Designing User-Centric Explanations for Medical Imaging with Informed Machine Learning

Luis Oberste[1(✉)], Florian Rüffer[1], Okan Aydingül[1], Johann Rink[2], and Armin Heinzl[1]

[1] University of Mannheim, Mannheim, Germany
{loberste,frueffer,oaydingu,aheinzl}@mail.uni-mannheim.de
[2] University Medical Centre Mannheim, Mannheim, Germany
johann.rink@medma.uni-heidelberg.de

Abstract. A flawed algorithm released in clinical practice can cause unintended harm to patient health. Risks, regulation, responsibility, and ethics shape the demand of clinical users to understand and rely on the outputs made by artificial intelligence. Explainable artificial intelligence (XAI) offers methods to render a model's behavior understandable from different perspectives. Extant XAI, however, is mainly data-driven and designed to meet developers' demands to correct models rather than clinical users' expectations to reflect clinically relevant information. To this end, informed machine learning (IML) utilizes prior knowledge jointly with data to generate predictions, a promising paradigm to enrich XAI with medical knowledge. To explore how IML can be used to generate explanations that are congruent to clinical users' demands and useful to medical decision-making, we conduct Action Design Research (ADR) in collaboration with a team of radiologists. We propose an IML-based XAI system for clinically relevant explanations of diagnostic imaging predictions. With the help of ADR, we reduce the gap between implementation and user evaluation and demonstrate the effectiveness of the system in a real-world application with clinicians. While we develop design principles of using IML for user-centric XAI in diagnostic imaging, the study demonstrates that an IML-based design adequately reflects clinicians' conceptions. In this way, IML inspires greater understandability and trustworthiness of AI-enabled diagnostic imaging.

Keywords: Explainable Artificial Intelligence · Informed Machine Learning · Action Design Research · Medical Image Analysis · User-Centric Design

1 Introduction and Problem Statement

Advances in artificial intelligence (AI) have led to human-level performance in time-consuming routine tasks such as medical image analyses, which is highly beneficial for more efficient and qualitative medical diagnostics [1]. However, the lack of interpretability, known as the 'black-box' challenge, faced by medical AI raised fears of having

A. Gerber and R. Baskerville (Eds.): DESRIST 2023, LNCS 13873, pp. 470–484, 2023.
https://doi.org/10.1007/978-3-031-32808-4_29

unintended as well as unethical consequences. Thus, it has become a serious barrier to its widespread adoption in clinical routine [2]. A flawed algorithm in this high-stake domain can cause large-scale harm to patients [3]. Meanwhile, frameworks have been proposed to develop and apply AI in an ethical, legal, and morally responsible manner [4]. These ascribe a fundamental role to explainability for a safe clinical application of AI [5]. Since clinicians must be able to justify their decisions, supporting the output of a model with explanations that they can fully understand and trust is of vital importance for clinical application [6, 7]. Explainable artificial intelligence (XAI) involves several techniques to render models' behaviors and outputs understandable from different perspectives. Despite the surge of developed XAI methods and their potential to realize responsible AI, their utility in healthcare is still questioned [8]. One critic is that current XAI designs purely rely their results on the features, examples, or patterns detected in the input data, e.g., through attribution-based methods to generate heat maps. These do not reveal the reason why exactly a model found an area useful for a diagnosis decision [8]. Such data-driven explanations have been claimed by clinicians as inadequate to identify appropriate interventions [9] and mismatching human conception [10]. To make matters worse, current XAI initiatives do not pay enough attention to whom the explanations are targeted [11], and mostly only meet developers' demands, e.g., to debug and enhance models [12].

Clinicians, in contrast, expect explainable systems to leverage validated and clinically relevant information that reliably supports their medical decision-making [9]. Meanwhile, it has been concluded that integrating prior knowledge into machine learning systems is essential for improved explanations of their functioning [13]. Techniques from informed machine learning (IML) have been developed to learn from both data and a separate source of knowledge [14]. While it lacks understanding of their precise effect on explainability [15], they bear the potential to integrate medical information for more informed explanations. Thus, this paper aims to answer the following research questions: *How can IML be used to provide more clinician-centric explanations for medical image analysis? How can effective explanatory information be integrated to help clinicians make more informed diagnosis decisions?*

There is a lack of studies involving respective users in the design process and evaluations within real applications [16]. However, evaluating medical XAI's effectiveness requires human-in-the-loop experiments [7]. Therefore, we will follow an Action Design Research-based approach to facilitate our interdisciplinary collaboration with radiologists [17, 18] and aim to develop a novel IML-based system for explainable diagnostic image analysis. On this basis, we aim to contribute to the XAI challenge by designing user-centric explanations that are congruent to clinical users' expectations and evaluating the effectiveness in a true-to-life setting in a team of real clinicians.

2 Theoretical Background and Related Work

2.1 AI-Enabled Computer-Aided Diagnosis Systems

Based on predictive statistical models, AI-enabled tools can support clinicians in diagnosis decisions with a second objective opinion [19–21]. So-called computer-aided *diagnostics* (CADx) flourished in disciplines that rely on the interpretation of images,

including radiology [22]. In contrast to conventional CADx designs, deep architectures of contemporary AI have been able to overcome the highly specialized and effortful feature extraction by learning features from input images by themselves [21, 23]. During technological progress, human-AI collaboration is leading to various new forms of interaction, with research and clinical implementations in their infancy [2, 24, 25]. Successfully translating AI-enabled CADx into practice requires purposeful designs of clinician-AI interactions that build conviction and trust [20, 26]. For fostering clinical trust, transparency has emerged as a crucial element, as even accurate diagnoses that are not understood are likely to be ignored by clinicians [2, 27].

