User-Centric Explainability in Healthcare: A Knowledge-Level Perspective of Informed Machine Learning

Luis Oberste and Armin Heinzl

***Abstract*—Explaining increasingly complex machine learning will remain crucial to cope with risks, regulations, responsibilities, and human support in healthcare. However, extant explainable systems mostly provide explanations that mismatch clinical users' conceptions and fail their expectations to leverage validated and clinically relevant information. A key to more user-centric and satisfying explanations can be seen in combining data-driven and knowledge-based systems, i.e., to utilize prior knowledge jointly with the patterns learned from data. We conduct a structured review of knowledge-informed machine learning in healthcare. In this article, we build on a framework to characterize user knowledge and prior knowledge embodied in explanations. Specifically, we explicate the types and contexts of knowledge to examine the fit between knowledge-informed approaches and users. Our results highlight that knowledge-informed machine learning is a promising paradigm to enrich former data-driven systems, yielding explanations that can increase formal understanding, convey useful medical knowledge, and are more intuitive. Although complying with medical conception, it still needs to be investigated whether knowledge-informed explanations increase medical user acceptance and trust in clinical machine learning-based information systems.**

***Impact Statement*—The majority of investigations of the explainability challenge are being conducted from a developer-oriented focus, typically summarizing end-users based on their role or machine learning expertise. However, users are far more heterogeneous, with varying backgrounds, experiences, and needs. This motivates a recent surge of interest in explanations that account for multifaceted user requirements. However, how to effectively develop user-centric explanations is still unclear, and research lacks an understanding of which role users knowledge plays in developing satisfactory explanations. This synopsis acknowledges the potential of knowledge-informed machine learning for richer explanations. It is among the first to investigate how this strengthens user understanding from a knowledge perspective. It pinpoints knowledge characteristics of the fit between system explanations and users, which can guide the design of more user-centric clinical information systems.**

***Index Terms*—Artificial intelligence (AI) in medicine, explainable AI, human-centered AI, interpretable AI, knowledge-based systems, machine learning.**

Manuscript received 8 September 2022; revised 31 October 2022; accepted 3 December 2022. Date of publication 6 December 2022; date of current version 21 July 2023. This work was supported in part by Research Campus M^{2}OLIE and in part by the German Federal Ministry of Education and Research (BMBF) within the Initiative "Research Campus – Public-Private Partnership for Innovation", under Grant 13GW0387A. This article was recommended for publication by Associate Editor M. Popescu upon evaluation of the reviewers' comments. *(Corresponding author: Luis Oberste.)*

The authors are with the University of Mannheim, 68161 Mannheim, Germany (e-mail: oberste@uni-mannheim.de; heinzl@uni-mannheim.de).

Digital Object Identifier 10.1109/TAI.2022.3227225

I. Introduction

MACHINE learning-based information systems (ML-based IS) provide groundbreaking enhancements in many healthcare applications such as diagnosis, treatment selection, and surgical interventions, not least improving patient health [1]. Despite powerful predictive performance, their "black box" characteristic impedes them from fully exploiting their potential in practice [2], [3], as they are unable to explain their recommendations, decisions, or actions to human observers [4]. However, understanding the predictions and how these were generated is a particularly critical requirement in healthcare [5], where incorrect decisions made by machines can have life-threatening consequences for patients [6], [7]. Moreover, biases, diagnostic errors, as well as over- or underdiagnoses must be identified [8]. To counteract the opaqueness, research on "explainable artificial intelligence" (XAI) has accelerated. A vast number of XAI methods have been developed, and their impact and consequences are being researched from various angles [9]. Explanations provided by these methods display, for instance, comparable input-output pairs, decision trees, or feature importance measures; through partial-dependence plots, Shapley additive explanations, and other common techniques [5], [10], [11], [12]. While the complexity of data, models, and tasks increase [4], and recent legislation such as the EU's General Data Protection Regulation advances ethical and fairness considerations in artificial intelligence (AI) based decisions, XAI becomes evermore relevant to clinical IS.

II. Problem Statement and Research Goal

XAI techniques are frequently *data-driven* since they rely their results on the features or examples that make up the input data, such as probabilities or influence scores. Those explanations, however, may rather be useful summary statistics than faithful explanations of original models' actions [13] and leave the decision of what the relevance information might mean to the user [14]. Instead, relating them to more abstract concepts is better understandable and less misleading [15]. For deep learning models, the XAI community pursues generating local approximations [16], which, however, might be useful to experts for model correction but misleading to lay users [16]. The effectiveness of explanations further diminishes in clinical practice [17]. For instance, the relevance of "age" or "admission type" for a prediction is negligible if all potential follow-up

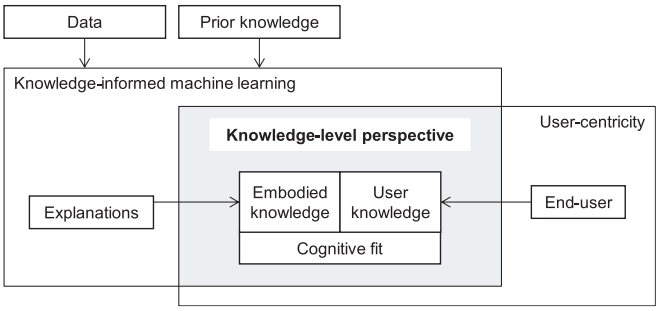

Fig. 1. Research focus and related concepts.

interventions are independent of age, or clinicians already know the reason why a patient is admitted, respectively [18]. Hence, data-driven XAI often mismatches human conception [19] and is not seen to give sufficient trust in a clinical setting [20].

Desiring explanations as well as understanding and trusting their rationales heavily depends on the user [11], [21], [22]. Most XAI investigations have been conducted in the computer science community [23] and their designs mainly reflect developers' needs [5], [24], [25]. In many cases, the type of user to whom explanations are targeted is misrepresented [26]. To make matters worse, non-ML-experts were considered as a *single* group of laymen [27]. However, recipients of explanations are far more heterogeneous, with physicians, patients, other medical staff, or managers involved in the use of medical AI systems, all with varying backgrounds and experiences [25]. The explanation requirements of individual users, who are typically not ML experts, are neglected. Implementing these demands is an urgent need for *user-centric* clinical ML-based IS [11], [28], [29], [30], [31]. Today, however, it is still unclear which XAI methods, and in which combination or situation, are effective for user-centric explainability [5], [26].

A key to more satisfying explanations is to align them with domain knowledge, as clinical users expect the systems to leverage validated and clinically relevant information [17], [20]. Data-driven approaches cannot use such knowledge jointly with the patterns learned from data to generate predictions. Meanwhile, the synergy in combining data-driven and knowledge-based systems demonstrated robustness, noise tolerance, and higher performance [32], [33], [34]. It has been concluded that integrating prior knowledge, e.g., encoded in medical ontologies, taxonomies, lexica, rules, or databases, into ML is also essential for better explanations of their functioning [19], [20], [35]. Likewise, the *cognitive fit theory* (CFT) postulates that a correspondence between the information on a task and the cognitive preferences of a user leads to effective and efficient problem-solving [36] and a higher influence of explanations on the user [37]. Thus, we find that a fit between the knowledge: fused into the ML system and embodied in explanations, and the explanation receiver is crucial for successful explainable clinical ML-based IS (see Fig. 1), and seek to answer the following questions.

1) *How can clinical knowledge-informed ML-based IS integrate prior knowledge to provide explanations?*
2) *How are the clinical users characterized and which knowledge-informed explanations are provided to them?*
3) *Which types of information are embedded in the explanations to be congruent to user expertise, i.e., improve knowledge-level fit?*

III. Conceptual Foundations

A. Explainable Artificial Intelligence and Interpretable ML

In many scholarly XAI papers, personal notions influence what comprises an explanation, lacking a well-defined vocabulary and standardized assessment criteria [11], [21], [38]. In general, an XAI system makes its functioning intelligible to a user [11], [39]. *Explainability* refers to all attempts to explain a model's prediction, uncover its inner workings, or present it with coherent expressions [40]; for instance, by providing textual or diagrammed explanations. There are philosophical [41] and medical [42] debates over what can be tolerated as an *explanation*, for instance, between evidence-based medicine (based on randomized controlled trials) and mechanistic reasoning (based on causal, e.g., physiological mechanisms between intervention and outcome) [43]. Clinical practice also tolerates statistically-based evidence that an intervention *does* work, even with mechanistic uncertainty *why* this is the case, so it was argued that similar requirements should be placed on clinical ML-based IS [44]. An explanation arbitrates between a human and a model [45] and is a statement, fact, or situation that renders a *target* (local, group-wise, or global) into some kind of understandable *pattern* (disease-related, statistical, or diagnostic) [46], [47]. Explanations that have a *local* target refer to patient-specific predictions, for instance, about an individual prognosis or diagnosis outcome. *Group-wise* explanations address population-level aspects, e.g., patient subgroups with a certain disease. Explanations with a *global* target unveil the workings of the entire model. A *disease* pattern, for instance, can be a pathophysiological mechanism for tuberculosis known from biomedical research. *Statistical* explanations highlight differences in data parts (e.g., between infected and non-infected patients). *Diagnostic* explanations are those a clinician would use when manifesting a disease based on a set of symptoms and justifying the diagnostic process [46]. In XAI, such patterns result from relationships contained in the data or knowledge, decision-relevant parts of the used representations, or active parts in the algorithmic model [38], [48], [49]. They can be found either by designing an inherently interpretable model underlying the system or by complementing a black-box model with an interpretable and faithful explanation, without accessing its inner workings [10], [48], [50], [51]. For the latter, a model is queried in a controlled way, e.g., by perturbing the input around a prediction of interest to reason about the local behavior. Occlusion methods cover parts of the input neutrally, e.g., grey squares in images. By comparing how the output changed, influential parts of the data are detected. *Interpretability* (or interpretable ML) was coined by the computer science community [10]. Gilpin et al. [41], for instance, suggest that interpretability is the extent to which explanations are understandable to humans. According to our view of *explainability* which also respects inherently

interpretable models, we use both terms synonymously [40], [52].

B. User-Centricity

Initial efforts for effective XAI have been conducted by developing methods to assess the quality of explanations, often distinguishing functionally-grounded, human-grounded, and application-grounded evaluations [53]. The former uses common metrics such as hit (match between predictions of black box and interpretable model), fidelity (the global ability of a posthoc explainer to mimic the black box), or complexity (e.g., number of premises in a rule). Human-grounded and application-grounded evaluations use simplified experiments with lay users and real users in a real-world application setting, respectively. Since the understanding of explanations heavily depends on the audience [11], [45], a growing body of work has recognized the role of users and their heterogeneous explainability needs. *User-centric* design [54] has been pursued to design effective XAI mechanisms [55] that unite the primary system's purpose with users' demands, taking their context, background knowledge, experiences, and expectations into account [56]. Resultant frameworks typically categorize them by their functional role in the AI application (e.g., operator, executor, developer, audience, deployer, or regulator [25], [28], [55], [57], [58]). or by their ML expertise (e.g., lay user, model user, developer [59], [60]). Derived needs may range from inspecting individual instances and features, monitoring performance, and validating model behavior, to seeking confidence and trust [27], [61], [62]. While most of the extant frameworks assigned needs to single user groups, Suresh et al. [27] frame these needs as independent components that may cut across all users.

Similarly, we characterize the domain knowledge and ML expertise of users and capture how these fit ML-based IS explanations from a knowledge-level perspective. In [27], we distinguish between knowledge *types* (formal, instrumental, personal), as well as *contexts* (ML, data domain, environment).

1) Formal knowledge refers to the understanding of codified information, such as documentation, rules, or laws, that is usually acquired through an educational process.
2) Instrumental knowledge denotes the ability to apply formal knowledge.
3) Personal knowledge is constituted by experiences gained through participation in specific domains. Each type can equivalently refer to a context: to research, develop, operate, or deploy ML; to collect, organize, analyze, or communicate data; and regarding the physical, social, or cultural environment of the AI application.

C. Knowledge Graphs and Knowledge-Informed ML

While data-driven ML might not conform to constraints provided by natural laws, regularities, or available knowledge [63], *hybrid* architectures have been advocated for decades to link symbolic operations with ML mechanisms [64]. Attention has been paid to analytical learning to make generalizations from data examples with domain theory [65], rule-based approaches as per inductive logic programming [66], and statistical relational learning to acquire relational information from data [67]. The hybrid paradigm resulted in a large volume of contributions from several communities inside and outside ML, including Semantic Web, natural language processing (NLP), neural-symbolic integration, and cognitive science.

