Multimodal Explanations for User-centric Medical Decision Support Systems

Bettina Finzel[1], **David Elias Tafler**[1], **Anna Magdalena Thaler**[1], **Ute Schmid**[1]

[1]Cognitive Systems, University of Bamberg
An der Weberei 5
96047 Bamberg, Germany

Abstract

Based on empirical evidence indicating that different types of explanations should be used to satisfy different users in terms of their information needs and to increase trust in the system, we motivate the use of multimodal explanations for decisions made by a machine learning model to support medical diagnoses. We present a system through which medical professionals or students can obtain verbal explanations for a classification by means of a dialogue and to which they can make queries to get prototypical examples in the form of images showing typical health conditions. Our approach could be used for validating algorithmic decisions using a human-in-the-loop method or for medical education.

Introduction

In medical diagnostics, Deep Learning is increasingly used to classify patient data. Due to the condition that such systems must be transparent, great progress has been made in the field of explainable artificial intelligence (XAI) in recent years. Work on visual explanations for example for the classification of human tissue (Hägele et al. 2020), malaria probes (Schallner et al. 2019) and facial expressions for patients suffering from pain (Rieger et al. 2020) shows that the methods developed can be used to reveal which features a deep neural network has found relevant for classification. However, relational information in terms of complex relationships between features of the data was not used to make the classification decision. As recent works have emphasized, medical diagnosis is often based on examining relational data, thus, a model that is able to incorporate these and to explain its decision in a relational manner, is key (Bruckert, Finzel, and Schmid 2020; Schmid and Finzel 2020; Holzinger et al. 2021). Moreover, the focus of these works has not been to present different explanations, that is, in varying modalities. Being able to make decisions based on complex relationships and being able to explain them in as many ways as possible are two important aspects of building systems that are not only transparent but also understandable to human decision makers. Multi-modal explanations take different angles in explaining a classification and thus may satisfy different users in terms of their varying need for information. Such systems can empower the user to validate the system and to remain in control of decisions, which is a crucial requirement in medicine (Tizhoosh and Pantanowitz 2018).

Recent research uses inductive logic programming (ILP) to train models that can be explained to the user in a comprehensible way and which are capable to deal with complex relational data. In contrast to visual explanations that can transfer only the information of occurrence or absence, a relational approach such as ILP can express arbitrarily complex relationships, for example spatial and temporal relations as well as recursion (Schmid 2018). In addition to the training data, ILP can be enriched by existing background knowledge that is by expert knowledge, be it for training or correction of learned models (Schmid and Finzel 2020). In the past, there have been prominent transparent systems that made decisions based on relational data, such as MYCIN (Shortliffe 2012). However, the focus there was more on building expert systems and less on satisfying different users through multimodal explanations.

To close this gap, we build on the findings from empirical research on multimodal explanations as well as on a recent research work that combines two different explanation approaches, namely verbal, dialogue-based as well as visual, image-based explanations. We use our approach for the first time to classify and explain medical data.

Related Work

Explainable Artificial Intelligence (XAI) aims to make AI systems, their decisions and actions comprehensible. Given the high stakes involved in medical diagnosis and human health in general, it is obvious that AI systems applied in this domain need to be understood by the user. To this end, an AI system may produce explanations of various types and formats. Building on previous work on combining verbal and visual explanations in artificial intelligence (Finzel et al. 2021), we here apply these concepts to the medical domain. We explore the potential use of different kinds of explanations given by a diagnostic decision support system for assessing primary tumors in tissue samples. Specifically, we consider dialogue- and image-based explanation, allowing for step-wise exploration of the reasons behind a classification outcome as well as displaying images on demand that show prototypical examples of health conditions.

The fact that there exists more than one type of explanation suggests that not every explanation fits every situation. For instance, to explain the classification of an object as belonging to one of two categories that only differ by color, a visual explanation might be preferred over a verbal explanation based on the nature of the problem. In contrast to this, visual classification tasks that rely on relational information rather than a simple presence or absence of features might require additional verbal explanations to improve the joint performance, trust in the decision aid system and to correctly counteract faulty system predictions (Thaler and Schmid 2021).