2.2 Explainable Artificial Intelligence in Healthcare

Explainability (or interpretability) is associated with the extent to which a user of an AI system is able to understand the reasoning behind an output generated by it [20, 28]. Such techniques attempt to explain a model's prediction, uncover its inner workings, or represent it using coherent expressions [29]. Typically, one can develop XAI either by designing an intrinsically interpretable model or by complementing a black-box model with a post-hoc explanation method, offering local or global explanations [20, 30]. Among the most used explainability techniques in various medical imaging tasks and modalities are heat maps [31]. These highlight how much each region of an image influences a model's disease prediction [22]. Like most XAI techniques, these are *data-driven* [32] since the explanations are generated purely from the data underlying the ML pipeline. These do not offer any information on how salient regions contribute to the result, and the same regions can be highlighted for contradicting predictions [20]. Any highlighted region leaves it up to the clinician to find an explanation, e.g., whether an airspace opacity, the shape of an organ, or a particular pixel has triggered the decision [33]. Moreover, *post-hoc* XAI methods (like heat maps) are based on approximations which cannot ensure complete accountability [20]. Complementing data-driven models with validated medical knowledge is thus a key to enhanced interpretability of medical imaging predictions [34, 35].

2.3 Informed Machine Learning

Integrating prior knowledge into ML can increase performance, robustness, and lower demands for large amounts of data [36]. While data-driven models might not conform to the knowledge of the domain and context, IML uses a separate source of information and integrates it into an ML pipeline [15]. It has been recently revisited to provide users with more meaningful and useful explanations [13, 37]. According to recent reviews [14, 15, 37], the spectrum of IML approaches ranges from the integration of scientific knowledge via common sense to expert knowledge in the respective field of application, based on various representations such as logical rules, equations, or knowledge graphs. A common way to integrate prior knowledge is to add a regularization term to a model's loss function which guides its learning toward desired outcomes. In healthcare, IML techniques are particularly promising to utilize the medical knowledge already available, such as disease-symptom relations or expert decision criteria, ultimately improving human understanding of ML models [30].

2.4 User-Centric Design

Clinical users have different expectations of the appropriateness of information than systems designers [38]. Their sustained use of predictions depends, among others, on effectiveness rather than the amount of information, and whether these anticipate a clinically significant change in a patient's condition. This implies to only report relevant results and avoid repeated or uncertain prompts, e.g., by using thresholds. However, clinical users are often not included in the development process [27], leading to unexpected dynamics and ambiguity when having the tools implemented into the high-judgment work of radiologists [39]. Information systems (IS) research has recognized that user-centric design is one of the central challenges for XAI to meet the needs of users who are typically domain but not technical experts [12]. Notable studies revealed that XAI models should reflect an analytic process like that of evidence-based medical decision-making [40], allow users to connect outputs to existing clinical processes [9, 41], and allow novices to become experts [42]. Hence, developing and evaluating user-centric XAI requires application-grounded experiments to justify whether goals are achieved in real-world settings [43].

3 Research Method and Process

3.1 Action Design Research-Based Approach

Design Science Research is concerned with designing and evaluating innovative IT artifacts for practically relevant challenges and with deriving prescriptive design knowledge [44, 45]. In particular, Action Design Research (ADR) places special emphasis on the interaction with stakeholders during artifact development [17]. That is, the interaction between researchers and stakeholders (here: radiologists) is reflected in the artifacts' redesign and the evaluation takes place in continuous interaction throughout each design cycle [18]. While XAI systems have been driven primarily by technological capabilities [39, 46], ADR helps us to intertwine XAI development and its integration into the highly loaded clinical routine, by closely collaborating with a clinic of radiology and nuclear medicine at a cooperating university hospital. The original ADR concept offers many opportunities for interpretation of how to conduct the 'build, intervene, and evaluate' engagement [18]. Critics argue that in different stages of the intervention, various artifact creations emerge gradually from which design knowledge should be captured at all levels. Hence, we conducted an elaborated ADR (eADR) process with all ADR activities in each intervention [18]: *problem formulation and planning, artifact creation, evaluation, reflection, and learning.*

The first activity aims to identify and conceptualize the research object [17]. Initial interviews concerned (a) the clinical background of radiology imaging, (b) problematic diseases and the help of XAI, (c) existing CADx support, as well as (d) appropriate modes and desired characteristics of explanations. We also conducted a refined literature search to identify solution possibilities and plan the design. Artifact creation is the essential building activity of various forms like models, tools, principles, features, or instantiations, depending on the current stage of the process [18].

The artifact was continuously evaluated and iteratively refined according to meeting expert expectations in practice [17, 47], through biweekly on-site presentations for ad-hoc feedback and semi-structured interviews for each ADR stage. All interviews were conducted with two resident radiologists (3 and 4 years of work experience), prior to half-hour job shadowing. Finally, the implementation stage was also evaluated in an informal expert group of radiotherapists, physicists, radiologists, and CADx software developers. Reflection and learning activities shift from a case-specific solution to broader medical XAI research: First, we constantly compared our artifact to the literature streams of IML and medical imaging to plan the next ADR cycles. To overcome the limited design knowledge accumulation observed in the IS community [44], we have aimed at complementary design knowledge. To this end, we collected meta-requirements (MRs) from prior research as theory-derived MRs are abstract and valid for more than one artifact [48]. We then derived generalizable learnings and formalized them as design principles (DPs), a set of prescriptive instructions for an artifact design that addresses MRs and follows established templates [48].

