KI - Künstliche Intelligenz (2024) 38:157–167
https://doi.org/10.1007/s13218-024-00835-y

PROJECT REPORTS

Human-Centered Explanations: Lessons Learned from Image Classification for Medical and Clinical Decision Making

Bettina Finzel[1]

Received: 28 August 2023 / Accepted: 9 January 2024 / Published online: 14 February 2024

Abstract
To date, there is no universal explanatory method for making decisions of an AI-based system transparent to human decision makers. This is because, depending on the application domain, data modality, and classification model, the requirements for the expressiveness of explanations vary. Explainees, whether experts or novices (e.g., in medical and clinical diagnosis) or developers, have different information needs. To address the explanation gap, we motivate human-centered explanations and demonstrate the need for combined and expressive approaches based on two image classification use cases: digital pathology and clinical pain detection using facial expressions. Various explanatory approaches that have emerged or been applied in the three-year research project "Transparent Medical Expert Companion" are shortly reviewed and categorized in expressiveness according to their modality and scope. Their suitability for different contexts of explanation is assessed with regard to the explainees' need for information. The article highlights open challenges and suggests future directions for integrative explanation frameworks.

Keywords Explainable Artificial Intelligence · Digital healthcare · Human-centered explanations · Explanation expressiveness · Explanation frameworks

1 Introduction

In recent years AI-based image classification has become an important tool for high stakes decisions, e.g., to support or even improve medical diagnoses based on image or video material [21, 52]. However, the use of AI is still largely limited to the field of research and development [23]. For widespread use in medical and clinical facilities, several requirements must first be met. Among the most important characteristics to be mentioned here are trustworthiness and reliability as well as transparency of AI systems, the latter being a prerequisite for fulfilling the first two requirements or making them measurable [7, 18]. It is particularly important to look at how confident or uncertain a system is in its decision and what reasons led to a particular outcome [3, 34]. Explainable Artificial Intelligence (XAI), which has grown into an entire branch of research, addresses precisely these aspects and has developed many methods to explain AI systems, especially for image classification tasks [47].

Explainable image classification with Convolutional Neural Networks (CNNs) is a vivid research field [47]. CNNs produce complex models that are opaque and not very comprehensible to humans [7]. This is both a challenge and an obstacle for developers of decision support systems in medicine and clinical settings, for experts who are to assess whether decisions are appropriate to the domain and made for the right reasons and strategies, but also for novices who want to learn how and why an AI system derives its decisions and also which features in a data set lead to the classification outcome itself. Most of the existing explanatory methods for making CNNs transparent and comprehensible have so far been built by and for developers [47]. These methods usually use highlighting on input data to emphasize which pixels in an image contributed positively or negatively to a classification.

Although, visualizations as explanations for image classifications match the modality of the input data (images), they are not always sufficient to make all relevant properties transparent [17, 46]. For example, visualizations cannot express absence as images are always composed of the same

✉ Bettina Finzel
bettina.finzel@uni-bamberg.de

[1] Cognitive Systems, University of Bamberg, Bamberg, Germany

amount of pixels, regardless of whether these pixels show an object of interest or not. Although negation can be displayed indirectly via a color-based visualization of negative importance derived from the parameters of a model, the meaning of relevance is not unambiguous. Its meaning depends on the correctness of a classification decision and the properties of contrasting class(es). Furthermore, relations between image areas may play a role, where opposing classes share a common set of properties but differ in the spatial composition or temporal sequence of properties. This holds especially for images from video data. The current visualization methods are limited in their expressiveness in this regard. Relations in comparison to visualizations can usually be easily verbalized and translated into natural language, which is an advantage over visualizations [7, 46]. Nevertheless, visualizations may efficiently provide relevant information "at a glance", making sequential processing of natural language explanations obsolete. Moreover, the explanatory power varies dependent on whether an explanation should provide *global* information about a model (i.e., how a class decision is generally derived) or whether, at a *local* level, only the classification of instances should be explained [3, 47]. Scaffolding between the two levels, that is switching from general, abstract information to more specific, detailed information and vice versa, can also be helpful or even required for understanding [19, 33]. In that sense, there is no "one size fits all" explanation method [49]. In order to explain the decision of an opaque CNN model, we thus suggest to use *multimodal explanations*, i.e., a combination of methods that provide various representations and which are more expressive all together.

Another important aspect is the human-centricity of explanations [34, 49], especially in high-stakes application fields such as medical or clinical diagnosis [10, 26]. The act of explaining constitutes human behavior and is closely linked to linguistic expression and, as explained above, can be supplemented by other modalities such as visualizations in order to convey the most appropriate information to the recipients of explanations, the explainees [17, 47]. The human-centered perspective addresses the *information need* of explainees. What is being explained and what is of interest? How should it be explained so that explainees understand the transmitted information? The information need varies based on the target explainees. Experts in a knowledge domain want to be able to validate a model and are therefore more interested in checking the performance of a model or the quality of explanations produced. Experts may even want to take corrective action and adapt the model, the explanations, or the data. Novices in a knowledge domain are primarily interested in how a model came to its classification decision and which properties from the data set lead to its prediction as they need to learn from the model and presented explanations about the domain. Developers share requirements with experts and novices, since they have the knowledge of how a classification method works (mechanistic understanding according to [39]), but first have to learn how the method behaves on given data to build up knowledge about the requirements of the domain (functional understanding according to [39]). Explanations that are understandable for humans lead to a higher perceived transparency and reliability of AI and thus increase trust in such systems [51]. Thus, human explainees may benefit from a bi-directional, human-guided framework (see Fig. 1) that allows model decisions to be explored, understood and, if