Knowledge-informed ML learns from a hybrid information source, by integrating, enriching, or combining a dataset with a separate source of knowledge [63], [68], [69], ranging from scientific knowledge (principles from a scientific discipline), world knowledge (common sense) to intuitive expert knowledge (gained implicitly through experiences) [63]. It requires some form of explicit representation such as logical rules, equations, constraints, simulations, or knowledge graphs (KGs). A KG is a large semantic network that integrates information by interlinking data instances (entities) with relations (edges), arranged as triplets (head, relation, tail). Possible entity classes (concepts) and relation types are defined by an ontology [70], [71]. A taxonomy is a standardized vocabulary that classifies terms into a tree-like hierarchy without directional relationships [72], such as the well-known International Classification of Diseases (ICD) which categorizes diseases, injuries, conditions, and external causes.

Knowledge can be incorporated into the *data*, *algorithm* (during model and hyperparameter configuration, learning algorithm, or inference), or *explanation*, but also as *human feedback* [63], [69]. KGs, for instance, can be represented as continuous, low-dimensional vectors by embedding techniques. In general, embeddings convert data into a low-dimensional space while capturing their semantics. Graph embeddings specifically preserve graph information, including its structure, nodes, and attributes. As input to ML, they serve as a computationally efficient way to solve graph analytics problems such as node classification and link prediction [73]. The attention mechanism gained interest to additionally learn how important related tail entities are for a particular node [74], helping a model not only focus on the embedded node itself but include valuable information from neighbors.

An example of model configuration is a Bayesian Network, whose structure, relations, and probabilities experts can predefine. Knowledge infusion during model learning can be achieved using regularizing terms in loss functions to penalize patterns that are inconsistent with the knowledge [35]. On an explanation level, knowledge integration can happen, e.g., by correcting rules. Humans themselves can be involved in the loop and teach ML models. Automation and self-service approaches empower non-ML experts to perform data analytics [75], e.g., through lossless data visualizations for visual model discovery [76] or active learning [77]. Visual analytics research developed interactive interfaces to explore underlying data, alter representations, refine ML models, and detect biases.

D. User Acceptance in Knowledge-Based ML Systems

System explanations not only improve usability but also promote more positive user perceptions and ultimately, system acceptance [37], [78]. Regarding individuals' decision

performance, CFT states that a "cognitive fit" exists between user and machine if the mental representation of the problem (determined by a person's knowledge, skills, and experience with the domain and tool) and the task to be solved (the presentation format and instructions of a task) involve the same type of information [36]. In such a situation, decision-makers do not need to transform their mental representation into a form suitable for accomplishing a particular task and fewer cognitive resources are required [79]. In XAI, this fit has been presupposed for *meaningful* explanations to humans [52], increasing user engagement, interaction, and influence of explanations [37]. Hence, knowledge embodied in explanations should match user preferences, encouraging a cognitive fit and facilitating the acceptance of explainable ML-based IS.

E. Related Work

Beyond extensive reviews in the XAI field [10], [11], further contributions have been made to state-of-the-art XAI methods in terms of their purpose, challenges, mechanisms, taxonomies, and objectives (e.g., [40], [80], [81], [82], [83]). Tjoa and Guan [40] discussed a broad range of medical XAI mechanisms, by distinguishing perceptive interpretability from interpretability by mathematical structures. Increasingly, neural-symbolic techniques are used to explain neural networks [84], and research addressed the extraction of symbolic representations from them [81], [85]. In comprehensive reviews of explainable ML [80], [86] advantages and examples of hybrid AI have been described that are centered on ontologies and KGs. A knowledge-level perspective on different forms of knowledge integration through Semantic Web technologies, to facilitate explanations was taken in [87]. Calegari et al. [88] offered a very detailed technical overview of hybrid XAI systems, categorizing the papers according to the goals of trustworthiness, causality, and informativeness. Another related study [89] compared data-driven and knowledge-aware XAI based on methodological properties, evaluation metrics, and application context. In visual analytics, opportunities through visualization and interaction to involve end-users with different expertise, namely AI novices, data experts, and AI experts, were discussed [90]. While these reviews are highly valuable, past research lacks: a cross-sectional review of prior knowledge integration in ML: while assessing how resulting explanations fit users from a knowledge-level perspective; and considering the healthcare domain. In extending prior surveys on medical XAI, we aim to examine knowledge-informed XAI for user-centric explanations of ML-based IS.

IV. Methodology

We conducted a systematic, cross-sectional literature review by searching IEEE Xplore, ACM Digital Library, PubMed, Web of Science, as well as ScienceDirect. Our search strategy followed a structured approach [91] and was based on the PRISMA framework [92] (see Fig. 2). Search terms included "explainability," "healthcare," and "AI," as well as topics or techniques to involve additional knowledge (e.g., knowledge-based, ontology, Semantic Web, interactive). For each, a broad range of syntactical variants and synonyms were included. The search was restricted to English articles published between 2016 and 2021. The identified studies were screened for relevance in terms of a healthcare-related topic and paper type (except panels, tutorials, doctoral consortia, or surveys). The remaining articles were finally assessed for eligibility. We included papers that propose an ML-based IS in a healthcare-related context with explicit attempts to explain its reasoning or present the system interpretable to users while utilizing a separate source of knowledge. We excluded papers without an ML-based application, a domain-specific explanation attempt, or the integration of external knowledge.

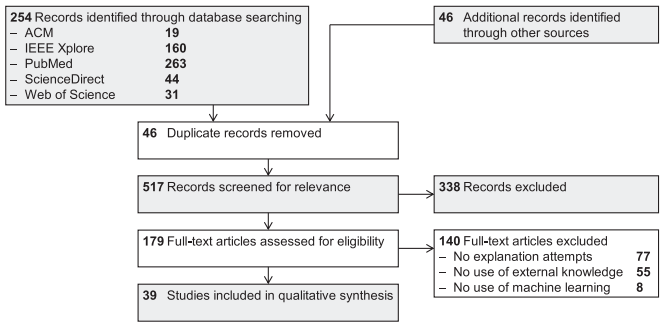

Fig. 2. Literature search results (schema adapted from Moher et al. [92]).

We structure our review as follows. In Section V-A, we examine the types and sources of knowledge utilized for informed ML tasks, as well as the way this information is integrated. In Section V-B, we study the characteristics of the explainable clinical ML-based IS about how information is presented to the user, including the types and scope of explanations, the mechanism they are evaluated, as well as the targeted users. Finally, we identify the knowledge occurring in the explanations of ML-based IS according to the knowledge-level framework [27] in Section V-C. Ultimately, this allows pinpointing in which cases an explanation meets users' understanding. We discuss how this additional knowledge contributes to user-centricity, and pose future directions for explainable, clinical ML-IS in Section VI.

V. Results of the Review

We synthesized 39 papers for our review, including 13 ML classification tasks regarding *general medical* XAI problems, 19 *disease-specific* XAI, and 7 *ICD coding-related* XAI, as given in Table I. General medical XAI tasks, such as predictions of drug interaction [93] and disease diagnosis [94] have been developed from early on, using manually authored ontologies to compute drug similarities and disease probabilities, respectively. More recent approaches linked data items to KG concepts to perform further computations such as disease prediction [95], medical text classification [12], [96], or doctor recommendation [97]. Others involved human interaction in disease progression modeling and adverse event prediction [98], [99], the former to associate clinical measures with different stages of diabetes development, and the latter to predict events like bleeding in prospective three days. In contrast to predicting a diagnosis from

TABLE I
MEDICAL XAI TASK BASED ON UTILIZED PATIENT DATA

General		
Adverse event prediction	Ta:	[98]
Disease prediction	Ta:	[94], [95], [125], [134], [143]
	Tx:	[103]
	Im, Ta:	[144]
Disease progression modeling	Ta:	[99]
Doctor recommendation	Ta:	[97]
Drug-interaction prediction	Ta:	[93]
Text classification	Tx:	[12], [96]
Specific health condition		
Atrial fibrillation prediction	Ta:	[19], [128], [133]
Breast cancer prediction*	Ta:	[122], [145]
Circulation disease prediction	Ta:	[20]
Colon cancer prediction	Im:	[127]
CAD prediction	Ta:	[116]
Diabetes prediction	Ta:	[104]
Heart failure prediction	Ta:	[74], [106]
High-cost risk prediction of COPD-affected patients	Ta:	[111]
Lesion annotation	Im:	[109]
Lung nodule classification	Im:	[110]
Mortality prediction of AKI-D-affected patients	Ta:	[117]
Skin cancer prediction	Im:	[100]
	Im, Ta:	[129]
Spinal disease prediction	Im:	[126]
Stroke prediction	Ta:	[118]
ICD-related coding		
Sequential diagnosis prediction	Ta:	[101], [102], [105], [107], [119], [120]
Text classification	Tx:	[108]

AKI-D = acute kidney injury requiring dialysis, CAD = coronary artery disease, COPD = chronic obstructive pulmonary disease, Im = Image, Tx = Text, and Ta = Tabular data.

*General classification which was evaluated for breast cancer prediction.

a broad set of diseases, most recent knowledge-informed ML classified specific diseases, for instance, images of pigmented lesions into benign, suspicious, or malignant skin cancer [100], highlighting the interest in specialized explainable ML-based IS. Finally, ICD-related coding was a technically more specific medical XAI task, in which the system outcome is a diagnostic code according to an ICD standard. Sequential diagnosis prediction generated disease codes for the next visit based on the codes given by past visits. As demonstrated by Choi et al. [101] and Ma et al. [102], the hierarchical knowledge of ICD can be captured to learn meaningful representations of the data.

A. Findings Regarding Knowledge Integration for Knowledge-Informed ML (RQ1)

This RQ explores the: types of knowledge sources and their integration into ML-based IS.

1) Types of Knowledge Sources:

a) External knowledge: Prior knowledge often originated from freely available clinical and biomedical knowledge bases [96], [101] or medical textbooks and encyclopedias [103]. BioPortal, for instance, contains more than 1000 sources for, among others, abnormal phenotypes related to diseases and genes (e.g., diabetes mellitus complications and symptoms [104]). The ICD-9-CM, used in U.S. public health, was a common taxonomical source of diagnostic information, organizing diagnostic codes (e.g., diabetes with ketoacidosis) as a hierarchy with "is-a" relations to their parents (e.g., disease of other endocrine glands). Complementary taxonomies served to collapse ICD codes into clinically meaningful categories [101], [102], [105], [106], [107]. ICD-related taxonomies allowed for the replacement of medical codes in the data with an upper, more abstract level of the hierarchy, to reduce the label space of over 14 000 full-length ICD-9 diagnosis codes [101], [105], [107].

External medical knowledge was also acquired as a KG representation. Teng et al. [108], for instance, harvested general knowledge facts as triple data from wiki contributions to build a KG. The constructed ontology consisted of disease, symptom, medicine, surgery, and examination entities. Since KG content not always perfectly matches annotations in the data, NLP techniques were required to match the phrases in the text with concepts in a KG [12], for instance, matching short phrases describing each ICD code with diagnosis information in the data. Graph embeddings, corresponding to ICD codes, made latent representations more effective for medical text classification, improving the model's terminological understanding [108]. However, the approach did not consider the hierarchical nature of the ICD taxonomy and predicted infrequent codes less accurately. For specific medical terms, Yan et al. [109] used information about body parts, types (such as nodules or liver mass), and attributes (such as intensity or shape) from a radiology lexicon to build a lesion-specific KG.

To integrate multiple different and potentially heterogeneous KG sources, ontology definition, semantic annotation, and ontology linking were performed [93], [95], [96]. For instance, Yu [95] used DBpedia class terms to define the ontology and acquired text data about common health conditions by Web crawling. It was designed to: overcome shortcomings of approaches that match knowledge sources based on word occurrences (potentially missing concepts due to writing styles) and provide probabilistic outcomes. Their solution directly mapped meaningful terms to the corresponding concepts in the ontology and extracted graph triplets for probability computations. Based on symptoms, disease predictions were generated as locally interpretable graphs with promising accuracy on different disease-specific data subsets, though the interpretability for doctors was not evaluated. Finally, the ontology linking step served for identifying connections between different ontologies and KGs. Fokoue et al. [93] integrated various drug-related public web sources—which were incomplete in isolation—for drug interaction prediction, including drug-gene-treatment relations, interacting genes, as well as gene-disease relations. With this KG, drug similarities were computed and collected in candidate feature vectors for downstream logistic regression. The explainability of such a KG was not evaluated.

b) Human sources: Knowledge *manually* endorsed by human experts was observed in three ways. First, medical experts crafted diagnosis-specific features which they additionally annotated to the data, such as the shape, property, and size of lung nodules for malignancy classification [110]. Li et al. [19] formulated this as an object detection task: Key feature points in electrocardiograph (ECG) data, that experts use to judge instances, were learned. The increased performance demonstrated the effectiveness of guiding a deep network with domain knowledge for disease prediction. In contrast, detecting

more than one type of expert ECG pattern turned out less effective and to potentially distract the model.