We further want to point to the importance of the person requesting the explanation. The idea of tailoring explanations to and evaluating them based on the goals of the explainee is not new in cognitive science (Leake 1991) and has been supported by more recent empirical evidence. For example (Vasilyeva, Wilkenfeld, and Lombrozo 2015) showed that people prefer explanations (formal, mechanistic, teleological) that are consistent with their goals. Also in the field of XAI, the users' inter- and intraindividual differences have been recognized as important to the development and improvement of XAI (e.g. (Gunning and Aha 2019; Miller 2019; Kulesza et al. 2015)).

In medicine, the potential applications of AI are manyfold, spanning diagnostics, therapeutics, population health management, administration, and regulation (He et al. 2019). In order to make such systems transparent, their explanations need to consider certain user characteristics, such as their expertise and goals. Especially in applications with potentially extreme consequences, such as decision support for diagnosing illnesses, the diagnostician needs to understand the system's recommendations to make well-informed decisions. Along these lines, (Holzinger et al. 2019) have differentiated between explainability, a more technical attribute of an algorithm, and causability, a feature of explanations that describes how well an explanation can transfer causal understanding to a human user. In order to increase causablilty of medical decision support systems, we combine different kinds of explanations. The user can request various explanations via a conversational interaction with the system and thus control the transmission of understanding that is explanation.

Multimodal Explanations for Medical Decision Making

In this section we show how our multimodal explanation approach can be applied to the medical use case of primary tumor staging. We first introduce the medical terminology and concepts for tumor staging and present examples for verbal, dialogue-based explanations as well as visual, prototype-based explanations accordingly.

Primary Tumor Classification in Colon Tissue Samples

The task of classifying tumors requires different competencies and diagnostic steps. The main tasks involved are tumor staging and grading (Wittekind, Bootz, and Meyer 2004).

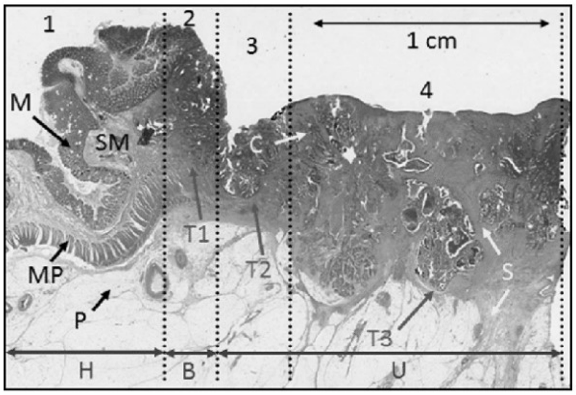

Figure 1: An example of a colon tissue sample under the microscope containing different stages (T1-T3) of tumors in accordance to the widely used TMN staging system (Wittekind, Bootz, and Meyer 2004) with different colon tissues involved: mucosa (M), submucosa (SM), muscularis propria (MP) and pericolic adipose tissue (P). The image was taken from (Pierangelo et al. 2013) for illustration of our use case.

While staging refers to determining the extent of a tumor (location, size and spreading in different layers of tissue), grading examines the abnormality of the appearance of tumor cells and tumor tissue. In this paper, we focus on the task of staging for primary tumors that is determining invasion depth of an original tumor in human colon tissue layers. We therefore look at spatial relationships between a tumor and its surrounding tissue layers. The most widely used system for tumor staging is the TMN staging system (Wittekind, Bootz, and Meyer 2004). This system is used to denote the stage of a tumor in pathology reports. The letters T, M and N are combined with further letters or numbers to indicate the exact stage. We focus on the T category, which is concerned with the size and the extent of the main tumor, also called the primary tumor. If a primary tumor is found in the colon tissue, it is assigned with one of five possible stages: Tis, T1, T2, T3 or T4. The higher the number after the T, the bigger the extent of the tumor. Tis stands for carcinoma in situ and denotes a tumor that hasn't yet extended to the next tissue layer. The stages can be further differentiated depending on the kind of tissue affected by the tumor (e.g. T4a, T4b). Note that there are further assignments, e.g. TX for tumors that cannot be assessed or T0 if there is no evidence for a tumor. We disregard these cases.