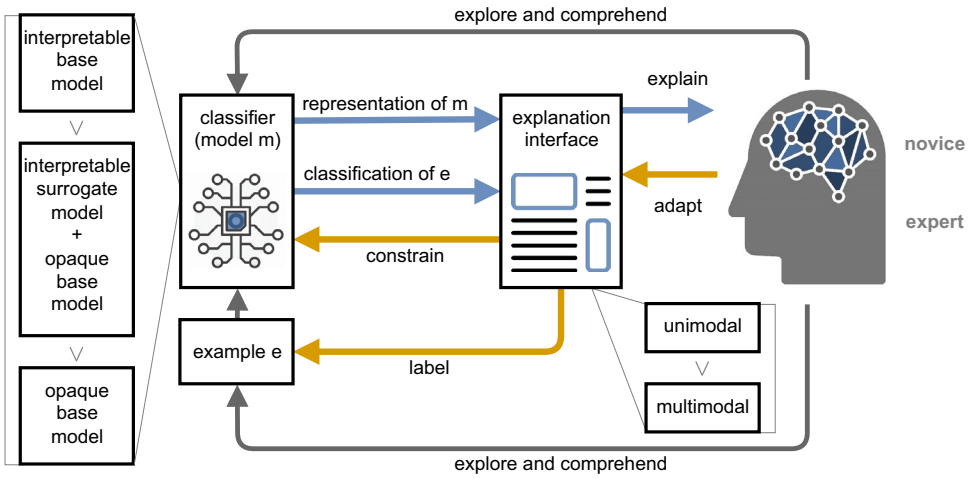

Fig. 1 A bi-directional explanation framework, where humans receive explanations of classification decisions either for an interpretable base model, an opaque base model or an opaque base model that is approximated by an interpretable surrogate model. Explanations are presented either as unimodal representations or multimodal representations. The human (expert or novice) can explore and comprehend the system, in particular the model, the classification of individual examples as well as given explanations through interacting with an explanation interface. The framework was adapted based on [46]

necessary, evaluated and corrected based on expressive explanations to built better human-AI partnerships [46].

In this article, we present explanatory methods and our lessons learned each from a medical and a clinical use case. The article first introduces related background in Sect. 2, presents methods for building expressive explanations with varying scope and modality in Sect. 3 and summarizes lessons learned from the *Transparent Medical Expert Companion* research project in Sect. 4. Open challenges and suggestions for future directions are briefly discussed in Sect. 5. Section 6 concludes the article.

2 Background

2.1 Explaining Image Classifications

As shown in various current review articles, the use of CNNs in medical and clinical image classification has increased significantly during the last years [3, 4, 57]. Due to the opacity of CNN-based models, the development of explanatory methods for medical and clinical research has also increased [9, 57]. Presenting all state-of-the-art explanatory methods for image-based diagnoses would miss the aim of this article. We therefore refer here to the most recent and very comprehensive overview articles [3, 4, 10, 12, 18, 32, 47, 54, 57] and limit ourselves to visual and relational explanatory methods that were used in the *Transparent Medical Expert Companion* project. These methods include Layer-wise Relevance Propagation (LRP) [36], Local Interpretable Model-agnostic Explanations (LIME) [43], Gradient-weighted Class Activation Mapping (Grad-CAM) [48], and Inductive Logic Programming (ILP) [37]. ILP is a relational interpretable machine learning approach that can be used to learn surrogate models for opaque base models [42] and which has been applied very successfully in relational domains like medicine and molecular biology in the past [5]. For a global and extended view of LRP-based explanations, the clustering method SpRAy [30], which is based on spectral analysis and suitable for detecting *Clever Hans* predictions [24], was also applied and extended. All explanatory methods mentioned here, except for ILP, make use of highlighting on individual pixels or pixel groups in images. The scope of used methods is mostly local, except for ILP and SpRAy, which provide also global views on classification outcomes. The more critical reviews (see for example in [3, 12, 18]) pointed out that investigating only local explanations is not enough to assess the trustworthiness and reliability of models. They argue that global views on a model's behavior are needed and potentially further evaluation methods and quality measurements for audit. This underscores the need to combine different explanatory methods and to properly evaluate models.

A rigorous evaluation of models is particularly crucial in medicine and clinical use cases as balanced annotated data is often unavailable, data may be noisy and sparse or available gold standards may be prone to limited reliability as well as validity [12, 18]. Since deep learning based approaches like CNNs heavily rely on large amounts of data sets available, poor quality is a major threat to trustworthy and reliable diagnosis [12]. In the next section we present the two use cases to which these challenges partially apply and both use cases do thus benefit from expressive explanations.

2.2 Digital Pathology

Last year, Andrey Bychkov and Michael Schubert published a comprehensive report about the global decline in pathologists across all continents [8]. According to their findings, experts see possible solutions to combating the shortage of medical specialists in the promotion of staff training, but also in the development of digital assistance systems. This is in line with one of the goals of the *Transparent Medical Expert Companion* research project: to integrate an explainable classifier for microscopy data of the human intestine for the detection and staging of colon cancer into such an assistance system. The task of the classifier was to detect different types of tissue and their position in relation to each other to diagnose pathological changes.

Colon cancer is typically assessed by pathologists using medical classifications such as Wittekind's standard work on the TNM classification [56]. Figure 2 shows on the left how such a microscopy image of human tissue is composed and how such a tissue can be viewed in its hierarchical and spatial resolution by zooming in and out. The main challenge is that cancerous tissue may mix with tissue that would be considered healthy in the absence of a tumor, imposing the need for considering the context of tissues when assessing, e.g., the invasion depth of a tumor. The invasion depth can be characterized by analysing the spatial relations between different colon tissues, e.g., a tumorous sample is to be considered as containing a tumor of stage 2 if the tumor itself invades areas of the intestine muscle tissue. When explaining image classifiers for digital pathology, it may therefore be of interest to use methods that can represent spatial relations as well as the hierarchical composition of tissues (a tissue is made of cells, each cell may contain a nucleus and further underlying morphological structures [35].