As a second way, experts' knowledge was incorporated into the construction of both ML models and KGs. A traditional approach was to pre-build Bayesian network structures, for instance, of diseases, their causes, symptoms, and demographics [94], as well as procedures and costs [111]. Domain experts could fill in a priori probabilities (e.g., of a disease) and conditional probabilities (e.g., of symptoms given a certain disease). Next, a model could infer posterior probabilities that a disease is present, given certain symptoms. Bayesian networks have been successfully applied in many medical studies [112] and were also combined with deep learning to generate image features automatically upfront [113]. Although experts may provide biased or imprecise information [114], the models can generate accurate and interpretable results [115]. Nevertheless, the approaches have not been evaluated to fulfill the explainability requirements of medical doctors and still must prove effective in practice. Other studies linked ML models with KGs [20], [97]. Sun et al. [20], for instance, manually built a KG of circulation diseases and risk factors, as their goal was to integrate evidence-based medical knowledge. The challenge was that not all patient features in the data could be mapped to the KG whilst graph-based embedding methods require all data to be arranged in a graph. Therefore, an autoencoder obtained separate patient embeddings, which were fused into a reinforcement learner (see section "integration through path-based ML"). While the achieved performance was as high as that of modern data-driven ML, this approach could not predict the onset of diseases whose causes are unclear, i.e., have undefined paths.

Finally, medical experts served with their preferences in several ML-based IS to help validate or control the outcomes (e.g., [100], [116], [117]). A human-centered approach to explaining support vector machines for coronary artery disease prediction [116] involved target users for rule validation. By visualizing attribute pairs with corresponding two-dimensional (2-D) support vector machine hyperplanes, rules were manually crafted and clarified by a medical expert. Validated rules were stored in a knowledge base and used for inference in a verbalized form. It was perceived as accurate and explainable, though evaluated by only one expert. Prentzas et al. [118] extracted rules from a tree-based model which they used as input to an argumentation system, where rules were incrementally tested for conflicts and prioritized by a medical expert. Their method reached an accuracy superior to a support vector machine for stroke prediction. As a specialty, the system outputs all possible options with corresponding explanations to the user. Although the utility of explaining not a single but all outcomes to the user was not practically evaluated, domain experts were expected to make more informed decisions. For involving humans in the loop, the field of visual analytics specifically contributed interactive interfaces, for instance, to explore medically important parameters within a patient's history [119], [120] or build and refine cohorts [99]. Medical experts could apply their domain knowledge to set specific hyperparameters and constraints, e.g., that the progression of chronic diseases cannot step backward [99]. Such interactive features were found useful by medical researchers to construct subgroups for clinical analyses and find relationships to health outcomes [99]. In other interactions [121], [122], experts could design classification boxes, to separate classes by reordering coordinates in a hierarchical process. Spruit and Vries [98] allowed multiple points of interaction along the ML pipeline, including data configuration, understanding, preparation, and ML modeling.

TABLE II
KNOWLEDGE INTEGRATION INTO DATA-DRIVEN ML-BASED IS

Knowledge graph embeddings	
Hierarchical ICD code embeddings	[101], [102], [106],
Medical concept embeddings	[74], [103], [106], [108], [117], [120]
Path-based models	
Causal KG with probability-based graph generation	[95]
Extract semantic paths for downstream ML model	[97]
Random walk over KG using reinforcement learning	[20]
Human-in-the-loop	
Integrate expert rules in a fuzzy rule base	[100]
Interactive controls of data and ML model	[98], [99], [119], [122], [145]
Postprocess ML outcomes based on medical knowledge	[116], [118]
Predefine ML model structure based on medical knowledge	[94], [111]
Semantic logic	
Knowledge-based semantic information for ML model	[93], [96], [104], [105], [107], [125], [126], [144]
Concept matching for frequent item set mining	[12]
Learn expert-related semantic features jointly	[19], [109], [110], [127], [128], [133], [134], [143]

2) Integration Into ML-Based IS: To use the acquired knowledge in ML, the ways of integration are presented subsequently, as given in Table II.

a) Integration through KG embeddings: KG embeddings found their way into explanations through attention mechanisms, to form the critical parts ML models look at. We found two applications: *hierarchical code embeddings* and *medical concept embeddings*. The first type is based on codes in the ICD taxonomy. One such approach [101] was based on a co-occurrence matrix taking the local context and global co-occurrences into account. Medical code embeddings were obtained as a combination of nodes and their ancestors via an attention mechanism. Finally, embeddings of codes that occur within a patient's stay were combined to visit representations that were used as input to a recurrent neural network (RNN) for sequential diagnosis prediction. The advantage was that final representations were more structured and consistent with the KG, compared to embeddings learned only through co-occurrence. A generalization of this approach [102] added a knowledge attention layer on top of the RNN output to learn and infuse ancestor code embeddings, which improved the performance even further. Regarding the interpretability of this approach, the structural quality of the embeddings increased and the attention layer allowed for observing the attention behavior of disease codes for individual patient visits. However, the interpretability was not evaluated by users.

Second, embeddings of *medical concepts* represented more complex semantic relations than hierarchies. Yin et al. [74]

gathered "causes" and "is-caused-by" disease relations from scientific publications, health portals, and online communities, and learned embeddings of corresponding medical diagnosis events. In addition, they encoded the time intervals between patient visits and fused them into a long short-term memory (LSTM) using an attention mechanism. As a specialty of their architecture, all output vectors of the model were concatenated in a global max-pooling operation to prevent the model from forgetting early medical events in long sequences. Medical knowledge led to higher accuracy compared to other attention mechanisms. At the same time, the method allowed computing the magnitude of each event's contribution to the predicted heart failure risk over time, distinguishing the contributions of the input data and the KG attention. This differed from former local, multilevel interpretability, which visualized contribution rates of visits and events within visits. The temporal dimension was emphasized to facilitate physicians' associations of past medical events with patient outcomes, though a human-grounded evaluation was not conducted.

Zhang et al. [106] embedded more disease relations, including "is-healed-by," "is-alleviated-by," and "risk-reduced-by," for heart failure risk prediction. The authors adopted two independent LSTM models in which they fused both visit embeddings and medical code embeddings. This helped the model to extract twofold KG information, through which performance increased and contribution rates could still be obtained. Outputs were weighted via an attention mechanism for the final heart failure risk prediction. Interpretability was facilitated by displaying a timeline of a patient visit based on event contributions, irrespective of the time intervals between the visits. This was evaluated by experts (see section "feature relevance"). Liu et al. [117] learned concept embeddings of abnormalities from a biomedical ontology. While previous graph attention models were targeted to discrete medical event data (e.g., diagnosis codes), the mortality prediction involved continuous temporal data (e.g., blood pressure measurements). Therefore, the authors proposed to match features in the data to KG concepts and grouped them based on the ontology, for instance, into cardiovascular and respiratory abnormalities. After group embeddings were obtained, LSTMs were leveraged for each feature group and combined via attention. The approach was specifically developed to learn both feature and time attention. This way, the local risk trajectory of a patient could be assessed by the overall importance of feature groups (through group-wise attention) and influential points in time before the outcome, though its explainability was not evaluated.

b) Integration through path-based ML: In this category, ML models directly operated with path-based information from the KG. *Probabilistic chains* of diseases were developed in [95]. Based on input conditions (e.g., symptoms, age, and gender), the graph generation method presented possible diseases (e.g., bronchitis) and causal relations to the input conditions, along with a probability distribution (e.g., cough: 0.0625 and age: 0.5, etc.). As an advantage of this approach, symptoms and diseases can be related to each other by long chains of evidence to explain system outcomes, including "causes" and "is symptom of' relations." An example of a chain is "*rheumatoid arthritis* $\rightarrow$ *psoriasis* $\rightarrow$ *psoriatic arthritis* $\rightarrow$ *myositis* $\rightarrow$ *inclusion body myositis*" [95]. An assumption due to the probabilistic nature of this approach was that input conditions are fully independent. It was assumed that this is not realistic in certain clinical cases, particularly when symptoms naturally occur together.

A way to use semantic paths with tabular has been demonstrated in [97]. First, possible semantic relations between patients and doctors (e.g., "doctor relates to the same disease of a patient") were defined. These, so-called interactive features, were extracted from a KG and constituted the inputs to a *downstream ML* model, together with individual features about doctors (e.g., consultation costs and number of patients). Given a patient and a doctor with both types of features, the deep neural network predicted whether the patient would consult the doctor or not. Using a model-agnostic technique called layer-wise relevance propagation [123], relevance scores of interactive and individual features were computed. The scores served as local explanations for the recommended medical doctors but were not evaluated by system users. Third, reinforcement learning was used in [20]. After mapping a patient's disease to a KG entity, the model was trained to start walking from there and stop at a finally predicted disease after T steps. Regarding the interpretability of this system, the predictions were made understandable through disease progression paths including transition probabilities, i.e., assigned probabilities to the paths between the given disease and the terminated (i.e., predicted) one. All possible disease progression paths were offered as explanations for a patient's final prediction. Although the explainability of the method was advocated by a doctor, the authors did not elaborate on an evaluation method.

c) Human-in-the-loop integration: Approaches of this category play a special role in XAI since they can interactively convince users instead of simply presenting explanations [122], [124]. In [100], the role of expert knowledge was twofold: On the one hand, the diagnostic ABCD framework for skin cancer was used that medical experts follow when assessing images. From this prior knowledge, *fuzzy rules* about the criteria asymmetry, border, color, and differential structures were generated. A human expert was finally involved in resolving rule conflicts, adding rules, or correcting them. The conceptual study, however, was neither implemented nor evaluated. Besides, it was argued that the expressions on which fuzzy logic systems operate (e.g., low, medium, and high) were more satisfactory than numeric values and that showing triggered rules increases explainability. In self-service approaches [121], [122], users were able to discover patterns with the help of compact but lossless data visualizations, i.e., preserving the multi-dimensionality of the data. Therefore, data were visualized by a graph that spanned sequences of 2-D coordinate systems [122]. An expert could iteratively use the system to change the order of coordinate pairs, swap coordinates, or perform non-linear scaling to find good class separation patterns. Wagle and Kovalerchuk [121] introduced inline coordinates, in which coordinates can be arbitrarily arranged on a common axis and reordered by a user. In both visualizations, classification boxes were constructed in a sequential, interactive process. For domain experts, this provided the opportunity to globally assess and influence the underlying

model, e.g., to modify its complexity and control overfitting. Based on the classification boxes, both approaches extracted and simplified analytical rules which are interpretable for humans. An interpretable and interactive visualization for sequential diagnosis prediction was proposed in [119]. The authors adapted the attention mechanism of an RNN to additionally incorporate information on time intervals between patient visits and enhance interactivity. As a specialty of their approach, the model uses two embedding matrices, one to compute attention weights and one to compute the final prediction output. This way, a medical expert could alter the influence of individual codes (retraining the model based on the second matrix) without affecting the overall attention weights.

Other types of human-in-the-loop approaches manually *validated extracted decision rules* against medical knowledge in a downstream task [116], [118] (see section "rule-based explanations"). Agarwal et al. [94] and Lin et al. [111] proposed to use *expert-predefined Bayesian networks* and learn the final structure and probabilities of causal relationships from data. Given a set of patient observations, the model by [94] inferred corresponding diseases, which were displayed with their probabilities and contextual information in terms of relevant follow-up tests and corresponding pathology labs. Although the interpretability was not evaluated, the probabilistic approach was expected to yield more confidence in the results. A graphical model was also favored over other ML models for predicting future high-cost patients [111] because, on the one hand, prior knowledge could be used to model influencing variables, such as whether the patient has taken systemic steroids. On the other hand, the network and its causal relationships between predictors could be illustrated. This easily showed information, for instance, that steroids were a strong predictor. Although interpretability was not evaluated, it allowed medical experts to affirm the results.

d) Integration of knowledge-based semantics: Background knowledge contains *semantic information* that can be encoded in downstream ML. This includes encoding patient data as signals in a KG [125] as well as using semantic similarities based on KGs, to mine similar data points (e.g., [96], [105], [107]) or to engineer features based on concepts related to the data [93], [104]. Panigutti et al. [105], [107] used a similarity score that is based on the depth of two nodes (ICD codes in the hierarchy) as well as their lowest common ancestor. The similarity score was used to compute visit-to-visit distances based on the ICD codes of each visit. These, in turn, were used to compute patient-to-patient distances. The method selected the closest k data points to generate a synthetic neighborhood that served to query the black-box model. Based on the augmented data, a surrogate, inherently-interpretable decision tree model was learned. As a result of the encoded ontological structure, the approach had a higher fidelity than surrogate models using data-driven neighborhood generation and perturbation strategies. It yielded a better approximation of the black-box model and, therefore, became more trustable. Finally, the decision tree was transformed into rules as a local explanation to doctors, though its explainability was not evaluated by humans. El-Sappagh et al. [104] designed a fuzzy rule-based system with semantic capabilities. The system used linguistic information from an ontology for the features contained in the rules, to consider semantically-related concepts across rules. Features were grouped for individual rule-based systems, before weighing their results for the final diagnosis. This surpassed traditional ML techniques and was expected to be medically intuitive and interpretable, whilst only rule complexity was evaluated. Han et al. [126] developed a neural-symbolic learning framework for spine disease classification. A KG of spatial correlations and locations of spinal structures was embedded into a neural network, so that semantic segmentation of spine images was improved. As an interpretability component, an unsupervised logical reasoning approach was applied to find pathological, causal effects between segmented structure and target diseases. Although the semantic representation and resulting if-then rules were considered interpretable and beneficial for the diagnosis, a user evaluation has not been performed.