Figure 1 shows a colon tissue sample, where healthy tissue and three of four stages are present (corresponding to the 4 zones separated by dotted lines). The leftmost zone contains healthy tissue that can be divided into mucosa (M), submucosa (SM), muscularis propria (MP) and pericolic adipose tissue (P). Zone 2 includes a T1 tumor (invading the mucosa and the submucosa), zone 3 a T2 tumor (extending to the muscularis propria) and zone 4 contains a T3 tumor (that is growing past the boundaries of the muscularis propria into the preicolic adipose tissue). The letters C and S included in Figure 1 denote tumor cells and tumor stroma, H, B and U denote further diagnostic areas, however, this

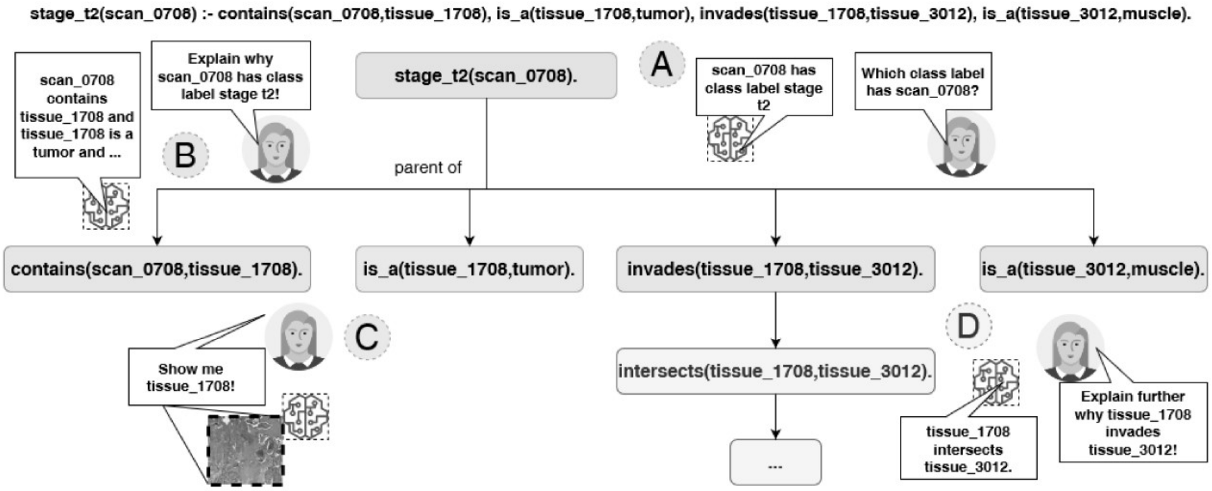
stage_t2(scan_0708) :- contains(scan_0708,tissue_1708), is_a(tissue_1708,tumor), invades(tissue_1708,tissue_3012), is_a(tissue_3012,muscle).
scan_0708 contains tissue_1708 and tissue_1708 is a tumor and ...
Explain why scan_0708 has class label stage t2!
B
stage_t2(scan_0708).
A
scan_0708 has class label stage t2
Which class label has scan_0708?
parent of
contains(scan_0708,tissue_1708).
is_a(tissue_1708,tumor).
invades(tissue_1708,tissue_3012).
is_a(tissue_3012,muscle).
C
Show me tissue_1708!
intersects(tissue_1708,tissue_3012).
...
D
tissue_1708 intersects tissue_3012.
Explain further why tissue_1708 invades tissue_3012!

Figure 2: An explanatory tree for *stage_t2(scan_0708)*, that can be queried by the user to get a local explanation of why scan_0708 is labeled as T2 (steps A and B). A dialogue is realized by further requests, either to get more visual explanations in terms of prototypes (step C) or to get more verbal explanations in a drill-down manner (step D).

is not important for the work presented here and therefore not explained further. To increase the readability of the following paragraphs for the reader who may not be familiar with medical terminology, we will name the different tissue types in the following subsections mucosa, submucosa, muscle and fat tissue.

Dialogue- and Prototype-based Explanations

Analogous to the example presented in previous work (Finzel et al. 2021), we can translate the given expert knowledge introduced by Figure 1 into background knowledge and train an ILP model on examples and the background knowledge to obtain rules for the classification of stages which can then be used to produce verbal explanations for a conversational dialogue with the user.

For cases similar to the example presented in Figure 1 we can get annotations of the different tissues manually or automatically (Schmid and Finzel 2020) and determine with the help of a spatial calculus, whether they intersect (Bruckert, Finzel, and Schmid 2020). Providing examples for each stage (T1-T4), where the corresponding background knowledge contains the information, which tissues intersect (tumor intersects mucosa for T1 example, tumor intersects muscle for T2 example, and so on) as well as providing negative, contrastive examples, we can derive a set of rules for each stage.