2.3 Affective Computing

Affective computing is a broad field that encompasses human emotion recognition and sentiment analysis as well as research on and applications of cognitive, empathetic and intelligent systems [55]. The term was coined in 1997 by Rosalind Picard and since then, this area of research has

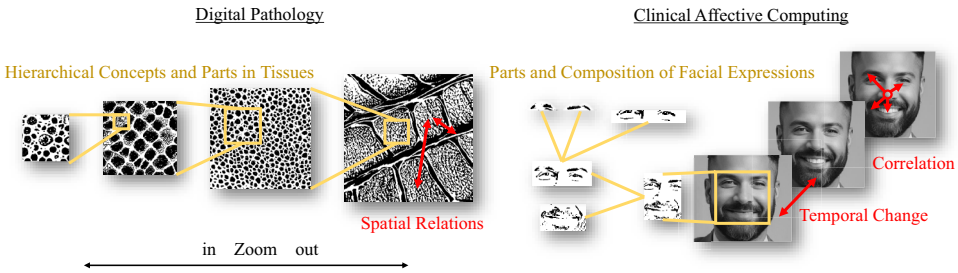

Fig. 2 Two use cases, one in digital pathology and another in clinical affective computing have been considered for the *Transparent Medical Expert Companion* project. To avoid any privacy issues, the data presented here was generated with a stable diffusion model. Any similarities in the resemblance with true tissue samples or facial recordings would be coincidental

found its way into clinical application scenarios [40]. The automatic analysis of facial expressions as a basis for emotion recognition is clinically relevant, for example, when people are unable to articulate their own emotions due to cognitive impairments [22, 29]. This is particularly relevant for the detection of pain and in individually tailored pain treatment. In this article, we focus on the use case of pain detection using human facial expressions, which are based on so-called action units [22]. Facial expression recognition is a suitable application use case for the development of explanation methods, since this task poses a major challenge for classification models due to the high degree of individuality of facial expressions and emotions, as well as the imbalance of the available data, missing annotations in existing databases or only sparsely representative data sets [31]. Explanation methods are therefore helpful to identify possible errors, biases or outliers in model behavior, or data.

The right half of Fig. 2 shows which properties of a data set are of interest for clinical facial expression recognition. For example, action units that often appear together in certain facial expressions, such as pain or other emotions, naturally correlate. In addition, facial expressions and thus also action units change over time. This concerns, among others, the intensity of an expression, the localization and the extent of an action unit on the face as well as the change in the correlation between different action units. A prerequisite for an AI system that provides trustworthy and reliable facial expression recognition is, on the one hand, that it can learn correlations of action units and their changes over time and, on the other hand, also recognizes that facial expressions not only consist of a composition of individual action units, but also that the action units themselves consist of individual, movable parts of the face, such as eyebrows, eyelids, and pupils [44]. The explanatory methods we present here were developed for CNNs trained on ready-to-use single image data or extracted video frames.

2.4 Human-Centered Explanations

A first step toward human-centered explanations is the question of who the explainees are, what knowledge they possess and what information they need [17, 47, 51]. The methods presented in this article are aimed in particular at medical and clinical experts, but are also suitable for explaining model behavior and classification criteria to novices.

Experts and novices differ in their information processing systems and capabilities [20]. Experts have acquired expertise in one subject, allowing them to recognize meaningful patterns in their field of expertise across different information modes. They are more efficient at recognizing and selecting problem-relevant information, although they initially invest more time in problem representation compared to novices. Experts show superior memory performance by being able to efficiently recall information from long-term and short-term memory. This superior performance is attributed to experts' ability to store knowledge in a principle-based manner, using knowledge schemata known for example as *chunks* to classify new information functionally [20].

Novices, on the other hand, must first build up knowledge and experience in a subject with the aim of eventually reaching the level of knowledge and competence of an expert. Novices are in general to be distinguished from laypersons. Although both groups must first receive information in order to be able to make more competent decisions, laypersons do not aim to become experts in a field. For example, medical intervention is dependent on the patient's consent. The patient should be aware of the possible consequences, but does not need to know exactly how the procedure is performed by an expert [6].

What are the implications of expertise research for human-centered explanations of AI-based classification? Psychological research has shown that for humans it is easier recognizing information or items than recalling them from memory [27, 53]. This may even apply to cognitively impaired people [25]. The advantage of recognition over

recall lies in the availability of informative cues. In this respect, psychological studies have shown that the performance of subjects who were only allowed to use recall approximated the performance of subjects tested under a recognition condition as the number of cues increased [53].

From this it can be concluded that explanations providing more information contribute better to the understanding of a classification, since fewer facts and relationships have to be recalled from memory. However, it is also to be expected that experts, due to the better cognitive linking and processing of information compared to novices, need less detailed explanations and can also gain an understanding of the behavior of a classification system from aggregated and more abstract material.

Good explanations range from much to little detail, depending on the information needs of the explainee [34]. Experts are usually interested in finding structures that are in line with their knowledge (which we refer to *validation by recognition*). Visual explanations in terms of highlighting or a verifiable summary may be sufficient for that purpose [54]. Novices rather want to know what structures are characteristic of a particular classification outcome or diagnosis (which is basically *learning from new information*). Relevant aspects need to be broken down in much more detail [11, 19, 33]. This may also apply to experts, for example in complicated or unclear cases [17].

Good explanations are selective and are limited to the essential aspects that the explainee needs to know [28, 34]. The most prominent strategies found in the explanation literature include explanations based on prototypes, contrastive explanations, or abstractions [47]. A dialogue-based interaction between the explainee and the explanation interface can help navigate through the spectrum of the level of detail or to switch between different explanation strategies that are available [16]. Prototype-based explanations are particularly suitable in domains where generalization to data is difficult, for example in the detection of skin diseases [41]. Contrastive explanations are particularly helpful where classes are very similar and can easily be confused with one another due to a few different characteristics [34]. This is true, for example, for facial expressions of pain versus disgust [29]. These are two conditions that are therefore best distinguished from one another by contrast [14]. The smallest difference in similar classes can be found on the basis of a contrastive explanation using near misses or counterfactuals [14, 34]. Just like the other two strategies, abstraction can help to aggregate information. Two approaches are generalization, i.e., abstraction from individual samples, or hiding and removing irrelevant information [13, 45].

The next section presents explanation methods that cover these aspects either individually or in combination with the aim to provide expressive, human-centered explanations.