TABLE III
TYPES OF EXPLANATIONS PRESENTED BY KNOWLEDGE-INFORMED ML

Presentation	User feedback absent	User feedback included
Disease probabilities and relationships	[20], [94], [95], [111], [125], [144]	[99]
Feature relevance	[74], [97], [101], [102], [103], [117], [129]	[106], [108], [119], [120]
Rule-based explanations	[100], [104], [105], [107], [118], [121], [122], [126]	[116]
Explanations based on semantic characteristics	[12], [93], [96], [109], [110], [127], [128], [133], [134], [143]	[19]

Semantic information could also be integrated through *textual matching* [12]. Moradi and Samwald [12], for instance, applied item set mining to explain black boxes but criticized that data-driven item sets rely on frequency, which might not properly reflect decision logic. For this reason, they extracted biomedical concepts from text using NLP techniques and matched them with KG concepts. Then, a confident item set mining method was applied which leveraged the strength of the relationship between concepts and class labels. It achieved higher fidelity than frequency-based item sets.

Another way of exploiting semantic knowledge was to *learn features jointly* with the classification task [19], [109], [110], [127], [128]. Thereby, approaches learned the classification label simultaneously with semantic characteristics through a combined loss function. Yan et al. [109] extracted semantic labels from radiology reports based on ontology concepts and formulated a multilabel classification task to predict them for lesion annotation of radiological images. In contrast to prior studies, this method expanded the label space by including missing parents according to an ontology. The label relevance was evaluated with radiologist annotations and showed higher performance than other multilabel classification methods.

B. Findings Regarding Provisioning and Targeting of Explanations to Users (RQ2)

The following section reviews how knowledge integration influences the explanations of ML-based IS, including: how users and their expertise were characterized and how it was

evaluated to meet their demands. As given in Table III, four clusters of presenting explanations emerged.

a) Disease probabilities and relationships: Like KG representations, one way to present and explain system outcomes is based on probabilities of diseases and their relations, such as disease transitions or co-occurrences. It can be achieved by path-based approaches (e.g., [20], [95]) and probabilistic graphical models (e.g., [94], [99]). As an example from [20], heart failure for a patient (probability of 32%) resulted from prior diagnoses of *hypertension, anemia, and diabetes* (with transition probabilities of 13%, 12%, and 11%, respectively). This approach was proposed to explain individual patient predictions to clinicians. Probabilistic networks allow for graphical presentations which can be interpretable—depending on their complexity. A Bayesian network [111] was illustrated as a whole and helped healthcare providers globally understand the high-cost probability and cost trajectory of patients, for instance, that co-occurring cancer increased the number of admissions. A hidden Markov model [99] learned disease progression as a series of transitions between hidden states. It was explained both locally (by viewing state transitions, health outcomes, and other variables of individuals) and group-wise (by cohort definition, e.g., based on genetic characteristics or biological attributes, and hypothesis testing). The visual analytics tool was designed for clinical researchers to investigate disease progression patterns in observational data collected from clinical studies and allowed them to define subgroups of patients and test their relationships with various health-related outcomes. Users' requirements and experiences with the interface were evaluated by four workshops and interviews with nine clinical researchers, respectively. The feedback confirmed perceived usefulness and reduced effort to data analysis. Although visualized more transparently, limitations were mentioned that hidden states remain difficult to interpret only based on features and are per se not intuitive, depending on the number of hidden states specified. As a merit of KGs, local disease progression paths were human-understandable and expected to be intuitive as well as advocated by clinicians for diagnosis tasks [20], [99].

b) Feature relevance: The relevance of elements in the input data was a common method to provide explanations. In the approaches that use KG embeddings, this was conducted by showing the amount of attention that elements received whereas the elements to which the algorithms attend differed. In Teng et al. [108], locally relevant text segments were presented as explanations for ICD code predictions. Interpretability was quantitatively evaluated based on a 5-word overlap of segments highlighted by experts and extracted from highly relevant and less relevant attention scores (< 0.8). It was accurate in 62.5% and 71.4% of the cases, respectively. Though it was evaluated only based on one example medical record, it confirmed that the method generally relies on important phrases in clinical text.

KG attention behavior could be examined by showing *ancestor relevance* as global [101] and local [102] explanations, i.e., the amount of attention that diseases and parental nodes receive. For example, a patient affected with myocarditis was predicted to be diagnosed with cardiac complications, probably because "coronary atherosclerosis" (a heart disease) received the highest attention in the KG [102]. The methods were designed to provide interpretable ICD code pre-selection to human coders, while their interpretability was not evaluated.

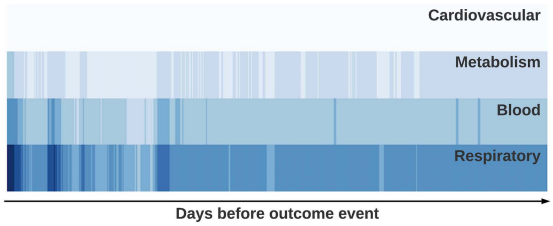

Fig. 3. KG-adjusted attention map, darker color means higher score. Adapted from Liu et al. [117].

Wang et al. [129] used diagnostic features of the ABCD framework as domain knowledge for skin images. These were computed arithmetically based on image segmentation using a pre-trained model and concatenated with patient demographic features for multi-modal skin classification. A posthoc explanator, Shapley additive explanations, was applied and indicated contributions of both patient and diagnostic features to the prediction. It was found to be easy to use for a dermatologist, though no evaluation has been conducted.

Yin et al. [74] computed both local influences of *input data* (medical events in patient history) and *knowledge* (events in the KG), to exhibit their contribution over time and help doctors better assess heart failure risk. As an example of local contribution rates of individual events [106], three codes (*"mitral valve disorders," "other primary cardiomyopathies," and "automatic implantable cardiac defibrillator in situ"*) that occurred in the last two visits of a patient led to a heart failure risk increase from 30.96% to 34.71%, meaning that later diagnoses were more decisive [106]. To verify the authenticity of such interpretations, a group of cardiologists was consulted, though no evaluation method was specified. The attention of feature groups over time was used as local explanations in [117] (see Fig. 3). The KG-based grouping was specifically designed to assist doctors in making timely decisions. It demonstrates the usefulness of not solely providing individual feature contributions, but an at-a-glance risk assessment.

Attention weights of features in the input data, in the time dimension, and of the ancestors of leaf nodes in the KG were visualized alongside patient history or presented as top ten lists. Interestingly, KG attention resulted in explanations to different users such as developers and clinical users, and for both groups, contribution rates of medical codes were provided. The approach that allowed to interactively manipulate attention weights [119] addressed a broad range of users, namely physicians, health professionals, and medical researchers. The expert feedback revealed that code-level contribution scores were too complex to be interpretable at first sight. Instead, interpretability was rather required when asked for it. As a result, the authors provided physicians with a simplified interface with only important events. The full functional range was dedicated to researchers for data exploration. Another visual analytics approach [120] visualized patient distribution and allowed for what-if analyses

to understand how medication changes cause variance in individual risk prediction results. A medical expert was asked to perform a set of tasks in a case study. Feedback pointed toward interaction costs of manual what-if analyses and cognitive load caused by visualizing the contribution of visits and codes, though the interface was perceived as useful and intuitive. Finally, an ongoing debate has been initiated on whether attention weights are causal indicators of input unit importance [130] since they were shown to significantly differ from other feature importance measures in certain cases [131], [132]. Besides attention, Yuan and Deng [97] applied a post hoc, model-agnostic method to extract feature importance and explain recommended physicians to users of healthcare consultation platforms. Importance scores were evaluated via quantitative sensitivity analysis by measuring the impact of removing individual features.

c) Rule-based explanations: Rule-based logic played a role when involving humans in the loop (e.g., [116], [118]) and providing rules as final explanations (e.g., [107], [122]). Thereby, sets of rules reflected how the global model functioned [121], and triggered rules explained individual patient decisions [100], [107]. From a surrogate decision tree for sequential diagnosis prediction [107], rules were extracted and presented as local explanations by visualizing the rule components in a timeline and highlighting the most informative codes. This was targeted at doctors and expected to be more informative since the ontology utilized in neighborhood generation led to higher fidelity and less complex rules in certain cases. Human-in-the-loop approaches involved domain experts such as physicians and healthcare providers both in the development of the systems and as users of rule-based explanations. Such interactions, for instance, were designed for non-technical users to define classification boxes in coordinate pair graphs [121], [122]. Extracted analytical rules reflected that if a data item falls within the rectangle, the corresponding class is assigned. As a strength of rules, computational techniques can be applied to simplify them. Though, not always were pure rules sufficient for medical users. Samuel et al. [116] explained coronary artery disease diagnoses by decision rules in two levels of detail. The goodness, satisfactoriness, and trustworthiness of both variants were rated by five medical experts using six items on a five-point scale. As a result, rules were understandable but solely naming the factors that triggered a rule was not perceived as convincing to explain diagnostic anomalies. Instead, additional background information, such as demographical data and the pharmaceutical history of the patient, increased satisfaction. Compared to what was also reported in CS [45], classification rules are easily understandable. Nevertheless, their textual representation does not immediately provide further information, such as the relevance of single attributes. Moreover, a rule explains only parts of the whole knowledge captured in a model. Another benefit of rule-based explanators was demonstrated in [105]. ICD codes that occurred in the rules of local patient explanations were computationally aggregated to explain over- and underdiagnoses in patient cohorts. The authors developed this approach for domain and healthcare facility experts aiming to audit the decision support system, although its effectiveness is yet to be evaluated by humans.

d) Explanations based on semantic characteristics: Semantic information can indirectly influence case-based explanations. This form of explanation discovers related samples in the data for comparison. Compared to data-driven distances, the distances of KG nodes or embeddings provided semantic similarity between the case to be explained and the related ones. Yan et al. [109] presented lesion examples that are similar according to the semantic labels in the radiology report, as a local explanation for radiologists. The method could retrieve lesions that looked different, but shared similar labels. This was proposed to guarantee that the radiologist understood the prediction, though the quality of the interpretable approach was purely evaluated based on predictive performance.

Other approaches extracted semantic wave patterns in ECG, which also experts use to judge atrial fibrillation [19], [128], [133]. Li et al. [19] first pretrained the model with these key features and then fine-tuned it for both feature detection and disease prediction tasks. The authors obtained contribution rates through a perturbation-based posthoc explanator and presented a contribution map as a local explanation to cardiologists. They computationally evaluated the approach by measuring the divergence between ground truth annotations and extracted regional importance. Cardiologist feedback was elicited, though no evaluation method was specified. This verified that the approach better matched the key feature distributions compared to the black-box model and that the tool is acceptable and attractive. However, perturbation-based strategies may generate random feature values that never appear in reality, resulting in meaningless explanations [12]. Bourgeais et al. [134] mapped biological concepts to neurons of a network, to predict phenotypes from gene expressions. A regularization term penalized connections that are not present in the gene ontology. Through layer-wise relevance propagation, one could extract neurons' biological significance in all internals of the model. Each hidden layer was visualized by semantic relevance scores for disease, disease subtype, and patient levels.