To avoid common inconsistencies in DPs, we derived 'action and materiality oriented' design principles that state *what* an artifact should provide and *how* it should be built to reach that [49]. Iterations are divided by eADR into *diagnosis, design, implementation*, and *evolution* stages. As the problem domain is largely identified and XAI models generally need to explain *how* and *why* the system generated a particular output [28], we chose an *objective-centered* entry point. To begin with the solution design, we explored options based on IML to develop user-centric and effective explanations for medical practice. After the diagnosis stage was conducted in July 2022, we started the ADR project with a design stage in August 2022, which evolved into an implementation stage in November 2022 and ended in January 2023.

4 Results of Designing a User-Centric XAI System

4.1 Synthesizing Literature for a Theory-Informed Artifact Design

In this diagnosis stage, we derived MRs from literature a priori to artifact instantiation [49], to guide the eADR process and inform the artifact design. As our study addresses user-centric explanations in medical imaging, we selected the key concepts of AI-enabled CADx, XAI, IML, and user-centric design to derive MRs and validated them in expert interviews, in line with ADR [48, 49]. For each concept, we constructed MRs that are abstract and tied to the solution objective [48], based on frequently cited studies as well as recent reviews within the respective communities (e.g., [22, 29, 32, 33], respectively) to build the knowledge base (cp. Sect. 2).

MR1 – Practically-Evaluated Diagnosis Support. AI-enabled CADx systems are highly accurate but require purposefully designed human-AI collaboration [25]. Thus, practical solutions and testing in clinical routine contexts are necessary [22].

MR2 – Inherent Explainability. Trustable and successful AI interactions require clinicians' understanding [20, 26]. A trustworthy XAI design in healthcare must embody an inherently explainable model to avoid pitfalls of post-hoc interpretability [20] and to overcome unimodal, data-driven XAI workflows [22, 33].

MR3 – Informed Machine Learning Component. Integrating medically relevant knowledge in machine learning models can overcome the limits of data-driven systems and enable knowledge-informed explanations [13, 15].

MR4 – User-Centric System Design. A user-centric XAI design must involve end-users from early on in development [27]. Clinician-centric XAI should provide information that is useful for patient-specific decision-making, and leverage validated medical knowledge [6, 11, 34, 35]. The explanations should mirror clinicians' way of thinking and integrate cleanly with their diagnostic workflow [40, 41], as well as provide information that users can connect to existing clinical processes [9, 42].

4.2 Design of a Conceptual XAI Prototype for Image Analysis

As it lacks knowledge on how IML can help to design user-centric and effective explanations for diagnostic imaging, we initiated the eADR process with a design stage.

Problem Formulation and Planning. Job shadowing and expert interviews helped us refine the problem and explore the design space of the artifact. After the radiologists validated the MRs, we pursued chest CT as an essential diagnostic tool across all patient populations. Thereby, pulmonary nodules are always a critical concern, as they occur in all age groups and are examined on every CT scan. The diagnostic difficulty consists of the detection of benign nodules as a sign of early lung cancer. The cooperating radiology clinic did not have an explainable CADx system in place. Instead, systems assist with the automated detection of nodule candidates and the semi-automated measurement of nodule size. As it is highly important to make a provable decision on whether a nodule has a risk of being malignant or not on the part of radiologists, we selected pulmonary nodules as a research case to explore XAI assistance. To guide the artifact, we reviewed key literature on PubMed, Google Scholar, and AIS eLibrary, using search terms (and a wide range of synonyms) related to disease prediction, medical imaging, XAI, and IML techniques. The search was restricted to English articles published between 2016 and 2022. Articles were included if they explain an image diagnosis by integrating, enriching, or combining a dataset with a separate source of medical knowledge, resulting in 11 relevant papers.

Prior AI studies of lung nodules are unable to explain the predicted malignancy [50]. One type of XAI extracts high-level concepts learned by deep learning [37], called concept attribution. It links model units, e.g., convolutional neural network (CNN) channels, with a separate set of human-friendly semantic concepts. The attribution, however, represents global post-hoc explanations (contrary to MR2) [51].

Other IML techniques build predictions based on pre-defined concepts [52]. These, however, rely classification only on a few expert-defined clinical concepts annotated to entire images, without combining deep image-based features. Another technique is multi-task learning (MTL) that learns multiple outputs simultaneously [37]. It has been used to learn diagnostic criteria jointly with the diagnosis, e.g., a dermoscopic 7-point checklist for skin cancer prediction [53]. As there are no connections between high-level outputs, one cannot extract mutual influence, e.g., between criteria and diagnosis [54].

One study overcame this by arranging two-tier predictions in a hierarchical CNN with a global loss function [55]. Here, the intermediate predictions of low-level semantic properties and learned image features of the CNN were concatenated into a high-level subnetwork for final diagnosis prediction. This way, the model learned from hybrid information, i.e., image and semantic features, both inherently influencing diagnosis prediction. Diagnostic criteria (e.g., the pattern or composition of nodules) are intuitive to radiologists since they are educated to apply them to chest images (MR4). Thus, we adapted a two-tier CNN [55] to not only predict whether a criterion is fulfilled but also provide its influence on a particular malignancy result.