3 Expressive Explanations with Varying Scope and Modality

Figure 3 categorizes selected methods from the *Transparent Medical Expert Companion* project that differ in scope, modality and ultimately in expressiveness. The y-axis for scope places local explanation methods at the bottom and global explanations at the top. Mixed, multiscope approaches, like explanatory dialogues, range over the whole y-axis. The axis for modality places visual explanation methods at the front, verbal explanations at the back and multimodal approaches in between. The further a method is to the right, the higher its expressiveness. The methods use different explanation strategies (such as prototypes, near misses, scaffolding) for which no additional dimension was added to the figure. In the following paragraphs, each approach

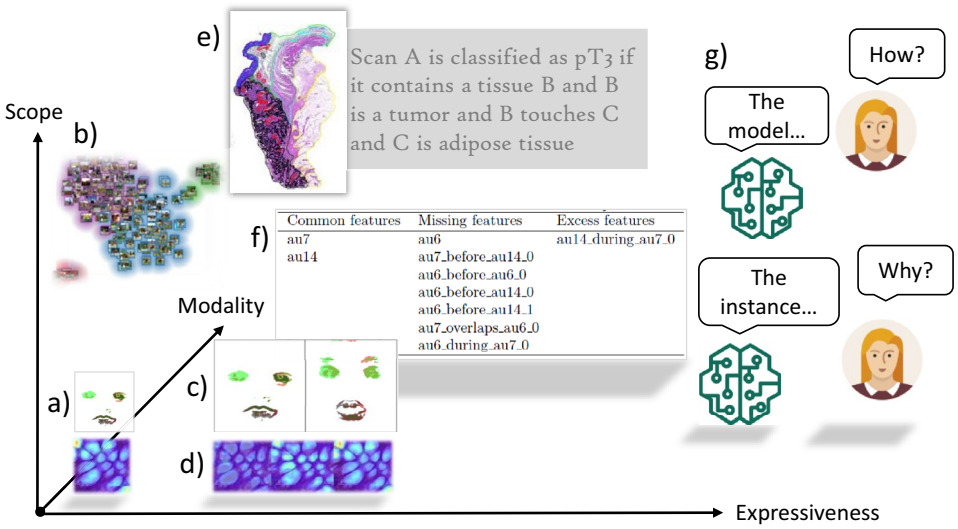

Fig. 3 A three-dimensional model that categorizes explanations according to their scope (from local to global, or both across the whole range of the y-axis), modality (front: visual, back: verbal, middle: multimodal), and their expressiveness (illustrated on an approximate range from low to high). Shadows indicate a bottom position with respect to the y-axis (local explanations). The other methods provide a more global view on classifier decisions or a mixture of local and global perspectives (e.g., an explanatory dialogue)

of interest is shortly introduced and as a summary listed according to its properties in Table 1.

3.1 Visual Explanations

The first explanation method we are going to introduce is LRP, a visual explanation technique. Montavon et al. [36] give a more detailed overview on LRP for interested readers. LRP is a technique that enables explainability for complex deep neural networks and that is particularly tested for CNN architectures from the VGG and ResNet family [30]. LRP involves propagating the prediction backwards in the neural network using specially designed propagation rules. LRP can be theoretically justified as a *Deep Taylor Decomposition* [36] and visualizes this decomposition in the form of heatmaps. This method was primarily developed to check the generalization capabilities of models for image classification based on visual and local explanations. Its placement in Fig. 3 is therefore at part (a) with the respective expressiveness of highlighting methods. This means that LRP provides an abstract view on the classification of instances mainly for experts.

On top of LRP the work of Lapuschkin et al. [30] proposes a cluster-based evaluation method. The authors introduce a semi-automated spectral relevance analysis over pixel relevance (SpRAy) and show that their approach is an effective method for characterizing and validating the behavior of nonlinear learning methods, like CNNs. They demonstrate that SpRAy helps to identify so-called *Clever Hans effects* [24] and thus to assess whether a learned model is indeed reliable for the problem for which it was designed. We consider SpRAy to be more global in scope than LRP, since the clusters produced hint at more general pattern in the classification and features of input images. We place it at part (b) in Fig. 3 above part (a) on the scope axis as it uses visualization as modality and may not be much more expressive than LRP or other highlighting methods. We consider it mainly relevant to experts.

Finzel et al. [13] make use of both, LRP and SpRAy based on the best practices and code bases introduced in [36] and [30]. Their work introduces a novel method for deriving temporal prototypes from CNN-based action unit predictions on video sequences. By clustering similar frames based on relevance computed with LRP, temporal prototypes are derived per cluster and input sequence. The prototypes offer an aggregated view of a network's frame-wise predictions while preserving the temporal order of expressions, thus reducing the evaluation effort for human decision makers for sequential data. The respective prototypes are visualized based on filtering and highlighting the most relevant areas when predicting specific action units in human faces. A quantitative evaluation demonstrates that temporal prototypes effectively capture and aggregate temporal changes in action units for a given emotional state throughout a video sequence [13]. Again, we consider this method to be mostly relevant to experts who want to validate a model. Its modality is visual, therefore we place it in Fig. 3 at part (c), in the neighborhood of part (a). However, due to their temporal resolution, we consider temporal prototypes to be more expressive than single-image highlighting. As illustrated in part (c) of Fig. 3 we can observe a change over time in sequential data, e.g., facial expressions. In an aggregated view that helps especially experts to efficiently validate a model.