The examples show that although semantic features were learned, models remained opaque and required posthoc explanators. Despite that, explanations were more intuitive when highlighting semantically-important elements in line with the observation process of experts, like in the ECG example. Domain-specific features are convincing and understandable to domain experts [128]. They make them feel more confident, and visualizing their importance is appreciated [19]. Inherently interpretable text classification, in contrast, was achieved by semantically-enhanced association rules [12], providing local (patient-wise) and group-wise (class-specific) item sets of biomedical concepts from the KG and their confidence scores. For instance, '*combination chemotherapy is the cornerstone of treatment that confers a meaningful survival benefit for patients with small-cell lung cancer*' was classified as "*treatment for disease*" with a confidence of 1 due to the concept set {small cell lung carcinoma, combination chemotherapy}. The method produced fewer item sets with higher fidelity than other posthoc explanators but was not evaluated by users.

e) Evaluation of explanations: In most proposed knowledge-informed systems, explanations were assessed in subjective or

TABLE IV
KNOWLEDGE EMBODIED IN EXPLANATIONS OF KNOWLEDGE-INFORMED MACHINE LEARNING-BASED INFORMATION SYSTEMS

Context	Formal knowledge	Instrumental knowledge	Personal knowledge
ML	*A1)* KG contribution [74] Semantic model internals [134], [143]	*A2)* Contribution map [74], [106], [117], [127] Knowledge-related text fragments [103], [108] Semantic item sets [12] Semantic neighborhood [107] Interactive model discovery [122]	*A3)* Intuitive representation [101], [102] State-level user interaction [99]
Data	*B1)* KG exploration [93] KG chain [20], [95], [125] Pathological relations [126] Semantic KG features [97]	*B2)* Fuzzy rules with semantic features [104] Expert-defined feature relations [111] Informative arguments [118] Expert-validated rules [116] Self-service user interaction [98], [145]	*B3)* Experts' key features [19], [109], [110], [128], [129], [133], [144] Expert-driven fuzzy rules [100] Feature-level user interaction Related evidence [93] Semantically-related cases [96], [109]
Env.		*C2)* Diagnostic awareness, context information [94] Semantic interoperability [104]	*C3)* Group-specific auditing [105]

Demonstrates suitable contexts and types of the knowledge that is embodied. Schema adapted from Suresh *et al.* [27].

computational (functional-grounded) forms, only rarely with human subjects. In the former cases, authors discussed the outcomes of their models informally and presented reasoned arguments why an explanation was meaningful. For instance, Teng et al. [108] argued why highlighted text segments were medically relevant for selected examples, but also quantitatively measured how many times medical experts found the local explanation correct. When quantitative metrics were used as explainability proxies, it concerned the complexity of rules or item sets [12], [104] or the predictive accuracy of the explainable model [118]. Despite conjectural functionality, justifying whether explanations achieve understanding and trust in *real-world settings* is a fundamental challenge in XAI research [82]. This is crucial in the healthcare domain, as it has been reported that the expectations of system designers, which information will build users' trust, largely differ from the results in clinical practice [124]. Therefore, research already stressed scientific rigor by defining and evaluating interpretability [53]. Since human evaluation with real end-tasks provides strong evidence of success, user studies are important to collect feedback and evaluate the quality of explanations. However, this rigor was rarely the case in the reviewed knowledge-informed approaches. A drawback is that human evaluation methods are more effortful and costly [53]. Nevertheless, even the feedback from one expert yielded important insights into the goodness and satisfactoriness of textual explanations [116]. Two cases elicited more solid feedback for relevance-based explanations. In this context, practitioners verified that the contribution rates of medical events were consistent with cardiologists' diagnosing habits [106] and that expert-based ECG patterns were favored over purely data-driven ones [19].

C. Findings Regarding the Fit Between Embodied Knowledge and User Knowledge (RQ3)

This section addresses how knowledge embodied in explainable ML-based IS benefits users' understanding. Therefore, we match the prior knowledge in explanations to the explainability needs on a knowledge level, according to the type and context of knowledge [27]. Since often neither users nor the types and contexts of knowledge were specified, we assessed two guiding criteria when reviewing the literature: First, which missing information does the explanation add to users? Second, what knowledge is expected from a user? We mapped these characteristics into the framework as given in Table IV, to examine the fit between explanations and users.

a) ML context: Formal knowledge (A1) refers to researching, developing, operating, and deploying an ML-based IS. A KG interpretation method [74] could distinguish the contribution rates of the data and the KG attention. This helped analyze the effects of hyperparameter changes on the KG contribution and final model performance. Since this reasoning is involved in analyzing model behavior, we expect that it yields a more profound ML understanding for users seeking an external, calculatory influence of the KG. By visualizing the importance of biological concepts per each hidden layer [134], one could semantically investigate the internals of the model structure. It was meant to be interpretable for physicians and biologists, but was not evaluated. Instrumental ML-related knowledge (A2), on the other hand, occurred when the utilized knowledge mainly served as a tool for successive qualitative improvements of the models and their explanations. Such instrumentalization was used for semantic item set explanations [12]. Although the decision sets reflected concepts that one would recognize when being familiar with the data domain, the interpretability aimed to detect wrongly learned relations. Following that, further adaptions of the data, model architecture, or hyperparameters can be initiated to improve model quality, corresponding to an instrumentalization of ML expertise on the part of the user. The same form of knowledge was also involved in the interpretation of text segments [108], guiding the model's attention toward additional tissue information for pathology-level interpretation [127], and utilizing a KG to group features [117]. As these approaches involve knowledge to find better models or explanations, it requires instrumental ML familiarity of users to associate further ML modifications with the explanations.

Personal knowledge was addressed on the ML level (A3) when interpreting the KG-based attention behavior. It was analyzed that embeddings align with the KG structure [101], [102] and the attention mechanism resulted in more medically meaningful and intuitive representations. Although the intuition of explanations depends on personal experiences [27], such global logic is mainly meant for ML developers. In the visual analytics approach of a hidden Markov model [99], personal interpretations of medical experts played an important role.

The tool was designed to disclose the model's hidden states and their clinical meaning, and it supported experts to define and drive analytical workflows. Although this approach allowed for interactively discovering how inferred states may fit clinical stages, the workshops showed that participants' computational background and experiences influenced the ad-hoc understanding of the model's layered, probabilistic nature and how fast insights could be acquired. As described, these approaches embodied formal, instrumental, and personal aspects on the ML level. However, it proved to be useful for users to be familiarized with the data domain. As the example of sequential diagnosis prediction [106] revealed, the model behavior could be easily associated with doctoral diagnostic habits, so we find these types of knowledge-informed explanations helpful for both data- and ML-familiarized users.

b) Data context: Regarding formal knowledge in the data context (B1), KGs can provide causal relationships in the data to analyze the reasons for system outcomes. This offered traceable forms of medical knowledge, for instance, patient disease trajectories based on probabilistic chains of symptoms and diseases [20], [95], [125] or pathological relations to a target disease based on an embedded KG of spinal structures [126]. In another case where users were provided with the KG (including the properties of drugs and their relationships) the underlying data interconnectedness of the system was indicated [93]. Thus, the explanations provided by these methods rely on causal probabilities and validated medical knowledge, fostering formal understanding. Instrumental knowledge (B2) is gained through demonstration and practice. Along this vein, experts in the data domain were involved in developing, interacting, and improving the explanations of ML-based IS. The users of these systems interacted with interpretable artifacts such as rule bases and Bayesian networks, for instance, to eliminate unrealistic relations and resolve dilemmas. Next to improved performance, this allowed them to "actively understand" by taking more informed decisions [118] and to rely on their medical expertise to discover hidden patterns in the data [116]. Thereby, self-service systems had a special orientation toward the backgrounds and experiences of their users. The tools were designed for healthcare professionals without requiring exhaustive data mining knowledge, emphasizing the interpretability for non-technical users [98], [121].

Some explanations of knowledge-informed ML-based IS were oriented toward personal knowledge (B3). Such approaches guided explanations to clinically meaningful points in ECG data that cardiologists prefer [19], [128] or presented lung nodule characteristics (e.g., texture or sphericity) [110] and lesion annotations [109] that radiologists report when interpreting images in diagnosis. These characteristics also enabled case-based reasoning by semantically-close cases. First, they demonstrated explanations that imitated how experts make decisions in their work routine so that practitioners could evaluate whether they comply with their prior knowledge. Second, they could correct feature importance scores according to their opinion [119]. As an advantage of this interaction, no ML understanding was required for users since they did not need to update model parameters directly. Although such expert-mimicking explanations lacked human-subject evaluation, they seem intuitive and useful for decision-makers with expertise in the data domain.

c) Application environment: A less prominent way to design explanations was to take the application context into account. An instrumental aspect (C2) was given by contextual information in a Bayesian network [94]. The system was designed to consider a broad range of diseases and include the tests that a patient may undergo as well as laboratories where to perform them. The interpretable system should mitigate a lack of knowledge of diseases and symptoms that consequently get unnoticed in practice. Another instrumental aspect was to take interoperability with other medical IS into account [104], which demonstrated that other data sources were supported through the use of standardized medical terminology for medical concepts. Moreover, we observed that personal issues (C3) of the societal environment were addressed in rule-based explanators [105]. Through this method, a domain expert could investigate potential misdiagnosis in personally interested patient subgroups. Accordingly, the user could audit the model whether to trust it or whether fairness issues might be present in the corresponding healthcare facility. Although debatable whether the explainability of these systems did increase, it has demonstrated that knowledge-informed approaches could mitigate knowledge issues in the application environment.

VI. Discussion

A. Summary of Knowledge-Informed ML

Informed ML-based IS covered a wide range of knowledge sources, ranging from general disease information (e.g., disease and symptom relations) to in-depth knowledge about special conditions (e.g., characteristics of lung nodules). ICD-related information was easily retrievable since annotations of diagnoses and procedures were commonly present in the data as standardized ICD codes per patient. The assumption behind this source is that the codes are good surrogates for the patient's actual health condition; albeit subject to numerous potential sources of error in coding actual patient conditions [105]. To gather knowledge from humans, user interactions carried out annotation tasks, KG construction, and probability definition [94], [111]. The interactions provided useful sources for ML-based IS, with both ML engineering-related (e.g., [98], [116], [122] and rather health-related aspects (e.g., [19], [119]).

External KG sources were often attained for general-purpose information, transformed into graph-based embeddings [74], [101], [102], [106], [108], [117], and fused into neural networks via an attention mechanism [117], [127]. This helped models find useful information on interrelated diagnosis codes [102], words [108], or concept groups [117]. Often, advantages regarding reduced data demand for infrequent codes [101] or improved performance [74] were emphasized. However, the way the integration affects a medical user's interpretability was less prominent. Nevertheless, one could derive the contribution rates of individual embedded diagnosis events when considering a patient's visit history. These, in turn, could be compared to the contribution of their ancestors (e.g., [101]), grouped [117], or

TABLE V
KNOWLEDGE-LEVEL XAI BENEFITS OF KNOWLEDGE-INFORMED MACHINE LEARNING-BASED INFORMATION SYSTEMS

Context	Formal knowledge	Instrumental knowledge	Personal knowledge
ML	KG-attribution to model internals	Knowledge encoded in model components	Intuition about KG influence
Data	Causal and medical KG relationships	Medical expertise-informed or interactive system outcomes	Imitation of medical decision-making, medical expert-driven decision aids
Env.		Medical application context	Awareness of social issues in healthcare

split up into data and knowledge shares [74]. Fewer approaches utilized path-based ML techniques, e.g., for disease progression, or humans in the loop for the development of explanations. With KGs offering semantic distances, data-driven models and explanations could benefit from medical guidance through semantic loss functions [19], [110]. It illustrates that ML-based IS could exert the knowledge directly, as in joint learning, or indirectly, as in case-based explanations.

B. Knowledge-Level Contributions to User-Centric XAI

Next, we discuss to what extent the shift from data-driven to knowledge-informed ML was beneficial to a knowledge-level understanding of users. The generalized contributions to user-centric XAI are given in Table V.

In the *ML context*, we realized that former data-driven techniques (known to be developer- and ML technicians-oriented [24]) were improved with knowledge-driven computations. In the case of item set explanations, this resulted in a shift from frequent to semantically-related item sets. The attention mechanism was extended with KG-based techniques to group or improve low-dimensional representations based on hierarchical information. Lastly, contribution rates were adapted from data elements to KG entities. Even by embodying additional medical knowledge and qualitative improvements, such forms of explanations largely remained in ML terms. Although the users were often left unspecified, the tasks that were proposed involved technical interpretations or configurations, which would facilitate a cognitive fit for ML-familiarized users. On this basis, we can highlight ML-level benefits in terms of: formally attributing KG semantics to model internals; providing tools to encode prior knowledge in explanations, and, on a personal side, and intuitively highlighting the influence of knowledge on a model.