Artifact Creation. Given an image of a possible nodule, our first intended IML prototype predicted intermediate characteristics and the final malignancy risk. Hereby, we integrated all low-level characteristics of nodules reported in the related literature, including subtlety, internal structure, calcification, sphericity, margin, lobulation, spiculation, as well as texture [23]. To increase transparency, we used more granular labels per criterion (e.g., non-solid, partly-solid, or solid texture) compared to the binary labels used in [55]. To extract their numerical importance for nodule malignancy, we adapted the integrated gradients XAI technique since it applies to any differentiable model and does not require modifications of the original network [56].

Evaluation. We gathered feedback from the clinicians on the usefulness of information integrated into the conceptual prototype (MR1) and the correspondence to the diagnostic workflow (MR4). First, radiologists had differing opinions and prioritizations regarding explanatory criteria (explanators). The size, number, localization, texture, as well as morphological features of nodules (shape and margin) were most important across all interviews. Notably, the first three features were not covered by our artifact, though being highly relevant. We attributed this to the fact that reviewed XAI approaches applied classification to individual nodule regions rather than overall lung volumes due to intended performance advancements. Thereby, ML was utilized to only predict features that conventional CADx could not provide rather than using them jointly. Second, radiologists validated that these criteria were effective to explain malignancy. Some of them were perceived as "an ultra-detailed perspective (…) of a very small structure." Though granularity was appropriate, it likely differs from other medical specialties. Another radiologist suggested that explanators should not be too detailed as a patient may have *many* nodules. Hence, relevant diagnostic criteria should be highlighted, and information overload must be avoided.

Reflection. The patient-wise and diagnostic nature of explanations were approved. Though, two important explanators (number and position of nodules) could not be predicted when classifying individual nodule images. This turned out problematic: If radiologists let the AI classify regions of supposed nodules, they already have made a preselection (using diagnostic criteria) and the explanations would become redundant.

Learning. We could generalize the learnings from a lung nodule-related design to medical imaging tasks as design principles that address MRs as follows.

DP1: In order to provide appropriate explanations, AI tasks need to be designed in an end-to-end manner. As indicated, integrating medical knowledge for malignancy

prediction (MR3) did not per se render explanations appropriate. Applying classification to image regions implied that a radiologist had to preselect problematic areas. Consequently, explanations become obsolete. Thus, explanations should provide information unknown to clinicians (MR1) and perform both detection and classification.

DP2: In order to provide diagnostically relevant information, the system should provide criteria tailored to users' medical specialties. Finding the right information to be integrated into IML was not trivial (MR2). Morphological criteria of lung nodules were appropriate for the domain of radiologists in which they are practically applied (MR4). Other explanators (e.g., the number and size of nodules) aim at a more common medical understanding and are helpful to general practitioners assessing chest CT images. Thus, designing informative explanations requires adjustment to the medical understanding and analytical decision-making of the specialty (MR2, MR4).

4.3 Implementation of a Conceptual XAI Prototype for Image Analysis

We continued with development-centered activities and aimed to implement a functional IML artifact to demonstrate a useful instantiation of the solution class.

Problem Formulation and Planning. In this stage, we dealt with the additional automated detection of nodules and the refinement of the explanators based on the end-to-end design. CNNs can successfully be applied end-to-end (from the input of a raw CT volume to the classification), whereas determining the right network input for medical imaging is very complex: One way is to extract slides from the 3D volume as usual 2D inputs [57]. Then, slices are analyzed independently and many of them are not informative. Region of interest-based inputs are only useful to diseases that do not span over multiple areas. One can also apply 3D CNNs directly to a whole CT volume. This fully integrates spatial information but is prone to overfitting due to one-per-patient samples. Hence, we planned to use 3D patches for creating a nodule detection model, prepended to the explainable classification task. This has resulted in a larger sample size since many 3D patches can be extracted and used as input to a single CNN model. Reflecting on the novel set of possible explanators with radiologists, patient details, such as demographics and clinical history, were found inappropriate as they had been already known before image interpretation. Radiologists found "recommendations using the Fleischner table" more useful, as this guideline is referred to determine follow-up biopsy or CT intervals depending on the texture, number, and size of nodules. Thus, we planned to provide Fleischner-related criteria to increase artifact utility and facilitate follow-up decision-making.

Artifact Creation. As an IML artifact, we implemented a functional prototype consisting of nodule detection (Fig. 1, upper lane) and classification (lower lane). Both tasks were implemented in python based on PyTorch and pylidc libraries, using ConvNet architectures[1]. From the public LIDC-IDRI dataset[2], 1,017 low-dose lung CT scans associated with 1,012 (both positive and clean) patients and 1,388 nodules (each annotated from at least three physicians) were taken. CT scans were transformed to

[1] https://github.com/facebookresearch/ConvNeXt.

[2] https://wiki.cancerimagingarchive.net/x/rgAe.

Hounsfield scale, filtered [−1000, 500], and scaled to [0, 1]. To prevent data leakage, we used patient-level splits for training, validation, and test data (75%, 10%, and 15%). Both models were separately trained using PyTorch's 'AdamW' optimizer.

