Although deep learning (DL) has achieved impressive results in various applications, the current challenge for applying ML in the medical field is less related to improving the algorithmic backbone or improving the models by increasing the amount of training data, "but to disentangle the underlying explanatory factors of the data in order to understand the context in an application domain" [10, p.2]. While Pearl has in particular emphasized the importance of causal reasoning for decision making [8], the use of causal inference methods in AI systems, in particular to support diagnosis, is rare [11]. So far there are only a limited number of CDSS that prioritise explainability [12], making this an important research topic.

Despite the fact that explainable AI (XAI) is prominently discussed in the recent AI literature, the notion of explainability remains vague [13] and disconnected from the actual needs of users and stakeholders, and very few systems are in productive use to gain experience on the usefulness of explanations in real-world settings. A lot of work has focused on explaining black-box models: In contrast to directly interpretable models, black-box models are explained by auxiliary methods, which is also referred to as post-hoc explanations [13]. It is argued, both in general [14] and for the special case of AI-driven CDSS [6] that interpretable models should be preferred over black-box models.

Establishing trust via a transparent, explanation-giving CDSS is an important avenue of research as doctors often perceive AI systems as potentially threatening their jobs [15], rather than recognising potential benefits in reducing diagnostic errors. In fact, it has been argued that doctors generally underestimate the risk of making such errors [5]. It is thus key to design systems in such a way that users are under control and can understand what the system is doing and why at all stages of an interaction. In fact, the goal is not to replace doctors by AI systems or, as Holzinger puts it, a "doctor-in-the-loop is indispensable" [10, p.2]. Thus, our prototype has been designed with the goal of a high degree of interactivity, so that the diagnostic decision is not reached by the system alone, but in interaction with a human user. Following Rohlfing et al. we call this paradigm a "co-constructive" approach to decision making and explanation giving [16].

Recently there has been a growing interest in interactive and visual explanations of (black-box) machine learning models, trying to achieve explainability through the analysis of the underlying model. For this purpose, different tools and software have been developed, such as the *modelStudio* software [17], the *What-If Tool* [18] or *explAIner* [19].

3. Methods

3.1. Data

The basis for our system is a three layer BN disease model [11], where binary nodes represent risk factors, diseases and symptoms, respectively, that are either present or absent. The BN

also encodes the causal relationships between risk factors, diseases (caused by risk factors) and symptoms (caused by diseases).

The data basis for our model is the data collected by Wardrope et al. [9]. Based on a retrospective self- and witness-report questionnaire study with 300 patients (100 each with epilepsy, syncope and PNES), the authors used a random forest approach to select the 36 most relevant features of their original 117 features collected. These 36 variables were then used to construct our BN.

3.2. CDSS prototype

The second step was to build a front-end for our model. The front-end was developed as a web application to ensure ease of use and access. Within the application, the user can input evidence about the patient, i.e. symptoms that are either present or absent. This process works sequentially: After each new user input, the different measures are recalculated and the user can ask for additional explanations (e.g. *How much would the presence of X change the probability of the diagnosis?*). After a new piece of evidence has been added by the user, the next most useful evidence is computed by the system and shown to the user.