In the *data context*, most knowledge-informed ML systems presented explanations independent of models' internal properties. Background knowledge helped in providing KG chains, expert interactions, or finding expert decision aids in the data. Benefits resulting from this were: explanations as KG relationships; finding patterns in the data with medical knowledge; and aligning explanations with expert decision-making. In personal knowledge-related forms of explanations, expert-inspired features provided end-to-end matching, both to inform the systems with their knowledge and to receive explanations that align with their decision-making. As supported by CFT [36], explanations that match the knowledge necessary to complete a task, e.g., key characteristics of lung nodules in medical imaging, allow a user to deploy given cognitive processes. Hereby, explanations based on expert decision aids might benefit most from high levels of cognitive fit, but not necessarily increase the understanding for less data-familiar users. Formal information on medical implications, like disease trajectories for a given patient, showed a close connection to causal explanations which were recently brought forward [8]. In XAI, causability refers to the extent to which an AI explanation achieves a specified level of causal understanding of a domain expert in a specified context of use [135]. As causability influences the perceived utility of explanations, the diagnostic reasoning provided by knowledge-informed approaches seems to be beneficial for the cognitive fit of expert users in the data domain.

Fewer methods addressed the knowledge of the *environment* to mitigate social issues, which could address a lack of disease awareness [94] and identify over- and underdiagnoses [105].

As we indicated, explanations were developed based on disease relations, features, rules, and semantics. From a user's point of view, the knowledge integration in ML-based IS did not necessarily lead to *informed* explanations conveying domain knowledge about the decision. Congruent with prior research [87], [136], knowledge integration is often based on semantic similarity functions, with a lack of user-adaptive or interactive explanation approaches as a result. We observed similar effects in synthetic neighborhood generation and case-based explanations. In contrast, research on visual analytics produced novel aspects through interactive visualizations. In the ML context, this addressed data reorientation for classification model discovery and exploring the states of hidden Markov models. As the interfaces visualized models in various, albeit intuitive ways, we perceive ML-related expertise in the foreground of grasping these types of explanations. In the data context, interactive methods have specifically addressed feature-level interpretation or self-service use. We can note that self-service approaches most prominently balanced prerequisite ML familiarity with the extent of model-related controls. Although they may be understood by medical experts with limited ML experiences, e.g., with no intuition about classification algorithms, they affect the global understanding of the model per se, compared to patient-specific outcomes.

It was also mentioned that time and effort are required to learn how to interpret and use the interactions [99]. The intense time pressure on doctors, however, is known to prevent them from conducting a comprehensive search or prolonged interrogation when dealing with AI tools [137]. With gaining experience through interaction, trust in the system can be enhanced [138] and the role of explanations might also change [139], [140]. It remains to be seen whether interactive approaches will prove themselves in time-constrained daily clinical settings. Eventually, we realize two major opportunities regarding the shift from data-driven to knowledge-informed explanations: First, causal models (formal knowledge in the data domain), contextual information (instrumental knowledge in the environment), and expert's key features (personal knowledge in the data domain)

explicitly convey domain knowledge not only in ML models but also in explanations, such that they comply with medical knowledge. Second, we found novel types of explanations through KGs for formal understanding. On an ML level, the contribution rates of a medical KG could be revealed in addition to medical events. On the data level, KG concepts and path-based explanations provided probability chains of diseases and symptoms using causal medical knowledge.

C. Implications for Research and Practice

Using a separate source of knowledge mostly surpassed data-driven XAI by performance and interpretability at the same time. Explanations by attention weights, feature importance scores, and case-based reasoning have been influenced by medical knowledge. A 2-D knowledge framework [27] helped to match which types and contexts of knowledge the explanations embodied. When adopting explainable knowledge-informed ML-based IS, practitioners should holistically consider: what information (type and context) the explanation shall transmit; and what background and experiences the target users have to understand it. As indicated, the design of explanations could separately address formal relationships, interactions, or personal intuition, each targeted to the contexts of ML, data, and environment.

Notably, explanations addressing environmental knowledge were rare. Ideas have been proposed to shape explanations more user-centric by communicating evidence, like facts or data, supporting the decision, and relating the evidence to decisions with contextual information [26]. Informing ML with the knowledge of the application context may have not yet reached its peak since aspects of the user's situation, task, or workflow may also provide valuable context for adapting explanations. For personal knowledge, this requires a system to pay respect to human expectations or experiences of the environment, an "under-explored opportunity for human-centered design" [27, p. 10]. Explanations that imitate and represent the personal decision-making behavior of experts can be intuitive and influence cognitive fit. As a limitation, this kind of knowledge may at the same time be most beneficial to expert users. It remains to be explored how it affects the acceptance of non-experts and users that are nonfamiliar with the corresponding context. Research should therefore address explanations that imitate not only medical decision-making but also the way medical doctors explain their decisions to patients.

Besides developing explainable, knowledge-informed ML, a critical point is to systematically evaluate the explanation methods among users. With user evaluation only present in 7 out of 39 papers, this lack of rigor is a critical deficiency in reviewed approaches and also general XAI research [141], [142]. Computational metrics have mostly been used as a proxy for the quality of results. Even if user feedback was gathered, the used evaluation instruments were rarely transparent in terms of their design, scales, and collected answers. Future studies must bridge the critical gap between XAI development and rigorous user studies. In this review, the target users of discussed explanations center on medical doctors and developers. For diagnostic predictions, there was no system targeting patients, who (as decision subjects) might particularly be interested in local explanations. One reason could be that patients should be able to understand the risks and benefits of treatments rather than the AI procedures [42] and that they are more interested in verifying "why" but not "how" a certain decision was made [61]. Another reason can be seen that current XAI methods are—simply put—not useful for patients, who typically possess little ML or data knowledge. As stated by Ghassemi et al. [14], local explanations are unreliable for justifying individual decisions. This indicates that user-centric considerations have not been able to stimulate the development of explanations from the beginning. However, frameworks to distinguish users, "why" they need explanations, and "what" to include are being developed. As also different domain experts need different kinds of explanations [26], the user's background is important to consider, which to date has not been the case. A knowledge-level framework helped to determine the type of knowledge integrated and embodied in explanations. This way, we should be better able to develop specialized explanations and help users understand them as appropriate to their situation.

VII. Conclusion

As an important step toward user-centricity, this conceptual study shed light on the relation between knowledge-informed ML, explainability, and user aspects from a knowledge level. Although the functionality of ML-based IS has been improved with knowledge-driven computations, the knowledge often remained hidden in the resulting explanations. Nonetheless, we explicated the major benefits of knowledge-informed ML-based IS which allow for formal diagnostic explanations through knowledge chains, utilize contextual information of the application environment, and reflect expert decision criteria. Thereby, we have been able to distinguish the types and contexts in which prior knowledge contributed to the understanding of users, in comparison to data-driven XAI.

Future research is encouraged to develop explanations under consideration of the user's background and experiences, with personal knowledge as well as the application environment as promising open characteristics to explore. Although knowledge-informed ML-based IS bring significant benefits, only 7 out of 39 papers conducted human evaluations. Hence, the effectiveness and utility for users in real-world settings still need to be investigated, to allow more profound conclusions about their explainability. Nevertheless, we deem the knowledge-informed ML paradigm as an enabler for tackling the challenges of extant user-centric XAI. If we can do so, AI applications likely inspire more trust in clinical settings.

References

[1] K.-H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," *Nature Biomed. Eng.*, vol. 2, no. 10, pp. 719–731, 2018.

[2] C. M. Cutillo, K. R. Sharma, L. Foschini, S. Kundu, M. Mackintosh, and K. D. Mandl, "Machine intelligence in healthcare-perspectives on trustworthiness, explainability, usability, and transparency," *NPJ Digit. Med.*, vol. 3, pp. 1–5, 2020.

[3] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine," *Nature Med.*, vol. 28, no. 1, pp. 31–38, 2022.

[4] O. Biran and C. Cotton, "Explanation and justification in machine learning: A survey," in *Proc. IJCAI Workshop Explainable Artif. Intell.*, 2017, pp. 8–13.

[5] L. Alam and S. Mueller, "Examining the effect of explanation on satisfaction and trust in AI diagnostic systems," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, pp. 1–15, 2021.

[6] E. Zihni et al., "Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome," *PLoS One*, vol. 15, no. 4, 2020, Art. no. e0231166.

[7] B. Heinrichs and S. B. Eickhoff, "Your evidence? Machine learning algorithms for medical diagnosis and prediction," *Hum. Brain Mapping*, vol. 41, no. 6, pp. 1435–1444, 2020.

[8] T. Ploug and S. Holm, "The four dimensions of contestable AI diagnostics - A patient-centric approach to explainable AI," *Artif. Intell. Med.*, vol. 107, 2020, Art. no. 101901.

[9] K. Bauer, M. von Zahn, and O. Hinz, "Expl(Ai)Ned: The impact of explainable artificial intelligence on cognitive processes," *SSRN J.*, 2021.

[10] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[11] A. Barredo Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020.

[12] M. Moradi and M. Samwald, "Explaining black-box models for biomedical text classification," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 8, pp. 3112–3120, Aug. 2021, doi: 10.1109/JBHI.2021.3056748.

[13] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.

[14] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *Lancet Digit. Health*, vol. 3, no. 11, pp. e745–e750, 2021.

[15] W. Samek and K.-R. Müller, "Towards explainable artificial intelligence," in *Lecture Notes in Computer Science, Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, Eds., Berlin, Germany: Springer, 2019, pp. 5–22.

[16] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining explanations in AI," in *Proc. Conf. Fairness, Accountability, Transparency*, 2019, pp. 279–288.

[17] M. Jacobs et al., "Designing AI for trust and collaboration in time-constrained medical decisions: A sociotechnical lens," in *Proc. Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–14.

[18] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: Contextualizing explainable machine learning for clinical end use," in *Proc. 4th Mach. Learn. Healthcare Conf.*, 2019, pp. 359–380. [Online]. Available: https://proceedings.mlr.press/v106/tonekaboni19a.html

[19] X. Li et al., "Domain knowledge guided deep atrial fibrillation classification and its visual interpretation," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, 2019, pp. 129–138.

[20] Z. Sun, W. Dong, J. Shi, and Z. Huang, "Interpretable disease prediction based on reinforcement path reasoning over knowledge graphs," Oct. 2020. [Online]. Available: http://arxiv.org/pdf/2010.08300v1

[21] G. Ras, N. Xie, M. van Gerven, and D. Doran, "Explainable deep learning: A field guide for the uninitiated," *J. Artif. Intell. Res.*, vol. 73, pp. 329–397, 2022.

[22] A. Papenmeier, D. Kern, G. Englebienne, and C. Seifert, "It's complicated: The relationship between user trust, model accuracy and explanations in AI," *ACM Trans. Comput.-Hum. Interact.*, vol. 29, no. 4, pp. 1–33, 2022,.

[23] J. M. Alonso, C. Castiello, and C. Mencar, "A bibliometric analysis of the explainable artificial intelligence research field," in *Communications in Computer and Information Science, Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations*, J. Medina et al., Eds., Berlin, Germany: Springer, 2018, pp. 3–15.

[24] K. Bauer, O. Hinz, W. van der Aalst, and C. Weinhardt, "Expl(AI)n it to me – explainable AI and information systems research," *Bus. Inf. Syst. Eng.*, vol. 63, no. 2, pp. 79–82, 2021.

[25] C. Meske, E. Bunde, J. Schneider, and M. Gersch, "Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities," *Inf. Syst. Manage.*, vol. 39, no. 1, pp. 53–63, 2020, doi: 10.1080/10580530.2020.1849465.

[26] M. Ribera and A. Lapedriza, "Can we do better explanations? A proposal of user-centered explainable AI," in *Proc. Joint ACM IUI Workshops*, 2019.

[27] H. Suresh, S. R. Gomez, K. K. Nam, and A. Satyanarayan, "Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–16.

[28] U. Bhatt et al., "Explainable machine learning in deployment," in *Proc. Conf. Fairness, Accountability, Transparency*, 2020, pp. 648–657.

[29] S. Khemlani and P. N. Johnson-Laird, "Why machines don't (yet) reason like people," *Künstliche Intelligenz*, vol. 33, no. 3, pp. 219–228, 2019.

[30] D. Gunning and D. Aha, "DARPA's explainable artificial intelligence (XAI) program," *AI Mag.*, vol. 40, no. 2, pp. 44–58, 2019.

[31] T. A. Schoonderwoerd, W. Jorritsma, M. A. Neerincx, and K. van den Bosch, "Human-centered XAI: Developing design patterns for explanations of clinical decision support systems," *Int. J. Hum.-Comput. Stud.*, vol. 154, 2021, Art. no. 102684.

[32] S. Chari, D. M. Gruen, O. Seneviratne, and D. L. McGuinness, "Directions for explainable knowledge-enabled systems," in *Knowledge Graphs for Explainable Artificial Intelligence: Foundations, Applications and Challenges*, Amsterdam, The Netherlands: IOS Press, 2020, pp. 245–261.