Model 1 (detection). For voxel-wise training of the nodule candidate detector (Fig. 1, top left box), we selected 1,095 positive (and applied thirtyfold random shift augmentation) and 116.4k random negative (i.e., nodule-free) 3D patches, in the size of 40 × 40 × 40 mm. We trained the model using binary cross-entropy loss. To detect final nodule locations, we applied the model to whole lung CT volumes using a sliding window, as common in medical imaging [58]. From the outputs in different windows, we kept the bounding boxes of ones with a sigmoid probability above 0.98. If two neighboring boxes are predicted as positives, we only kept the median one. Final locations were used as input to subsequent classification (Fig. 1, lower left box).

Model 2 (IML-based classification). Each detected candidate nodule is classified like [55]. It was trained based on the positive instances from the dataset used for model 1 (class distribution: 11.1%, 11%, 45.8%, 18.5%, and 13.6%, from benign to malignant), using a common classification loss (cross-entropy). We discarded internal structure, lobulation, and subtlety to match identified criteria. We added a size feature (nodule diameter) which we binned analogous to Fleischner guidelines [<4, 4–8, 8–20, 20–30, > 30]. Low-level image features (ConvNet output) were used to compute each nodule-level diagnostic criterion through fully connected layers (Fig. 1, blue lines). Image and criteria representations were merged and fed into another block of fully connected layers to compute the malignancy probability (golden lines).

XAI Output. To generate lung-level explanations, we extracted the number and position (positive bounding boxes) of nodules from model 1 (green lines). From model 2, we used diagnostic features, along with brief corresponding labels as well as their numerical importance, and malignancies as nodule-level explanations. The combination of both levels was used for patient-level summarization: Overall malignancy risk reported at least one malignant nodule being present and, if yes, the average malignancy. As favored by the radiologists, only the most malignant nodule was initially explained, and less malignant ones were displayed on request. Rule-based recommendations were added as per Fleischner guidelines, e.g., multiple 6–8 mm large nodules with solid texture and low malignancy risk should be CT screened after 3–6 months.

Evaluation. In semi-structured interviews, we used exemplary diagnosis explanations and validated complexity, medical relevance, and the effect on the routine. Radiologists advocated the multi-level approach, i.e., regarding overall malignancy as well as each nodule candidate. They perceived nodule-specific criteria and their importance as satisfactory to understand malignancy results, whereas one preferred even shorter explanations. All positively stressed that they can quickly compare criteria with their assessment and adopt them for documentation in case they agree. Overall, the system was found as a quality-enhancing cross-check to not overlook any nodes, with helpful explanatory information. It offered well-prepared characteristics, helping both residents to justify diagnoses as well as chief residents to verify reports. From an organizational view, this can qualitatively improve follow-up activities since high-risk patients could automatically be referred to further imaging or biopsy for confirmation.

For a quantitative evaluation, we applied metrics commonly used in image classification [57]. In detection (model 1), we achieved an F-score of 0.71, accuracy of 0.99, precision of 0.62, and recall of 0.84. Although the precision is relatively lower (probably due to the realistic test dataset), our detection is still highly accurate with comparably few false positives, compared to state-of-the-art results in deep learning [50]. The downstream model yielded a root-mean-square error (RMSE) of 0.79 for malignancy classification, which is low compared to the feature's standard deviation of 1.19. For performance comparison, we also binarized the features to match [55]. Malignancy (0.86), calcification (0.92), margin (0.77), texture (0.88), and sphericity (0.59) all resulted in equal or higher binary classification accuracy compared to [55]. Size and spiculation were predicted with an accuracy of 0.93 and 0.86, respectively.

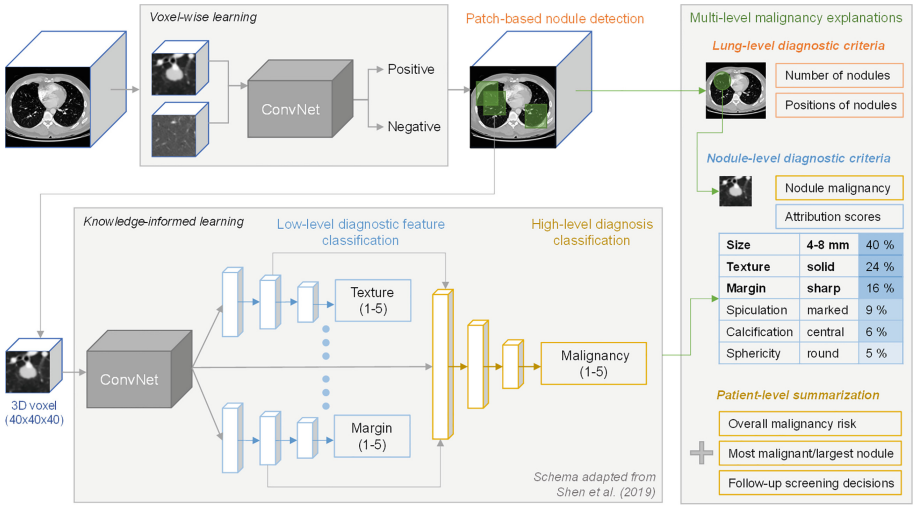

Fig. 1. Resulting user-centric XAI for lung nodule analysis based on IML.