Mohammed et al. [35] propose a method to enhance the transparency of automatic tissue classification by leveraging the technique of Grad-CAM as introduced in [48]. In their work, Mohammed et al. go beyond the commonly used visualizations of the final layers of a CNN in order to identify the most relevant intermediate layers. The researchers collaborate with pathologists to visually assess the relevance of these layers, and they find that the intermediate layers are particularly important for capturing important morphological structures that are necessary for accurate class decisions. They provide a user-friendly interface that can be used by medical experts to validate any CNN and its layers. By offering visual explanations for intermediate layers and taking into account the expertise of pathologists, the authors claim to provide valuable insights into the decision-making process of neural networks in histopathological tissue

Table 1 Overview of selected explanation methods applied and developed during the research project *Transparent Medical Expert Companion*

Explanation method properties	Publication
Modality	
Visual (pixel importance)	[13, 30, 35, 36]
Verbal (relations)	[46]
Multimodal (visualizations and relations)	[16, 17]
Scope	
Local	[36]
Global	[30]
Mixed	[13, 16, 17, 35, 46]
Interaction	
Dialogue	[16, 17]
Correction	[44, 46]
Type of information aggregation	
Contrastive explanation	[14]
Prototype-based explanation	[13]
Cluster-based explanation	[13, 30]
Classifier base model	
Interpretable base model	[16, 17, 46]
Opaque base model with interpretable surrogate model	[46]
Opaque base model	[13, 30, 35, 36]

classification [35]. We place this method at part (d) of Fig. 3 as it provides more than a single-image visual explanation. Compared to the previously introduced temporal prototypes, it provides a spatial resolution based on the extraction of highlights from different intermediate layers of a CNN. It uncovers hierarchical and part-based compositions learned by a CNN. We consider it therefore as a method that offers more expressiveness than single-image explanations.

3.2 Expert-Based Correction and Verbal Explanations

The concept of mutual explanations for interactive machine learning is introduced and demonstrated in an article by Schmid and Finzel [46] as part of an application named *LearnWithME* for comprehensible digital pathology [7]. Their work is motivated by combining deep learning black-box approaches with interpretable machine learning for classifying the depth of tumor invasion in colon tissue. The idea behind their proposed approach is to combine the predictive accuracy of deep learning with the transparency and comprehensibility of interpretable models, in particular ILP [37, 38]. Specifically, the authors present an extension of the ILP framework *Aleph* [50] to enable interactive learning. Medical experts can ask for verbal explanations. They can correct labels of classified examples and, moreover, correct explanations. Expert knowledge is injected into the ILP model's learning process in the form of user-defined constraints for model adaptation [46]. We place this approach at the back of Fig. 3 at part (e) as it provides verbal explanations. We consider it more expressive than visualization techniques. This approach is obviously tailored to experts in a domain. Nevertheless, since verbal explanations can convey relational information, *LearnWithME* can be also helpful for novices who want to understand more complex relationships in classification outcomes.

Another approach exploits domain-knowledge in the form of correlations between ground truth labels of facial expressions in order to regularize CNN-based classifications of action units in video sequences [44]. This approach could be enhanced, for example, by visual explanations, like those introduced in [13], or verbal explanations for temporal relationships between facial expressions, similar to [14] to serve the information needs of domain experts.

3.3 Contrastive Explanations

In their work, Finzel et al. [14] present an approach for generating contrastive explanations to explain the classification of similar facial expressions, in particular, pain and disgust from video sequences. Two approaches are compared: one based on facial expression attributes and the other based on temporal relationships between facial expressions within a sequence. The input to the contrastive explanation method is the output of a rule-based ILP classifier for pain and disgust. Two similarity metrics are used to determine most similar and least similar contrasting instances (near misses versus far misses) based on the coverage of sample features with and without considering the coverage by learned rules. The results show that explanations for near misses are shorter than those for far misses, regardless of the similarity metric used [14]. We place this method at the back of Fig. 3 at part (f) as it provides contrastive verbal explanations. We consider it to be a little bit more expressive than verbal explanations generated for example by the previously introduced *LearnWithME*. Contrastive explanations shed light into the decision boundary of models and may therefore be of value for experts and novices.

3.4 Dialogue-Based Interaction and Multimodal Explanation

Based on empirical evidence suggesting that different types of explanations should be used to satisfy different information needs of users and to build trust in AI systems [51], Finzel et al. [17] implement multimodal explanations (visual and verbal) for decisions made by a machine learning model. They apply an approach to support medical diagnoses, that was first introduced on artificial data [16]. The method enables medical professionals and students to obtain verbal explanations for classifications through a dialogue, along with the ability to query the system and receive prototypical examples in the form of images depicting typical health conditions from digital pathology. The authors suggest that this approach can be used to validate algorithmic decisions by incorporating an expert-in-the-loop method, or for medical education purposes [17]. We consider this approach to be the most expressive within the selection of methods presented here. We place it rightmost in Fig. 3 at part (g). It encompasses multimodal explanations as well as scaffolding between global and local explanations in a drill-down manner [16, 17].

We want to point out that most of the work presented here provides repositories with respective code bases. If data could not be made available due to data protection, the necessary, licensed resources are either listed in the respective publications or small, simulated data sets are included in order to at least provide testable code.

4 Lessons Learned from Explainable Medical and Clinical Image Classification

This section outlines the lessons learned from the *Transparent Medical Expert Companion* project by discussing the methods introduced in the previous section.

4.1 There is a Lack of Meaning and Semantics in Visual Explanations

All methods presented in Sect. 3.1 are based on some form of visual highlighting of relevant pixel areas in input images to approximate the reasoning for the classification of a CNN. Although, domain experts can interpret these visual explanations, the meaning of highlights might not be unambiguous and rather dependent of whether the ground truth label of an image matches the label predicted by a CNN and whether contrasting classes share common properties or not. Especially, if explanations should also serve novices, it is necessary to assign some meaning to important pixels. Currently, novel approaches for concept-based explanations are being investigated. For example, there exists a concept-level extension to LRP [1, 2].

4.2 Assessing the Faithfulness of Models with the Help of Visual Explanations is Challenging

Another lesson learned from applying and evaluating the methods from Sect. 3.1 was that visual explanations may not always display relevance indicating faithfulness to the properties of classes (e.g., due to noise) without providing semantic grounding (see Sect. 4.1). Relevance is usually spread across all pixels in an image, where only a few are truly important for a classification outcome. Thus, it remains a challenge to assess the faithfulness of CNNs to the domain of interest. Therefore, more application-grounded evaluation techniques are needed to support experts and developers in the process of model improvement. For instance, a technique developed during the project computes the aggregation of pixel relevance inside the boundaries of polygonal areas overlayed on human faces in order to evaluate facial expression classification. The goal is to approximate the faithfulness of a CNN to domain-knowledge in the form of facial expressions which get localized via landmarks and for which the relevance assigned to respective pixel areas is put into ratio with the total relevance in an image [15].