[33] A. Bennetot, J.-L. Laurent, R. Chatila, and N. Díaz-Rodríguez, "Towards explainable neural-symbolic visual reasoning," Sep. 2019. [Online]. Available: http://arxiv.org/pdf/1909.09065v2

[34] A. D. Garcez, M. Gori, L. C. Lamb, L. Serafini, M. Spranger, and S. N. Tran, "Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning," May 2019. [Online]. Available: http://arxiv.org/pdf/1905.06088v1

[35] M. Gaur, K. Faldu, and A. Sheth, "Semantics of the black-box: Can knowledge graphs help make deep learning systems more interpretable and explainable?," *IEEE Internet Comput.*, vol. 25, no. 1, pp. 51–59, Jan./Feb. 2021.

[36] I. Vessey, "Cognitive fit: A theory-based analysis of the graphs versus tables literature," *Decis. Sci.*, vol. 22, no. 2, pp. 219–240, 1991.

[37] J. S. Giboney, S. A. Brown, P. B. Lowry, and J. F. Nunamaker, "User acceptance of knowledge-based system recommendations: Explanations, arguments, and fit," *Decis. Support Syst.*, vol. 72, pp. 1–10, 2015.

[38] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 44, pp. 22071–22080.

[39] V. Bellotti and K. Edwards, "Intelligibility and accountability: Human considerations in context-aware systems," *Hum.–Comput. Interaction*, vol. 16, no. 2-4, pp. 193–212, 2001.

[40] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021.

[41] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Analytics*, 2018, pp. 80–89.

[42] J. J. Wadden, "Defining the undefinable: The black box problem in healthcare artificial intelligence," *J. Med. Ethics*, pp. 764–768, 2021.

[43] J. Howick, P. Glasziou, and J. K. Aronson, "Evidence-based mechanistic reasoning," *J. Roy. Soc. Med.*, vol. 103, no. 11, pp. 433–441, 2010.

[44] L. G. McCoy, C. T. A. Brenna, S. S. Chen, K. Vold, and S. Das, "Believing in black boxes: Machine learning for healthcare does not need explainability to be evidence-based," *J. Clin. Epidemiol.*, vol. 142, pp. 252–257, 2022.

[45] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, 2019.

[46] M. Lemoine, "Explanation in medicine," in *The Routledge Companion to Philosophy of Medicine*, 1st ed. M. Solomon, J. R. Simon, and H. Kincaid, Eds., Evanston, IL, USA: Routledge, 2016, pp. 310–323.

[47] L. Arbelaez Ossa, G. Starke, G. Lorenzini, J. E. Vogt, D. M. Shaw, and B. S. Elger, "Re-focusing explainability in medicine," *Digit. Health*, vol. 8, 2022, Art. no. 20552076221074488.

[48] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdiscip. Rev. Data Mining Knowl. Discov.*, vol. 9, no. 4, 2019, Art. no. e1312.

[49] P. Korica, N. E. Gayar, and W. Pang, "Explainable artificial intelligence in healthcare: Opportunities, gaps and challenges and a novel way to look at the problem space," in *Lecture Notes in Computer Science, Intelligent Data Engineering and Automated Learning*, H. Yin, et al., Eds., Berlin, Germany: Springer, 2021, pp. 333–342.

[50] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[51] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies," *J. Biomed. Inform.*, vol. 113, 2021, Art. no. 103655.

[52] A. Bibal and B. Frénay, "Interpretability of machine learning models and representations: An introduction," in *Proc. Eur. Symp. Artif. Neural Netw., Comput. Intell. Mach. Learn.*, 2016, pp. 77–82.

[53] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," Feb. 2017. [Online]. Available: http://arxiv.org/pdf/1702.08608v2

[54] M. Förster, M. Klier, K. Kluge, and I. Sigler, "Fostering human agency: A process for the design of user-centric XAI systems," in *Proc. 41st Int. Conf. Inf. Syst.*, 2020.

[55] M. Langer et al., "What do we want from explainable artificial intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research," *Artif. Intell.*, vol. 296, 2021, Art. no. 103473.

[56] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "XAI-explainable artificial intelligence," *Sci. Robot.*, vol. 4, no. 37, 2019.

[57] C. Zednik, "Solving the black box problem: A normative framework for explainable artificial intelligence," *Philosophy Technol.*, vol. 34, no. 2, pp. 265–288, 2021, doi: 10.1007/s13347-019-00382-7.

[58] J. Gerlings, M. S. Jensen, and A. Shollo, "Explainable AI, but explainable to whom?," Jun. 2021. [Online]. Available: http://arxiv.org/pdf/2106.05568v1

[59] J. J. Ferreira and M. S. Monteiro, "What are people doing about XAI user experience? A survey on AI explainability research and practice," in *Lecture Notes in Computer Science, Design, User Experience, and Usability. Design for Contemporary Interactive Environments*, A. Marcus and E. Rosenzweig Eds., Berlin, Germany: Springer, 2020, pp. 56–73.

[60] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau, "Visual analytics in deep learning: An interrogative survey for the next frontiers," *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 8, pp. 2674–2693, Aug. 2019, doi: 10.1109/TVCG.2018.2843369.

[61] R. Tomsett, D. Braines, D. Harborne, A. Preece, and S. Chakraborty, "Interpretable to whom? A role-based model for analyzing interpretable machine learning systems," Jun. 2018. [Online]. Available: http://arxiv.org/pdf/1806.07552v1

[62] H. Felzmann, E. F. Villaronga, C. Lutz, and A. Tamò-Larrieux, "Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns," *Big Data Soc.*, vol. 6, no. 1, 2019.

[63] L. von Rueden et al., "Informed machine learning – A taxonomy and survey of integrating prior knowledge into learning systems," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 614–633, Jan. 2023, doi: 10.1109/TKDE.2021.3079836.

[64] G. Marcus, "The next decade in AI: Four steps towards robust artificial intelligence," Feb. 2020. [Online]. Available: http://arxiv.org/pdf/2002.06177v3

[65] T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.

[66] S. Muggleton, *Inductive Logic Programming*. New York, NY, USA: Academic, 1992.

[67] S. Džeroski, "Relational data mining," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds., Berlin, Germany: Springer, 2010, pp. 887–911.

[68] K. Ziegler et al., "Injecting semantic background knowledge into neural networks using graph embeddings," in *Proc. IEEE 26th Int. Conf. Enabling Technol., Infrastructure Collaborative Enterprises*, 2017, pp. 200–205.

[69] K. Beckh et al., "Explainable machine learning with prior knowledge: An overview," May 2021. [Online]. Available: http://arxiv.org/pdf/2105.10172v1

[70] L. Ehrlinger and W. Wöß, "Towards a definition of knowledge graphs," *SEMANTiCS*, vol. 48, no. 1–4, 2016.

[71] H. Paulheim, "Knowledge graph refinement: A survey of approaches and evaluation methods," *Semantic Web*, vol. 8, no. 3, pp. 489–508, 2016, doi: 10.3233/SW-160218.

[72] A. Blumauer, "From taxonomies over ontologies to knowledge graphs," 2014. [Online]. Available: https://semantic-web.com/from-taxonomies-over-ontologies-to-knowledge-graphs/

[73] G. Mai, K. Janowicz, and B. Yan, "Combining text embedding and knowledge graph embedding techniques for academic search engines," in *Proc. Joint 4th Workshop Semantic Deep Learn.*, 2018, pp. 77–88. [Online]. Available: http://ceur-ws.org/Vol-2241/paper-08.pdf

[74] C. Yin, R. Zhao, B. Qian, X. Lv, and P. Zhang, "Domain knowledge guided deep learning with electronic health records," in *Proc. IEEE Int. Conf. Data Mining*, 2019, pp. 738–747.

[75] D. Dellermann, A. Calma, N. Lipusch, T. Weber, S. Weigel, and P. Ebel, "The future of human-AI collaboration: A taxonomy of design knowledge for hybrid intelligence systems," May 2021. [Online]. Available: http://arxiv.org/pdf/2105.03354v1

[76] B. Kovalerchuk, M. A. Ahmad, and A. Teredesai, "Survey of explainable machine learning with visual and granular methods beyond quasi-explanations," in *Studies in Computational Intelligence, Interpretable Artificial Intelligence: A Perspective of Granular Computing*, W. Pedrycz and S. M. Chen, Eds., Berlin, Germany: Springer, 2021, pp. 217–267.

[77] Q. Wang and R. S. Laramee, "EHR STAR: The state-of-the-art in interactive EHR visualization," *Comput. Graph. Forum*, vol. 41, no. 1, pp. 69–105, 2022.

[78] E. Rader, K. Cotter, and J. Cho, "Explanations as mechanisms for supporting algorithmic transparency," in *Proc. Conf. Hum. Factors Comput. Syst.*, 2018, pp. 1–13.

[79] D. Bačić and R. M. Henry, "Task-representation fit's impact on cognitive effort in the context of decision timeliness and accuracy: A cognitive fit perspective," *AIS Trans. Hum.-Comput. Interaction*, vol. 10, pp. 164–187, 2018.

[80] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *J. Artif. Intell. Res.*, vol. 70, pp. 245–317, 2021.

[81] J. Townsend, T. Chaton, and J. M. Monteiro, "Extracting relational explanations from deep neural networks: A survey from a neural-symbolic perspective," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3456–3470, Sep. 2020,.

[82] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (XAI): A survey," Jun. 2020. [Online]. Available: http://arxiv.org/pdf/2006.11371v2

[83] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable artificial intelligence: An analytical review," *Wiley Interdiscip. Rev. Data Mining Knowl. Discov.*, vol. 11, no. 5, 2021, Art. no. e1424.

[84] P. Hitzler, F. Bianchi, M. Ebrahimi, and M. K. Sarker, "Neural-symbolic integration and the semantic web," *Semantic Web*, vol. 11, no. 1, pp. 3–11, 2020.

[85] S. Marton, S. Lüdtke, and C. Bartelt, "Explanations for neural networks by neural networks," *Appl. Sci.*, vol. 12, no. 3, 2022, Art. no. 980.

[86] J.-X. Mi, A.-D. Li, and L.-F. Zhou, "Review study of interpretation methods for future interpretable machine learning," *IEEE Access*, vol. 8, pp. 191969–191985, 2020, doi: 10.1109/ACCESS.2020.3032756.

[87] A. Seeliger, M. Pfaff, and H. Krcmar, "Semantic web technologies for explainable machine learning models: A literature review," in *Proc. 8th Int. Semantic Web Conf*, 2019, vol. 2465, pp. 1–16.

[88] R. Calegari, G. Ciatto, and A. Omicini, "On the integration of symbolic and sub-symbolic techniques for XAI: A survey," *Intelligenza Artificiale*, vol. 14, no. 1, pp. 7–32, 2020.

[89] X.-H. Li et al., "A survey of data-driven and knowledge-aware eXplainable AI," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 1, pp. 29–49, Jan. 2022, doi: 10.1109/TKDE.2020.2983930.

[90] J. Ooge, G. Stiglic, and K. Verbert, "Explaining artificial intelligence with visual analytics in healthcare," *Wiley Interdiscip. Rev. Data Mining Knowl. Discov.*, vol. 12, no. 1, 2022, Art. no. e1427.

[91] J. Webster and R. T. Watson, "Analyzing the past to prepare for the future: Writing a literature review," *MIS Quart.*, vol. 26, no. 2, pp. xiii–xxiii, 2002.

[92] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *Plos Med.*, vol. 6, no. 7, 2009, Art. no. e1000097.

[93] A. Fokoue, O. Hassanzadeh, M. Sadoghi, and P. Zhang, "Predicting drug-drug interactions through similarity-based link prediction over web data," in *Proc. 25th Int. Conf. Companion World Wide Web*, 2016, pp. 175–178.

[94] P. Agarwal, R. Verma, and A. Mallik, "Ontology based disease diagnosis system with probabilistic inference," in *Proc. 1st India Int. Conf. Inf. Process.*, 2016, pp. 1–5.

[95] H. Yu, "Health causal probability knowledge graph," in *Proc. 7th Int. Conf. Bioinf. Res. Appl.*, 2020, pp. 49–58.