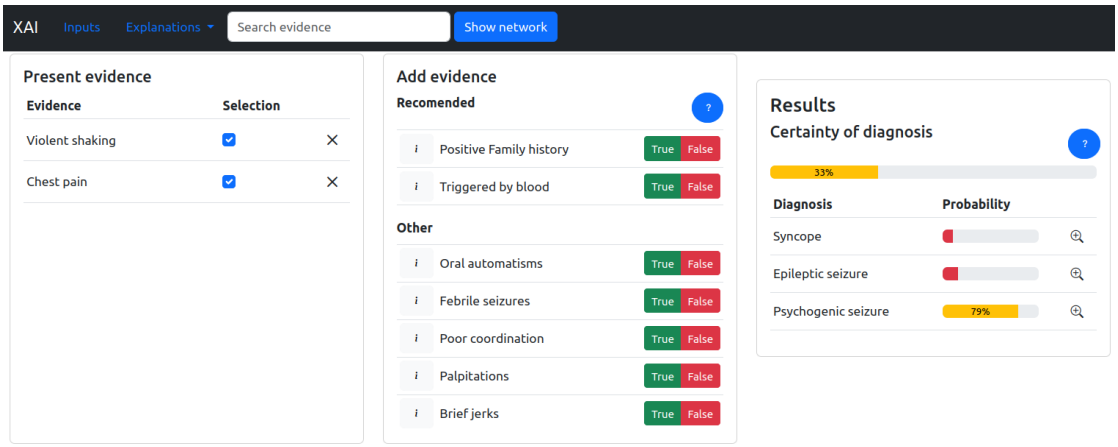

Figure 1: Main window of our prototype. In the example shown, two features (presence of *chest pain* and *violent shaking*) have been selected as evidence. In addition to the probabilities of the diagnoses and the certainty of the diagnosis, two features (*positive family history* and *triggered by blood*) that would reduce the uncertainty of the diagnosis most, are shown to the user. The user can input further evidence, check the available explanations (cf. figure 2) or view more detailed information (magnifying glass and question mark symbols). The prototype is available online at: https://webtentacle1.techfak.uni-bielefeld.de/xai-interaction/.

Within the prototype, different types and levels of explanations are available to the user. Besides a view of the underlying BN and SHAP (SHapley Additive exPlanations) values for feature importance [20], causal explanation trees (CET) [21] and most relevant explanations (MRE) [22] (a partial instantiation of the three diagnoses that maximises the generalised Bayes factor (GBF) as relevance measure) have been implemented as explanation methods.

In addition, counterfactual explanations, calculated via the *expected sufficiency* (symptoms to

persist if all other causes of the symptoms would switch off) and *expected disablement* (symptoms which would switch off, if the disease would not be present) [11] or *pertinent positives* (features minimal sufficient to justify the classification) and *pertinent negatives* (features that would alter the classification if added) [23], are incorporated in our prototype.

4. Future work

Currently, the data and features in our data set are limited. In spite of this, the accuracy of the baseline BN is 80.23%, which represents a reasonable performance given that our goal is not to have a fully automatic approach but one in which the doctor is kept in the loop and makes the final decision. In order to improve the model and to increase the number of parameters in the model, we are working on the annotation of a bigger data set comprising of more than 2000 outpatient letters and video EEG recordings in cooperation with the epilepsy centre at the University Hospital Bochum, Germany. A crucial next step is the evaluation of our prototype with respect to two important aspects: User needs regarding helpful explanations and the usability of the whole system. So far, the decision regarding which explanations and information to include in the prototype is based on inspiration from the literature and on what the BN can compute. As a next step, we will conduct a user study with epileptologists to determine which explanations are most helpful for this target group and what queries they are interested in. A second study will be conducted to evaluate the usability of the system, with a particular focus on the degree of interactivity preferred by users and whether they prefer to have as much information as possible displayed directly or retrieve pieces of information only on request.

Once the development of the whole system is complete, the key question of whether or not such an interactive, co-constructive system can foster trust in CDSS and help overcome user concerns can be addressed.


The best diagnosis is epileptic seizure NOT present and syncope NOT present and psychogenic seizure present. This diagnosis is **probable** (GBF: 12.25) and the next diagnosis epileptic seizure present and syncope present has a **moderate difference** (11.59) to the best diagnosis.

GBF	Description
12.25	epileptic seizure: NOT present syncope: NOT present psychogenic seizure: present
0.66	epileptic seizure: present syncope: present
0.38	epileptic seizure: present psychogenic seizure: NOT present

Figure 2: Most relevant explanation for the example shown in figure 1.

Acknowledgments

Funded by the *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation): TRR 318/1 2021 – 438445824.

References

[1] G. R. Baker, P. G. Norton, V. Flintoft, R. Blais, A. Brown, J. Cox, E. Etchells, W. A. Ghali, P. Hébert, S. R. Majumdar, M. O'Beirne, L. Palacios-Derflingher, R. J. Reid, S. Sheps, R. Tamblyn, The Canadian Adverse Events Study: The incidence of adverse events among hospital patients in Canada, CMAJ 170 (2004) 1678–1686. doi:10.1503/cmaj.1040498.