This set of rules can be seen as a *global* explanation, meaning it explains the characteristics of a class. These rules contain variables and relationships between them that are satisfied by all positive examples and no negative examples. The background knowledge can be arbitrarily complex, consisting of either only singular properties or more sophisticated relationships, such as the definition of spatial relations and reasoning rules. Given the learned rules and the background knowledge, we create so-called explanatory trees that explain the classification of individual examples and can therefore be considered as *local* explanations (Finzel et al. 2021).

An exemplary rule from an ILP model that was trained to recognize tissue samples of stage T2, might state that a scan A is classified as stage T2, if it holds that A contains B and B is a tumor and B invades C and C is muscle tissue. This rule represented in the logic programming language Prolog would state: "*stage_t2(A) :- contains(A,B), is_a(B,tumor), invades(B,C), is_a(C,muscle).*". The upper case letters A, B, C are variables that can be substituted by lower case constants by applying the rule to the given positive examples, meaning that the background knowledge consisting of containment and spatial relationships satisfies the learned rule.

The explanatory tree we create to explain the classification of an individual example results in a structure presented in Figure 2 based on a logical proof procedure introduced in (Finzel et al. 2021). The class label is set to the root node of the explanatory tree and the reasons for the class decision, given by the substitution of variables in the learned rule, determine the child nodes of the root node. For our colon tissue example individual parts of the rule (e.g. the *invades* relationship) can be explained by further background knowledge (in this case the definition of some spatial relationship *intersects* that was computed based on geometric properties of the input data). The explanatory tree can be traversed in a conversational manner (see Figure 2) to obtain verbal explanations for the reasons of a stage classification of a particular microscopy scan.

A special property of our approach is that we complement the verbal explanations by visual explanations in terms of prototypes, in cases, where a verbal explanations cannot be presented due to limits of expression (e.g. if the user wants to see how a certain tissue type looks like). Explanations by means of prototypes are based on the idea that categories, especially those without unambiguous necessary and sufficient criteria for including or excluding examples, can be represented by a central tendency of the category members, called a prototype (Rosch 1987). We chose prototypes for the implementation of a complementary explanation method besides verbal explanation, because research has shown that prototypes are relevant, among others in category learning (Minda and Smith 2001), scheme-inductive reasoning as a successful diagnostic reasoning strategy (Coderre et al. 2003) and expert teaching (Sternberg and Horvath 1995). In our model, prototypes are representative category members and displayed as images.

Having the explanatory tree as well as the images of the prototypes, the user can now traverse the explanatory tree and ask for prototypes through a dialogue with the system. Analogous to the approach presented first in (Finzel et al. 2021) the user can ask for a global explanation, e.g., "what does *stage T2* mean?". In order to make a request for *local* explanations, the user can pose the following requests (see Figure 2):

- Which class label has <example>? (reference A)
- Explain why <example> has class <class label>! (reference B)
- Show me <concept>! (reference C, displays a prototype)
- Explain further why <relation>! (reference D, allows for drill-down of explanations)

Users can furthermore request to return to the last explanation in order to proceed with their search for answers on different branches of the explanatory tree. In (Finzel et al. 2021) the class labels describe only binary relationships. For the use case presented here, we extended the code to cover class labels for unary relations (the traditional nature of a class label).

The implementation, including files to train an ILP model, the code to create an explanatory tree, the images showing the prototypes as well as the dialogue-based interface are available via a git repository[1].

Discussion

Often interpretable approaches are seen as an alternative to explanation generation for black boxes: It has been argued that for high stake decision making, for instance in health care, interpretable models should be preferred over ex-post explanation generation for neural network models (Rudin 2019). It has been pointed out that explanations might be misleading and inspire unjustified trust (Babic et al. 2021). However, although interpretable models such as decision rules or ILP models are white box and therefore inspectable – providing explanations might be still necessary, even for high performing models. High accuracy does not imply that the model considered features that are relevant to the application domain. Furthermore, similar to computer programs, white box models might be inspectable in principle but can be too complex for easy comprehensibility. Explanation mechanisms that make use of different modalities and detail as the ones proposed in this paper are helpful or even necessary to communicate what the model considered as relevant. In our case, the model reached a high accuracy by covering all positive examples and no negative examples. Still, it is necessary that medical experts can inspect the reasons behind each decision as motivated before. Therefore, the presented conversational interface was integrated on top of the multimodal explanation approach.