[96] C. Pesquita, "Towards semantic integration for explainable artificial intelligence in the biomedical domain," in *Proc. 14th Int. Joint Conf. Biomed. Eng. Syst. Technol.*, 2021, pp. 747–753.
[97] H. Yuan and W. Deng, "Doctor recommendation on healthcare consultation platforms: An integrated framework of knowledge graph and deep learning," *Internet Res.*, vol. 32, no. 2, pp. 454–476, 2022.
[98] M. Spruit and N. de Vries, "Self-service data science for adverse event prediction in electronic healthcare records," in *Springer Proceedings in Complexity, Research and Innovation Forum 2020*, A. Visvizi, M. D. Lytras, and N. R. Aljohani, Eds., Berlin, Germany: Springer, 2021, pp. 517–535.
[99] B. C. Kwon et al., "DPVis: Visual analytics with hidden Markov models for disease progression pathways," *IEEE Trans. Visual. Comput. Graph.*, vol. 27, no. 9, pp. 3685–3700, Sep. 2021, doi: 10.1109/TVCG.2020.2985689.
[100] S. Lima, L. Teran, and E. Portmann, "A proposal for an explainable fuzzy-based deep learning system for skin cancer prediction," in *Proc. 7th Int. Conf. eDemocracy eGovernment*, 2020, pp. 29–35.
[101] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, "GRAM: Graph-based attention model for healthcare representation learning," in *Proc. Int. Conf. Knowl. Discov. Data Mining*, 2017, pp. 787–795, doi: 10.1145/3097983.3098126.
[102] F. Ma, Q. You, H. Xiao, R. Chitta, J. Zhou, and J. Gao, "KAME," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage.*, 2018, pp. 743–752.
[103] K. Zhang, L. Cai, Y. Song, T. Liu, and Y. Zhao, "Combining external medical knowledge for improving obstetric intelligent diagnosis: Model development and validation," *JMIR Med. Inform.*, vol. 9, no. 5, 2021, Art. no. e25304.
[104] S. El-Sappagh, J. M. Alonso, F. Ali, A. Ali, J.-H. Jang, and K.-S. Kwak, "An ontology-based interpretable fuzzy decision support system for diabetes diagnosis," *IEEE Access*, vol. 6, pp. 37371–37394, 2018, doi: 10.1109/ACCESS.2018.2852004.
[105] C. Panigutti, A. Perotti, A. Panisson, P. Bajardi, and D. Pedreschi, "FairLens: Auditing black-box clinical decision support systems," *Inf. Process. Manage.*, vol. 58, no. 5, 2021, Art. no. 102657.
[106] X. Zhang, B. Qian, Y. Li, C. Yin, X. Wang, and Q. Zheng, "KnowRisk: An interpretable knowledge-guided model for disease risk prediction," in *Proc. IEEE Int. Conf. Data Mining*, 2019, pp. 1492–1497.
[107] C. Panigutti, A. Perotti, and D. Pedreschi, "Doctor XAI," in *Proc. Conf. Fairness, Accountability, Transparency*, 2020, pp. 629–639.
[108] F. Teng, W. Yang, L. Chen, L. Huang, and Q. Xu, "Explainable prediction of medical codes with knowledge graphs," *Front. Bioeng. Biotechnol.*, vol. 8, 2020, Art. no. 867.
[109] K. Yan, Y. Peng, V. Sandfort, M. Bagheri, Z. Lu, and R. M. Summers, "Holistic and comprehensive annotation of clinically significant findings on diverse CT images: Learning from radiology reports and label ontology," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8515–8524.
[110] W. Xu, K. Wang, J. Lin, Y. Lu, S. Huang, and X. Zhang, "Knowledge-guided and hyper-attention aware joint network for benign-malignant lung nodule classification," in *Proc. IEEE Int. Conf. Image Process.*, 2020, pp. 310–314.
[111] S. Lin, Q. Zhang, F. Chen, L. Luo, L. Chen, and W. Zhang, "Smooth Bayesian network model for the prediction of future high-cost patients with COPD," *Int. J. Med. Inform.*, vol. 126, pp. 147–155, 2019.
[112] G. Zaharchuk, "Fellow in a box: Combining AI and domain knowledge with Bayesian networks for differential diagnosis in neuroimaging," *Radiology*, vol. 295, no. 3, pp. 638–639, 2020.
[113] M. T. Duong, A. M. Rauschecker, and S. Mohan, "Diverse applications of artificial intelligence in neuroradiology," *Neuroimaging Clin. North Amer.*, vol. 30, no. 4, pp. 505–516, 2020.
[114] A. C. Constantinou, N. Fenton, and M. Neil, "Integrating expert knowledge with data in Bayesian networks: Preserving data-driven expectations when the expert variables remain unobserved," *Expert Syst. Appl.*, vol. 56, pp. 197–208, 2016.
[115] S. Enamandram, E. Sandhu, B. H. Do, J. J. Reicher, and C. F. Beaulieu, "Artificial intelligence and machine learning applications in musculoskeletal imaging," *Adv. Clin. Radiol.*, vol. 2, pp. 285–297, 2020.
[116] S. S. Samuel, N. N. B. Abdullah, and A. Raj, "Interpretation of SVM using data mining technique to extract syllogistic rules," in *Lecture Notes in Computer Science, Machine Learning and Knowledge Extraction*, A. Holzinger, P. Kieseberg, A. M. Tjoa, and E. Weippl, Eds., Berlin, Germany: Springer, 2020, pp. 249–266.
[117] L. J. Liu, V. Ortiz-Soriano, J. A. Neyra, and J. Chen, "KGDAL: Knowledge graph guided double attention LSTM for rolling mortality prediction for AKI-D patients," in *Proc. ACM Conf. Bioinf., Comput. Biol. Biomed.*, 2021, pp. 1–10, doi: 10.1145/3459930.3469513.
[118] N. Prentzas, A. Nicolaides, E. Kyriacou, A. Kakas, and C. Pattichis, "Integrating machine learning with symbolic reasoning to build an explainable AI model for stroke prediction," in *Proc. IEEE 19th Int. Conf. Bioinf. Bioeng.*, 2019, pp. 817–821.
[119] B. C. Kwon et al., "RetainVis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records," *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 1, pp. 299–309, Jan. 2019, doi: 10.1109/TVCG.2018.2865027.
[120] R. Li, C. Yin, S. Yang, B. Qian, and P. Zhang, "Marrying medical domain knowledge with deep learning on electronic health records: A deep visual analytics approach," *J. Med. Internet Res.*, vol. 22, no. 9, 2020, Art. no. e20645.
[121] S. N. Wagle and B. Kovalerchuk, "Self-service data classification using interactive visualization and interpretable machine learning," in *Studies in Computational Intelligence, Integrating Artificial Intelligence and Visualization for Visual Knowledge Discovery*, B. Kovalerchuk, K. Nazemi, R. Andonie, N. Datia, and E. Banissi, Eds., Berlin, Germany: Springer, 2022, pp. 101–139.
[122] S. N. Wagle and B. Kovalerchuk, "Interactive visual self-service data classification approach to democratize machine learning," in *Proc. 24th Int. Conf. Inf. Visualisation*, 2020, pp. 280–285.
[123] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: An overview," in *Lecture Notes in Computer Science, Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, Eds., Berlin, Germany: Springer, 2019, pp. 193–209.
[124] O. Lahav, N. Mastronarde, and M. van der Schaar, "What is interpretable? Using machine learning to design interpretable decision-support systems," Nov. 2018. [Online]. Available: http://arxiv.org/pdf/1811.10799v2
[125] H. Chereda et al., "Explaining decisions of graph convolutional neural networks: Patient-specific molecular subnetworks responsible for metastasis prediction in breast cancer," *Genome Med.*, vol. 13, no. 1, pp. 1–16, 2021.
[126] Z. Han, B. Wei, X. Xi, B. Chen, Y. Yin, and S. Li, "Unifying neural learning and symbolic reasoning for spinal medical report generation," *Med. Image Anal.*, vol. 67, 2021, Art. no. 101872.
[127] C. Yan, J. Xu, J. Xie, C. Cai, and H. Lu, "Prior-aware CNN with multi-task learning for colon images analysis," in *Proc. IEEE 17th Int. Symp. Biomed. Imag.*, 2020, pp. 254–257.
[128] Y. Zhou et al., "Addressing noise and skewness in interpretable health-condition assessment by learning model confidence," *Sensors*, vol. 20, no. 24, 2020.
[129] S. Wang, Y. Yin, D. Wang, Y. Wang, and Y. Jin, "Interpretability-based multimodal convolutional neural networks for skin lesion diagnosis," *IEEE Trans. Cybern.*, vol. 52, no. 12, pp. 12623–12637, Dec. 2022.
[130] M.-Y. Kim et al., "A multi-component framework for the analysis and design of explainable artificial intelligence," *Mach. Learn. Knowl. Extraction*, vol. 3, no. 4, pp. 900–921, 2021.
[131] Y.-H. Sheu, "Illuminating the black box: Interpreting deep neural network models for psychiatric research," *Front. Psychiatry*, vol. 11, 2020, Art. no. 551299.
[132] S. Jain and B. C. Wallace, "Attention is not explanation," Feb. 2019. [Online]. Available: http://arxiv.org/pdf/1902.10186v3
[133] S. Sager et al., "Expert-enhanced machine learning for cardiac arrhythmia classification," *PLoS One*, vol. 16, no. 12, 2021, Art. no. e0261571, doi: 10.1371/journal.pone.0261571.
[134] V. Bourgeais, F. Zehraoui, M. Ben Hamdoune, and B. Hanczar, "Deep GONet: Self-explainable deep neural network based on gene ontology for phenotype prediction from gene expression data," *BMC Bioinf.*, vol. 22, 2021, Art. no. 455.
[135] A. Holzinger, A. Carrington, and H. Müller, "Measuring the quality of explanations: The system causability scale (SCS): Comparing human and machine explanations," *Künstliche Intelligenz*, vol. 34, no. 2, pp. 193–198, 2020.
[136] V. Balayan, P. Saleiro, C. Belém, L. Krippahl, and P. Bizarro, "Teaching the machine to explain itself using domain knowledge," Nov. 2020. [Online]. Available: http://arxiv.org/pdf/2012.01932v1
[137] S. Lebovitz, "Diagnostic doubt and artificial intelligence: An inductive field study of radiology work," in *Proc. Int. Conf. Inf. Syst.*, 2019, pp. 1–17.

[138] D. Branley-Bell, R. Whitworth, and L. Coventry, "User trust and understanding of explainable AI: Exploring algorithm visualisations and user biases," in *Lecture Notes in Computer Science, Human-Computer Interaction. Human Values and Quality of Life*, M. Kurosu, Ed., Berlin, Germany: Springer, 2020, pp. 382–399.
[139] A. Bussone, S. Stumpf, and D. O'Sullivan, "The role of explanations on trust and reliance in clinical decision support systems," in *Proc. Int. Conf. Healthcare Inform.*, 2015, pp. 160–169.
[140] F. C. Kitamura and O. Marques, "Trustworthiness of artificial intelligence models in radiology and the role of explainability," *J. Amer. College Radiol.*, vol. 18, no. 8, pp. 1160–1162, 2021.
[141] J. van der Waa, E. Nieuwburg, A. Cremers, and M. Neerincx, "Evaluating XAI: A comparison of rule-based and example-based explanations," *Artif. Intell.*, vol. 291, 2021, Art. no. 103404.
[142] W. Xu, M. J. Dainoff, L. Ge, and Z. Gao, "Transitioning to human interaction with AI systems: New challenges and opportunities for HCI professionals to enable human-centered AI," in *Proc. Int. J. Hum.–Comput. Interaction*, 2022, pp. 1–25.
[143] V. Bourgeais, F. Zehraoui, and B. Hanczar, "GraphGONet: A self-explaining neural network encapsulating the gene ontology graph for phenotype prediction on gene expression," *Bioinformatics*, vol. 38, no. 9, pp. 2504–2511, 2022.
[144] J. D. Rudie et al., "Subspecialty-level deep gray matter differential diagnoses with deep learning and Bayesian networks on clinical brain MRI: A pilot study," *Radiol. Artif. Intell.*, vol. 2, no. 5, 2020, Art. no. e190146.
[145] S. N. Wagle and B. Kovalerchuk, "Self-service data classification using interactive visualization and interpretable machine learning," 2021. [Online]. Available: http://arxiv.org/pdf/2107.04971v1

Luis Oberste received the M.Sc. degree in business informatics in 2019 from the University of Mannheim, Mannheim, Germany, where he is currently working toward the Ph.D. degree in information systems.

His research interest includes healthcare information systems and explainable AI in healthcare.

Armin Heinzl received the masters' degree in business administration from the University of Frankfurt, Frankfurt, Germany, and the habilitation degree in information systems from the WHU - Otto Beisheim School of Management, Vallendar, Germany.

He is a Professor and Chair Person in business informatics with the University of Mannheim, Mannheim, Germany. His research interests include conversational agents, mobile health apps for behavior change, AI augmented decision support, and computational creative systems.