4.3 Constraints and Their Quality Matter

The main lesson learned from the methods presented in Sect. 3.2 is that corrective feedback by humans is beneficial for improving models in certain cases, although probably limited in number. We found that constraining a model is especially helpful if a data set contains noise and the noise can be filtered by a constraint all at once [46]. This reduces the corrective effort for experts. Nevertheless, we would like to point out that corrective feedback may harm a model if the corrective decision is biased, uncertain, or contradicting parts of a model that are necessary to correctly detect other classes or sub-classes of instances. More general approaches as the one presented in [44] are a promising counter measure against locally false corrections, since they consider correlations between classes in the entire ground truth, however, as described before, especially medical and clinical data may not be free from erroneous labels, noise or sparse features, which could render correlations spurious.

Even if the application of constraints may not directly contribute to improving a model, constraints are still helpful in exploring and debugging the predictive behavior of a model and interactively examining the quality of the generated explanations and the input data [46]. For this purpose, for example a combination of injecting domain-specific local constraints and domain-independent global constraints, hard constraints as well as soft constraints into a model, could support the automated quantification of a model's faithfulness and uncertainty for experts or help with exploring alternative explanations for novices.

4.4 There is a Trade-Off Between Generating and Selecting Contrastive Samples for Explanation

Contrastive explanation methods like the one introduced in Sect. 3.3 seem especially helpful for experts as well as novices, as they provide the means to explore a decision boundary of a model. A general lesson learned here is the observation that for rather sparse data sets it might not be possible to *generate* contrastive explanations like near misses from the data distribution, since such a process may yield unrealistic samples. The approach, presented in [14] therefore *selects* samples from the given data set based on a similarity metric. In borderline cases, where instances from a data set do not have many commonalities, this may mean that produced explanations will not be minimal. For very similar classes, this may happen, however, probably in a neglectable amount of cases.

4.5 There is a Need for Interfaces that Bridge the Semantic Gap Between Different Modalities

Complementing visual explanations for image classification with appropriate verbal explanations produced by an interpretable surrogate model in order to express more complex relationships may not always be straightforward to do. It may be necessary to first match the content of and concepts behind a visualization with the message that is transferred by a verbal explanation. Neuro-symbolic approaches to image classification that integrate knowledge into the process of learning and explaining may provide a solution to identifying concepts and knowledgably putting them into relation.

4.6 Interpretable Models are not Easy to Understand by Default

For all methods described in Sect. 3.4 the main finding was that not only deep learning models are complex. Also interpretable (surrogate) models may not be explainable at once due to complex relations among individual concepts and rich conceptual hierarchies [16]. Scaffolding of reasons for a classification outcome, e.g., through a dialogue-based interaction between the system and the human user may help navigating the different decision paths and strategies of a model.

4.7 We Need More Integrative Explanation Frameworks

Ultimately, all the previous lessons learned illustrate the need for an integrative approach that provides various explanation strategies and flexible interaction interfaces for experts, novices and developers. This means that explainees should be able to switch between local and global explanations, choose between different domain-relevant explanation modalities and, if necessary, use constraints to explore the predictive behavior of a model and corresponding explanations according to their individual information need. The open challenge is to realise such a dynamic, integrative explanation framework that puts the human and the learning system into a beneficial explanatory dialogue.

5 Open Challenges and Future Directions

Open challenges concern mainly the lack of high quality data from the real world and in benchmarks that allow to assess the performance and quality of models and explanations in high-stakes applications such as medicine and clinical decision making. Furthermore it remains an open question how the quality of explanations and the faithfulness of models could be assessed and ensured for complex image classification models like CNNs. The most promising future directions that we currently see, is XAI evaluation from less technical perspectives including accuracy measures and pixel importance quantification toward more human-centered, concept-based as well as application-grounded metrics. This includes constructing and collecting appropriate benchmarks as well as developing valid, reliable and objective evaluation metrics. Deriving empirical evidence for explanation quality and beneficial interaction between AI and human experts and novices will pave the way for more integrative, dialogue-based solutions that allow learning and exchange in two directions: from the system to the human and back.

6 Conclusions

We presented and categorized explanation methods that have been developed and applied for image classification in digital pathology and affective computing during an interdisciplinary research project, involving medical and clinical experts. In particular, we discussed the expressiveness of visual, verbal and multimodal explanations differing in scope and their suitability to serve the information need of experts and novices in an application field. Our findings suggest that solely highlighting features in the input space to assess an image classifier's performance or reasons behind a decision may not be enough. We presented methods that extend unimodal visual explanations, e.g., by verbal explanations and dialogue-based interaction between a human and a classification model. We identified that prototypes may be especially beneficial as explanation methods when a global view is not feasible due to limited generalizability in the domain of interest. We present contrastive explanations, e.g., based on near misses, as suitable for cases, where instances are close to the decision boundary of a classifier. Finally, we conclude that multimodal explanations are helpful if various perspectives are needed for solving the task at hand, e.g., validating a model as an expert or learning from a model as a novice. The goal of this article was to report our lessons learned from designing explanations such that they support the understanding of model decisions from a human perspective. Our findings include in particular that there is a need for more semantically enriched, explorative, interactive and especially integrative explanatory approaches that provide explanation as a process in which different explainees may satisfy their information need in accordance to their current level of knowledge and understanding.

We see a research gap in the lack of human-centered and application-grounded benchmarks and metrics for explanation generation and evaluation as well as in empirical evidence that demonstrates the usefulness of explanations for experts and novices in medicine and clinical decision making. We believe that providing more human-centered and integrative explanation frameworks will pave the way to beneficial AI transparency and human understanding of and trust in AI.

Acknowledgements This work was partially funded by the German Federal Ministry of Education and Research under Grant FKZ 01IS18056 B, BMBF ML-3 (TraMeExCo, 2018-2021) and by the German Research Foundation under Grant DFG 405630557 (PainFaceReader, 2021-2023).

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability Data and code produced during the presented research project is made available in the referenced publications, if applicable.