With respect to decision support systems that are based on learned models, requirements have recently been stated (Bohanec 2021), in the light of which we want to discuss our approach. In his work Bohanec points out that there are five requirements that should be fulfilled. The first one is *correctness*, meaning that the model should provide correct (valid, right) information given the decision problem. Second, the model should fulfill *completeness*, a property that refers to considering all relevant aspects of the decision problem and providing answers for all possible inputs. Next, he mentions *consistency* in terms of logical and preferential consistency. Another important requirement is *comprehensibility* of provided information for the user. Finally, Bohanec mentions *convenience*, referring to easily accessible, timely information, appropriate for the task and the user.

Our implementation fulfills a part of these requirements by design. The underlying ILP algorithm that produced the model is complete and consistent with respect to the problem domain (Finzel et al. 2021). Furthermore, ILP output can be considered to be comprehensible for humans, especially since it is easy to translate it to verbal statements in the form of natural language (Muggleton et al. 2018). Furthermore, our approach heads towards convenience by presenting explanations in different modalities to suite different users, levels of understanding and tasks. Correctness is ensured at least by the deductive step, when explanatory trees are created from previously induced rules.

In such visually complex domains as tissue classification, near misses as a further type of example-based explanations (Rabold, Siebers, and Schmid accepted Aug. 2021) could be integrated into the explanation procedure presented here, since they can help to communicate information about the decision boarders between diagnostic categories. Our approach could be further extended for an application in medical education, which is an interesting field to use new explanatory techniques (Chan and Zary 2019), for example in histopathological diagnosis (Crowley and Medvedeva 2003).

Conclusion

Motivated by empirical evidence that indicates that multimodal explanations are beneficial for understanding, we presented an approach and its implementation that combines verbal, dialogue-based explanations with visual, prototype-based explanations in order to give insights into the reasons of a decision of a model trained to classify the stage of cancerous colon tissue samples. We applied inductive logic programming to generate this said model, an approach that

[1] Gitlab repository of our implementation of multimodal explanations (including two example data sets: a proof-of-concept data set from the animal world and a data set for colon tissue classification for T1-T4 stages): https://gitlab.rz.uni-bamberg.de/cogsys/public/multi-level-multi-modal-explanation

fulfills the requirements of completeness, consistency and comprehensibility by design. Further empirical investigations shall evaluate the helpfulness of our implementation in terms of convenience.

Acknowledgments

The work presented in this paper is funded by grant FKZ 01IS18056 B, BMBF ML-3 Transparent Medical Expert Companion (TraMeExCo), 2018-2021. We say many thanks to our project partners from the Fraunhofer IIS (Volker Bruns, Dr. Michaela Benz) and the University Hospital Erlangen (Dr. med. Carol Geppert, Dr. med. Markus Eckstein, and Prof. Dr. Arndt Hartmann, head of institute of pathology) who provided us with data our simulated case study is based upon and with the knowledge about assessing primary tumors in colon cancer microscopy scans.

References

Babic, B.; Gerke, S.; Evgeniou, T.; and Cohen, I. G. 2021. Beware explanations from AI in health care. *Science* 373(6552): 284–286.

Bohanec, M. 2021. *From Data and Models to Decision Support Systems: Lessons and Advice for the Future*, 191–211. Springer.

Bruckert, S.; Finzel, B.; and Schmid, U. 2020. The Next Generation of Medical Decision Support: A Roadmap Toward Transparent Expert Companions. *Frontiers in Artificial Intelligence* 3: 75.

Chan, K. S.; and Zary, N. 2019. Applications and challenges of implementing artificial intelligence in medical education: integrative review. *JMIR medical education* 5(1): e13930.

Coderre, S.; Mandin, H.; Harasym, P. H.; and Fick, G. H. 2003. Diagnostic reasoning strategies and diagnostic success. *Medical education* 37(8): 695–703.

Crowley, R. S.; and Medvedeva, O. 2003. A general architecture for intelligent tutoring of diagnostic classification problem solving. In *AMIA Annual Symposium Proceedings*, volume 2003, 185. American Medical Informatics Association.