Overall, the two-tier artifact (Fig. 1) provided higher granularity of explanators including their numerical influence, improving the practical relevance of explanations. Compared to binary nodule classification [55], IML is equally performant while ensuring radiologist-informed explainability. We have also gained a more realistic performance estimate and mitigated data leakage through patient-level holdout data.

Reflection. Contrary to initial assumptions, diagnostically relevant information with which clinicians are familiar was not automatically meaningful. Instead, definitions in close collaboration with clinicians showed context-specific usefulness. IML as a tool for knowledge integration was beneficial to clinician-centric and effective explainability. Not only could nodule classification be enhanced with informative explanations, but the detected characteristics also assisted in creating follow-up screening rules, further reducing manual work in nodule-related decision-making.

Learning. Further strands of image analyses, such as Alzheimer's disease detection, are seeking diagnostic-focused explanations [59]. Our developed artifact, therefore, may apply to any disease whose diagnosis relies on image interpretation. Among others, oncological diseases such as liver and bone metastases could also benefit from the proposed design of explanatory information as they adhere to structured diagnosis criteria, too. Thus, we could embody the IML-based solution within the general black-box problem of AI-based CADx systems. For the goal of designing user-centric and effective XAI for medical image analyses, we generalize two more learnings.

DP3: In order to provide useful explanations, the system should span medical explanations across ML tasks instead of model-driven explanations. Unlike common XAI approaches, model-specific explainability was not useful (MR1). In contrast, the mere nodule detection needed no extra explainability approach. Derived explanations centered around the facts of why a nodule has been predicted as malignant (MR4). As the overall system was deemed adequately explainable, explanations should be designed across ML tasks, not by what information a single ML model generates.

DP4: In order to provide knowledge-informed explanations, the system should distinguish information relevance based on the routine workflow. Probing the appropriateness of medical indicators (MR4) in practice yielded multiple insights. The patient context medically influences malignancy risk but was already known to radiologists before interpreting the image. Lung-level aspects were found appropriate, even with a more general medical understanding. Nodule-level aspects were found overly fine-grained and partly counterproductive for a busy routine, whereas this level of detail aided follow-up decision-making. Future XAI designs should consider the different levels of explanations entailing diverse implications for radiology routine (MR3).

5 Discussion and Conclusion

Data-driven explanations have not been able to provide clinicians with the appropriate understanding of AI-powered results. We hope to further contribute to the medical XAI challenge by actively involving clinicians in building a CADx system for image analysis. Our findings suggest that IML can surpass data-driven and post-hoc XAI and directly derive explanations from the inference process. Such 'intrinsic interpretability' benefits from desirable properties such as faithfulness, trustworthiness, and fidelity [60] since no black box needs to be approximated. While competitive performance could be achieved, IML involved a complex design of the network input to make end-to-end image detection and additional feature classification possible.

Another challenge entailed by IML is that explanatory benefits are not often readily apparent. While the two-tier hierarchy allowed for explainable feature predictions, other techniques such as knowledge graph integration may have less obvious effects.

Thanks to eADR, we have been able to gain valuable practical insights into malignancy assessment to constantly refine our artifact design. While quality frameworks are being developed and attribute many possible properties to explanations (e.g., [28]), our findings strongly support observations that explanations do not need to be exhaustive but must ensure usability when implemented in healthcare [5]. Radiologists did not

find all diagnostically relevant information meaningful but preferred actionable factors, though in dissenting granularity. Meanwhile, we provided explanations that span over two ML tasks, which ensured that the XAI system's purpose was useful. Consequently, workflow-level utility and system-level explainability were mutually dependent. In contrast to suggestions to additionally utilize patient demographic data [50], we observed that these were medically but not practically adequate. After all, we identified explanators that mimic radiologist diagnostics and integrate cleanly with the workflow. Considering the broader medical XAI field, we identified diseases of other specialties, including but not limited to oncology and neurology that are assessable through similar diagnostic criteria. These are very likely to benefit from IML, too. The proposed architecture provides flexibility to change the detection component to other diseases, intermediate features to other medical indicators, and follow-up rules. Ultimately, we proposed hands-on principles that facilitate the design of corresponding explanations. Nevertheless, our study has two limitations. First, nodule classification could encounter a consequential error if the first model produces a false detection, so the estimated performance may be optimistic. However, we expect this to have only a minor impact on the performance of our instantiation as we observed relatively few false positives. Second, the qualitative feedback on the explanatory information as part of our evaluation was limited to experts from one institution with many years of experience. Hence, we encourage research to practically test and evaluate the principles of IML design and resulting explainability methods in various healthcare settings. As the communication format may also affect explanation understanding, the variants to arrange such information within an interface could provide further important insights into the demands of clinical users.

For practically designing explanatory information in medical imaging, our study faced heterogeneous information according to what CADx systems semi-automatically obtain, what radiologists desire, what is obsolete due to the workflow, as well as what medical guidelines suggest. The order of importance was also different. In summary, we encourage future developments to unite the medical knowledge available with user-centric XAI purposes. Through a proposed multi-level explanation approach, medically relevant and actionable criteria could and should be consolidated. It follows that identifying the correct intersection areas may also reduce the manual efforts required to document image interpretation and aftercare decision-making.