Declaration

Conflict of interest The author declares that she has no conflict of interest.

References

1. Achtibat R, Dreyer M, Eisenbraun I, Bosse S, Wiegand T, Samek W, Lapuschkin S (2022) From "where" to "what": towards human-understandable explanations through concept relevance propagation. CoRR. https://doi.org/10.48550/arXiv.2206.03208. arXiv:2206.03208
2. Achtibat R, Dreyer M, Eisenbraun I, Bosse S, Wiegand T, Samek W, Lapuschkin S (2023) From attribution maps to human-understandable explanations through concept relevance propagation. Nat Mach Intell 5(9):1006–1019
3. Adadi A, Berrada M (2020) Explainable ai for healthcare: from black box to interpretable models. In: Bhateja V, Satapathy SC, Satori H (eds) Embedded systems and artificial intelligence. Springer Singapore, Singapore, pp 327–337
4. Albahri A, Duhaim AM, Fadhel MA, Alnoor A, Baqer NS, Alzubaidi L, Albahri O, Alamoodi A, Bai J, Salhi A, Santamaría J, Ouyang C, Gupta A, Gu Y, Deveci M (2023) A systematic review of trustworthy and explainable artificial intelligence in healthcare: assessment of quality, bias risk, and data fusion. Inf Fusion 96:156–191. https://doi.org/10.1016/j.inffus.2023.03.008
5. Bratko I, Muggleton SH (1995) Applications of inductive logic programming. Commun ACM 38(11):65–70. https://doi.org/10.1145/219717.219771
6. Bromme R, Rambow R (2000) Experten-Laien-Kommunikation als Gegenstand der Expertiseforschung: Für eine Erweiterung des psychologischen Bildes vom Experten. Psychologie 2000. Bericht über den 42. Kongress der Deutschen Gesellschaft für Psychologie in Jena 2000, Pabst Science Publishers
7. Bruckert S, Finzel B, Schmid U (2020) The next generation of medical decision support: a roadmap toward transparent expert companions. Front Artif Intell 3:507973. https://doi.org/10.3389/frai.2020.507973
8. Bychkov A, Schubert M (2023) Constant demand, patchy supply. https://thepathologist.com/outside-the-lab/constant-demand-patchy-supply
9. Chaddad A, Peng J, Xu J, Bouridane A (2023) Survey of explainable ai techniques in healthcare. Sensors 23(2). https://doi.org/10.3390/s23020634. https://www.mdpi.com/1424-8220/23/2/634
10. Chen H, Gomez C, Huang CM, Unberath M (2022) Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. NPJ Digit Med 5(1):156
11. Chi MT (1996) Constructing self-explanations and scaffolded explanations in tutoring. Appl Cogn Psychol 10(7):33–49
12. Dhar T, Dey N, Borra S, Sherratt RS (2023) Challenges of deep learning in medical image analysis-improving explainability and trust. IEEE Trans Technol Soc 4(1):68–75. https://doi.org/10.1109/TTS.2023.3234203
13. Finzel B, Kollmann R, Rieger I, Pahl J, Schmid U (2021) Deriving temporal prototypes from saliency map clusters for the analysis of deep-learning-based facial action unit classification. In: Seidl T, Fromm M, Obermeier S (eds) Proceedings of the LWDA 2021 workshops: FGWM, KDML, FGWI-BIA, and FGIR, Online, September 1–3, 2021, CEUR Workshop Proceedings, vol 2993, pp 86–97. CEUR-WS.org. https://ceur-ws.org/Vol-2993/paper-09.pdf
14. Finzel B, Kuhn PS, Tafler ED, Schmid U (2022) Explaining with attribute-based and relational near misses: an interpretable approach to distinguishing facial expressions of pain and disgust. In: Inductive logic programming: 31th international conference, ILP 2022, Cumberland Lodge, Windsor Great Park, United Kingdom, September 28–30, 2022, proceedings, vol 31. Springer, pp 1–12
15. Finzel B, Rieger I, Kuhn S, Schmid U (2023) Domain-specific evaluation of visual explanations for application-grounded facial expression recognition. In: Holzinger A, Kieseberg P, Cabitza F, Campagner A, Tjoa AM, Weippl E (eds) Machine learning and knowledge extraction. Springer Nature Switzerland, Cham, pp 31–44
16. Finzel B, Tafler DE, Scheele S, Schmid U (2021) Explanation as a process: user-centric construction of multi-level and multi-modal explanations. In: Edelkamp S, Möller R, Rueckert E (eds) KI 2021: advances in artificial intelligence—44th German conference on AI, virtual event, September 27–October 1, 2021, proceedings, lecture notes in computer science, vol 12873. Springer, pp 80–94. https://doi.org/10.1007/978-3-030-87626-5_7
17. Finzel B, Tafler DE, Thaler AM, Schmid U (2021) Multimodal explanations for user-centric medical decision support systems. In: Doyle TE, Kelliher A, Samavi R, Barry B, Yule SJ, Parker S, Noseworthy MD, Yang Q (eds) Proceedings of the AAAI 2021 fall symposium on human partnership with medical AI: design, operationalization, and ethics (AAAI-HUMAN 2021), virtual event, November 4–6, 2021, CEUR Workshop Proceedings, vol 3068. CEUR-WS.org. https://ceur-ws.org/Vol-3068/short2.pdf
18. Ghassemi M, Oakden-Rayner L, Beam AL (2021) The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 3(11):e745–e750
19. Graesser AC, McNamara DS, VanLehn K (2005) Scaffolding deep comprehension strategies through point &query, autotutor, and istart. Educ Psychol 40(4):225–234
20. Gruber H, Ziegler A (1996) Expertiseforschung. Theoretische und methodische Grundlagen, Opladen
21. Hägele M, Seegerer P, Lapuschkin S, Bockmayr M, Samek W, Klauschen F, Müller KR, Binder A (2020) Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. Sci Rep 10(1):1–12
22. Hassan T, Seuß D, Wollenberg J, Weitz K, Kunz M, Lautenbacher S, Garbas J, Schmid U (2021) Automatic detection of pain from facial expressions: a survey. IEEE Trans Pattern Anal Mach Intell 43(6):1815–1831. https://doi.org/10.1109/TPAMI.2019.2958341
23. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K (2019) The practical implementation of artificial intelligence technologies in medicine. Nat Med 25(1):30–36
24. Hernández-Orallo J (2019) Gazing into clever Hans machines. Nat Mach Intell 1(4):172–173. https://doi.org/10.1038/S42256-019-0032-5