Finzel, B.; Tafler, E. D.; Scheele, S.; and Schmid, U. 2021. Explanation as a process: user-centric construction of multi-level and multi-modal explanations. In *German Conference on Artificial Intelligence (Künstliche Intelligenz)*, (to be published). Springer.

Gunning, D.; and Aha, D. 2019. DARPA's explainable artificial intelligence (XAI) program. *AI Magazine* 40(2): 44–58.

Hägele, M.; Seegerer, P.; Lapuschkin, S.; Bockmayr, M.; Samek, W.; Klauschen, F.; Müller, K.-R.; and Binder, A. 2020. Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. *Scientific reports* 10(1): 1–12.

He, J.; Baxter, S. L.; Xu, J.; Xu, J.; Zhou, X.; and Zhang, K. 2019. The practical implementation of artificial intelligence technologies in medicine. *Nature medicine* 25(1): 30–36.

Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K.; and Müller, H. 2019. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9(4): e1312.

Holzinger, A.; Malle, B.; Saranti, A.; and Pfeifer, B. 2021. Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI. *Information Fusion* 71: 28–37. ISSN 1566-2535.

Kulesza, T.; Burnett, M.; Wong, W.-K.; and Stumpf, S. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, 126–137.

Leake, D. B. 1991. Goal-based explanation evaluation. *Cognitive Science* 15(4): 509–545.

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

Minda, J. P.; and Smith, J. D. 2001. Prototypes in category learning: the effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 27(3): 775.

Muggleton, S. H.; Schmid, U.; Zeller, C.; Tamaddoni-Nezhad, A.; and Besold, T. 2018. Ultra-strong machine learning: comprehensibility of programs learned with ILP. *Machine Learning* 107(7): 1119–1140.

Pierangelo, A.; Manhas, S.; Benali, A.; Fallet, C.; Totobenazara, J.-L.; Antonelli, M. R.; Novikova, T.; Gayet, B.; De Martino, A.; and Validire, P. 2013. Multispectral Mueller polarimetric imaging detecting residual cancer and cancer regression after neoadjuvant treatment for colorectal carcinomas. *Journal of biomedical optics* 18(4): 046014.

Rabold, J.; Siebers, M.; and Schmid, U. accepted Aug. 2021. Generating Contrastive Explanations for Inductive Logic Programming Based on a Near Miss Approach. *Machine Learning* .

Rieger, I.; Kollmann, R.; Finzel, B.; Seuss, D.; and Schmid, U. 2020. Verifying Deep Learning-based Decisions for Facial Expression Recognition. In *28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2020, Bruges, Belgium, October 2-4, 2020*, 139–144.

Rosch, E. 1987. Wittgenstein and categorization research in cognitive psychology. In *Meaning and the growth of understanding*, 151–166. Springer.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1(5): 206–215.

Schallner, L.; Rabold, J.; Scholz, O.; and Schmid, U. 2019. Effect of Superpixel Aggregation on Explanations in LIME - A Case Study with Biological Data. *CoRR* abs/1910.07856. URL http://arxiv.org/abs/1910.07856.

Schmid, U. 2018. Inductive Programming as Approach to Comprehensible Machine Learning. In *DKB/KIK@ KI*, 4–12.

Schmid, U.; and Finzel, B. 2020. Mutual Explanations for Cooperative Decision Making in Medicine. *Journal: KI-Künstliche Intelligenz* 2: 227–233.

Shortliffe, E. 2012. *Computer-based medical consultations: MYCIN*, volume 2. Elsevier.

Sternberg, R. J.; and Horvath, J. A. 1995. A prototype view of expert teaching. *Educational researcher* 24(6): 9–17.

Thaler, A. M.; and Schmid, U. 2021. Explaining Machine Learned Relational Concepts in Visual Domains-Effects of Perceived Accuracy on Joint Performance and Trust. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Tizhoosh, H. R.; and Pantanowitz, L. 2018. Artificial intelligence and digital pathology: challenges and opportunities. *Journal of pathology informatics* 9.

Vasilyeva, N.; Wilkenfeld, D.; and Lombrozo, T. 2015. Goals Affect the Perceived Quality of Explanations. *Cognitive Science* .

Wittekind, C.; Bootz, F.; and Meyer, H.-J. 2004. Tumoren des Verdauungstraktes. In Wittekind, C.; Bootz, F.; and Meyer, H.-J., eds., *TNM Klassifikation maligner Tumoren*, International Union Against Cancer, 53–88. Springer.