References

1. Pumplun, L., Fecho, M., Islam, N., Buxmann, P.: Machine learning systems in clinics – how mature is the adoption process in medical diagnostics? In: Proceedings of the 54th Hawaii International Conference on System Sciences (2021)
2. Johnson, M., Albizri, A., Harfouche, A.: Responsible artificial intelligence in healthcare: predicting and preventing insurance claim denials for economic and social wellbeing. Inf. Syst. Front. (2021)
3. Topol, E.J.: High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. **25**(1), 44–56 (2019)
4. Wiens, J., et al.: Do no harm: a roadmap for responsible machine learning for health care. Nat. Med. **25**(9), 1337–1340 (2019)
5. Arbelaez Ossa, L., Starke, G., Lorenzini, G., Vogt, J.E., Shaw, D.M., Elger, B.S.: Re-focusing explainability in medicine. Digital Health **8** (2022)

6. Markus, A.F., Kors, J.A., Rijnbeek, P.R.: 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. **113** (2021)
7. Payrovnaziri, S.N., et al.: Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. JAMIA **27**(7), 1173–1185 (2020)
8. Fernandez-Quilez, A.: Deep learning in radiology: ethics of data and on the value of algorithm transparency, interpretability and explainability. AI Ethics **3**(1), 257–265 (2022)
9. Jacobs, M., et al.: Designing AI for trust and collaboration in time-constrained medical decisions: a sociotechnical lens. In: Kitamura, Y., Quigley, A., Isbister, K., Igarashi, T., Bjørn, P., Drucker, S. (eds.) Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–14. ACM, New York (2021)
10. Li, X., Qian, B., Wei, J., Zhang, X., Chen, S., Zheng, Q.: Domain knowledge guided deep atrial fibrillation classification and its visual interpretation. In: Zhu, W., et al. (eds.) International Conference on Information and Knowledge Management, pp. 129–138. ACM, New York (2019)
11. Ribera, M., Lapedriza, A.: Can we do better explanations? A proposal of user-centered explainable AI. In: Proceedings of the IUI Workshops. ACM, New York (2019)
12. Bauer, K., Hinz, O., van der Aalst, W., Weinhardt, C.: Expl(AI)n it to me – explainable AI and information systems research. Bus. Inf. Syst. Eng. **63**(2), 79–82 (2021). https://doi.org/10.1007/s12599-021-00683-2
13. Gaur, M., Faldu, K., Sheth, A.: Semantics of the black-box: can knowledge graphs help make deep learning systems more interpretable and explainable? IEEE Internet Comput. **25**(1), 51–59 (2021)
14. Beckh, K., et al.: Explainable Machine Learning with Prior Knowledge (2021)
15. von Rueden, L., et al.: Informed machine learning - a taxonomy and survey of integrating prior knowledge into learning systems. IEEE Trans. Knowl. Data Eng. **35**(1), 614–633 (2021)
16. Doshi-Velez, F., Kim, B.: Considerations for evaluation and generalization in interpretable machine learning. In: Escalante, H.J., et al. (eds.) Explainable and Interpretable Models in Computer Vision and Machine Learning. TSSCML, pp. 3–17. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98131-4_1
17. Sein, M.K., Henfridsson, O., Purao, S., Rossi, M., Lindgren, R.: Action design research. MIS Q. **35**(1), 37–56 (2011)
18. Mullarkey, M.T., Hevner, A.R.: An elaborated action design research process model. EJIS **28**(1), 6–20 (2019)
19. Fernández-Loría, C., Provost, F., Han, X.: Explaining data-driven decisions made by AI systems: the counterfactual approach. MIS Q. **46**(3), 1635–1660 (2022)
20. Salahuddin, Z., Woodruff, H.C., Chatterjee, A., Lambin, P.: Transparency of deep neural networks for medical image analysis: a review of interpretability methods. Comput. Biol. Med. **140**, 105111 (2021)
21. Cheng, J.-Z., et al.: Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. Sci. Rep. **6**, 1–13 (2016)
22. Rajpurkar, P., Chen, E., Banerjee, O., Topol, E.J.: AI in health and medicine. Nat. Med. **28**(1), 31–38 (2022)
23. Hancock, M.C., Magnan, J.F.: Lung nodule malignancy classification using only radiologist-quantified image features as inputs to statistical learning algorithms. J. Med. Imaging **3**(4), 044504 (2016)
24. Grüning, M., Trenz, M.: Me, you and AI - managing human AI collaboration in computer aided intelligent diagnosis. In: SIGHCI 2021 Proceedings (2021)
25. Hinsen, S., Hofmann, P., Jöhnk, J., Urbach, N.: How can organizations design purposeful human-AI interactions: a practical perspective from existing use cases and interviews. In: Proceedings of the 55th Hawaii International Conference on System Sciences (2022)