[2] M. Soop, U. Fryksmark, M. Köster, B. Haglund, The incidence of adverse events in Swedish hospitals: A retrospective medical record review study, International Journal for Quality in Health Care 21 (2009) 285–291. doi:10.1093/intqhc/mzp025.

[3] K. K. Hall, S. Shoemaker-Hunt, L. Hoffman, S. Richard, E. Gall, E. Schoyer, D. Costar, B. Gale, G. Schiff, K. Miller, T. Earl, N. Katapodis, C. Sheedy, B. Wyant, O. Bacon, A. Hassol, S. Schneiderman, M. Woo, L. LeRoy, E. Fitall, A. Long, A. Holmes, J. Riggs, A. Lim, Making Healthcare Safer III: A Critical Analysis of Existing and Emerging Patient Safety Practices, Agency for Healthcare Research and Quality (US), Rockville (MD), 2020.

[4] S. Cheraghi-Sohi, R. Alam, M. Hann, A. Esmail, S. Campbell, N. Riches, Assessing the utility of a differential diagnostic generator in UK general practice: A feasibility study, Diagnosis 8 (2021) 91–99. doi:10.1515/dx-2019-0033.

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

[6] R. L. Pierce, W. Van Biesen, D. Van Cauwenberge, J. Decruyenaere, S. Sterckx, Explainability in medicine in an era of AI-based clinical decision support systems, Frontiers in Genetics 13 (2022) 903600. doi:10.3389/fgene.2022.903600.

[7] C. Panigutti, A. Beretta, F. Giannotti, D. Pedreschi, Understanding the impact of explanations on advice-taking: A user study for AI-based clinical Decision Support Systems, in: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 2022.

[8] J. Pearl, Causality, Cambridge University Press, 2000.

[9] 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, M. Reuber, Machine learning as a diagnostic decision aid for patients with transient loss of consciousness, Neurology: Clinical Practice 10 (2020) 96–105. doi:10.1212/CPJ.0000000000000726.

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

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

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

[13] 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, Y. Zhang, One Explanation

Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques, 2019. doi:10.48550/arXiv.1909.03012. arXiv:1909.03012.

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

[15] C. Krittanawong, The rise of artificial intelligence and the uncertain future for physicians, European Journal of Internal Medicine 48 (2018) e13–e14. doi:10.1016/j.ejim.2017.06.017.

[16] 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, 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 (2021) 717–728. doi:10.1109/TCDS.2020.3044366.

[17] H. Baniecki, P. Biecek, modelStudio: Interactive Studio with Explanations for ML Predictive Models, Journal of Open Source Software 4 (2019) 1798. doi:10.21105/joss.01798.

[18] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viegas, J. Wilson, The What-If Tool: Interactive Probing of Machine Learning Models, IEEE Transactions on Visualization and Computer Graphics (2019) 1–1. doi:10.1109/TVCG.2019.2934619. arXiv:1907.04135.

[19] T. Spinner, U. Schlegel, H. Schäfer, M. El-Assady, explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning, IEEE Transactions on Visualization and Computer Graphics (2019) 1–1. doi:10.1109/TVCG.2019.2934629. arXiv:1908.00087.

[20] S. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions, 2017. doi:10.48550/arXiv.1705.07874. arXiv:1705.07874.

[21] U. Nielsen, J.-P. Pellet, A. Elisseeff, Explanation Trees for Causal Bayesian Networks, 2012. doi:10.48550/arXiv.1206.3276. arXiv:1206.3276.

[22] C. Yuan, H. Lim, T.-C. Lu, Most relevant explanation in Bayesian networks, Journal of Artificial Intelligence Research 42 (2011) 309–352.

[23] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, P. Das, Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives, 2018. doi:10.48550/arXiv.1802.07623. arXiv:1802.07623.