26. Alam, L., Mueller, S.: Examining the effect of explanation on satisfaction and trust in AI diagnostic systems. BMC Med. Inform. Decis. Making **21**(1), 178 (2021)
27. Braun, M., Harnischmacher, C., Lechte, H., Riquel, J.: Let's get physic(AI)l - transforming AI-requirements of healthcare into design principles. In: ECIS 2022 (2022)
28. Vilone, G., Longo, L.: Notions of explainability and evaluation approaches for explainable artificial intelligence. Inf. Fusion **76**, 89–106 (2021)
29. Tjoa, E., Guan, C.: A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans. Neural Netw. Learn. Syst. **32**(11), 4793–4813 (2021)
30. Oberste, L., Heinzl, A.: User-centric explainability in healthcare: a knowledge-level perspective of informed machine learning. IEEE Trans. Artif. Intell. 1–18 (2022)
31. Saporta, A., et al.: Benchmarking saliency methods for chest X-ray interpretation. Nat Mach Intell **4**(10), 867–878 (2022)
32. Li, X.-H., et al.: A survey of data-driven and knowledge-aware explainable AI. IEEE Trans. Knowl. Data Eng. (2020)
33. Ghassemi, M., Oakden-Rayner, L., Beam, A.L.: The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit. Health **3**(11) (2021)
34. Zihni, E., et al.: Opening the black box of artificial intelligence for clinical decision support: a study predicting stroke outcome. PLoS ONE **15**(4) (2020)
35. Sun, Z., Dong, W., Shi, J., Huang, Z.: Interpretable Disease Prediction based on Reinforcement Path Reasoning over Knowledge Graphs (2020)
36. Choi, E., Bahadori, M.T., Song, L., Stewart, W.F., Sun, J.: GRAM: graph-based attention model for healthcare representation learning. In: ACM SIGKDD, pp. 787–795 (2017)
37. Deng, C., Ji, X., Rainey, C., Zhang, J., Lu, W.: Integrating machine learning with human knowledge. iScience **23**(11) (2020)
38. Lahav, O., Mastronarde, N., van der Schaar, M.: What is interpretable? Using machine learning to design interpretable decision-support systems (2018)
39. Lebovitz, S.: Diagnostic doubt and artificial intelligence: an inductive field study of radiology work. In: ICIS 2019 Proceedings (2019)
40. Tonekaboni, S., Joshi, S., McCradden, M.D., Goldenberg, A.: What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use (2019)
41. Evans, T., et al.: The explainability paradox: challenges for xAI in digital pathology. Futur. Gener. Comput. Syst. **133**, 281–296 (2022)
42. Pazzani, M., Soltani, S., Kaufman, R., Qian, S., Hsiao, A.: Expert-informed, user-centric explanations for machine learning. In: AAAI, vol. 36, no. 11, pp. 12280–12286 (2022)
43. Das, A., Rad, P.: Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey (2020)
44. vom Brocke, J., Winter, R., Hevner, A.R., Maedche, A.: Special issue editorial –accumulation and evolution of design knowledge in design science research: a journey through time and space. JAIS **21**(3), 520–544 (2020)
45. Peffers, K., Tuunanen, T., Niehaves, B.: Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research. EJIS **27**(2), 129–139 (2018)
46. Chari, S., Seneviratne, O., Gruen, D.M., Foreman, M.A., Das, A.K., McGuinness, D.L.: Explanation ontology: a model of explanations for user-centered AI. In: Pan, J.Z., et al. (eds.) ISWC 2020. LNCS, vol. 12507, pp. 228–243. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-62466-8_15
47. Gilpin, L.H., Testart, C., Fruchter, N., Adebayo, J.: Explaining Explanations to Society (2019)
48. Möller, F., Guggenberger, T.M., Otto, B.: Towards a method for design principle development in information systems. In: Hofmann, S., Müller, O., Rossi, M. (eds.) DESRIST 2020. LNCS, vol. 12388, pp. 208–220. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-64823-7_20

49. Chandra, L., Seidel, S., Gregor, S.: Prescriptive knowledge in IS research: conceptualizing design principles in terms of materiality, action, and boundary conditions. In: Proceedings of the 48th Hawaii International Conference on System Sciences, pp. 4039–4048 (2015)
50. Jassim, M.M., Jaber, M.M.: Systematic review for lung cancer detection and lung nodule classification: taxonomy, challenges, and recommendation future works. J. Intell. Syst. **31**(1), 944–964 (2022)
51. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: quantifying interpretability of deep visual representations. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3319–3327 (2017)
52. LaLonde, R., Torigian, D., Bagci, U.: Encoding visual attributes in capsules for explainable medical diagnoses. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 294–304. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_29
53. Murabayashi, S., Iyatomi, H.: Towards explainable melanoma diagnosis: prediction of clinical indicators using semi-supervised and multi-task learning. In: International Conference on Big Data, pp. 4853–4857. IEEE (2019)
54. Lucieri, A., Dengel, A., Ahmed, S.: Deep learning based decision support for medicine—a case study on skin cancer diagnosis (2021)
55. Shen, S., Han, S.X., Aberle, D.R., Bui, A.A., Hsu, W.: An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. Expert Syst. Appl. **128**, 84–95 (2019)
56. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Proceedings of the 34th ICML, vol. 70, pp. 3319–3328. PMLR (2017)
57. Wen, J., et al.: Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. Med. Image Anal. **63**, 101694 (2020)
58. Wu, J., Qian, T.: A survey of pulmonary nodule detection, segmentation and classification in computed tomography with deep learning techniques. J. Med. Artif. Intell. **2**, 1–12 (2019)
59. Dyrba, M., Hanzig, M., Altenstein, S., Bader, S., Ballarini, T., Brosseron, F., Buerger, K., et al.: Improving 3D convolutional neural network comprehensibility via interactive visualization of relevance maps: evaluation in Alzheimer's disease. Alzheimer's Res. Ther. **13**(1), 1–18 (2021)
60. Pintelas, E., Livieris, I.E., Pintelas, P.: A grey-box ensemble model exploiting black-box accuracy and white-box intrinsic interpretability. Algorithms **13**(1), 17 (2020)