25. Holdstock J, Mayes A, Gong Q, Roberts N, Kapur N (2005) Item recognition is less impaired than recall and associative recognition in a patient with selective hippocampal damage. Hippocampus 15(2):203–215. https://doi.org/10.1002/hipo.20046
26. Holzinger A (2016) Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inform 3(2):119–131
27. Katz D, Rhee L, Katz C, Aronson D, Frank G, Gardner C, Willett W, Dansinger M (2020) Dietary assessment can be based on pattern recognition rather than recall. Med Hypotheses 140:109644. https://doi.org/10.1016/j.mehy.2020.109644
28. Kulesza T, Stumpf S, Burnett MM, Yang S, Kwan I, Wong W (2013) Too much, too little, or just right? Ways explanations impact end users' mental models. In: Kelleher C, Burnett MM, Sauer S (eds) 2013 IEEE symposium on visual languages and human centric computing, San Jose, CA, USA, September 15–19, 2013, pp 3–10. IEEE Computer Society. https://doi.org/10.1109/VLHCC.2013.6645235
29. Kunz M, Peter J, Huster S, Lautenbacher S (2013) Pain and disgust: the facial signaling of two aversive bodily experiences. PLoS ONE 8(12):e83277
30. Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller KR (2019) Unmasking clever Hans predictors and assessing what machines really learn. Nat Commun 10(1):1096
31. Lautenbacher S, Hassan T, Seuss D, Loy FW, Garbas JU, Schmid U, Kunz M et al (2022) Automatic coding of facial expressions of pain: are we there yet? Pain Res Manag 2022
32. Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR (2022) Application of explainable artificial intelligence for healthcare: a systematic review of the last decade (2011–2022). Comput Methods Progr Biomed 226:107161. https://doi.org/10.1016/j.cmpb.2022.107161
33. McNeill KL, Lizotte DJ, Krajcik J, Marx RW (2006) Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. J Learn Sci 15(2):153–191
34. Miller T (2017) Explanation in artificial intelligence: insights from the social sciences. arXiv:1706.07269 [cs]
35. Mohammed A, Geppert C, Hartmann A, Kuritcyn P, Bruns V, Schmid U, Wittenberg T, Benz M, Finzel B (2022) Explaining and evaluating deep tissue classification by visualizing activations of most relevant intermediate layers. Curr Dir Biomed Eng 8(2):229–232. https://doi.org/10.1515/cdbme-2022-1059
36. Montavon G, Binder A, Lapuschkin S, Samek W, Müller K (2019) Layer-wise relevance propagation: an overview. In: Samek W, Montavon G, Vedaldi A, Hansen LK, Müller K (eds) Explainable AI: interpreting, explaining and visualizing deep learning, lecture notes in computer science, vol 11700. Springer, pp 193–209. https://doi.org/10.1007/978-3-030-28954-6_10
37. Muggleton SH (1991) Inductive logic programming. New Gener Comput 8(4):295–318. https://doi.org/10.1007/BF03037089
38. Muggleton SH, Schmid U, Zeller C, Tamaddoni-Nezhad A, Besold T (2018) Ultra-strong machine learning: comprehensibility of programs learned with ILP. Mach Learn 107(7):1119–1140
39. Páez A (2019) The pragmatic turn in explainable artificial intelligence (XAI). Minds Mach 29(3):441–459. https://doi.org/10.1007/s11023-019-09502-w
40. Picard RW (1997) Affective computing. MIT Press, Cambridge
41. Prabhu V, Kannan A, Ravuri M, Chaplain M, Sontag D, Amatriain X (2019) Few-shot learning for dermatological disease diagnosis. In: Machine learning for healthcare conference. PMLR, pp 532–552
42. Rabold J, Siebers M, Schmid U (2018) Explaining black-box classifiers with ILP-empowering LIME with aleph to approximate non-linear decisions with relational rules. In: Riguzzi F, Bellodi E, Zese R (eds) Inductive logic programming—28th international conference, ILP 2018, Ferrara, Italy, September 2–4, 2018, proceedings, lecture notes in computer science, vol 11105. Springer, pp 105–117. https://doi.org/10.1007/978-3-319-99960-9_7
43. Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?": explaining the predictions of any classifier. In: Krishnapuram B, Shah M, Smola AJ, Aggarwal CC, Shen D, Rastogi R (eds) Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco, CA, USA, August 13–17, 2016. ACM, pp 1135–1144. https://doi.org/10.1145/2939672.2939778
44. Rieger I, Pahl J, Finzel B, Schmid U (2022) CorrLoss: integrating co-occurrence domain knowledge for affect recognition. In: 26th international conference on pattern recognition, ICPR 2022, Montreal, QC, Canada, August 21–25, 2022. IEEE, pp 798–804. https://doi.org/10.1109/ICPR56361.2022.9956319
45. Sayer A (1982) Explanation in economic geography: abstraction versus generalization. Prog Hum Geogr 6(1):68–88
46. Schmid U, Finzel B (2020) Mutual explanations for cooperative decision making in medicine. Künstliche Intell 34(2):227–233. https://doi.org/10.1007/s13218-020-00633-2
47. Schwalbe G, Finzel B (2023) A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. Data Min Knowl Discov 1–59
48. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2020) Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis 128(2):336–359. https://doi.org/10.1007/s11263-019-01228-7
49. Sokol K, Flach PA (2020) One explanation does not fit all. Künstliche Intell 34(2):235–250. https://doi.org/10.1007/s13218-020-00637-y
50. Srinivasan A (2007) The Aleph manual. https://www.cs.ox.ac.uk/activities/programinduction/Aleph/aleph.html
51. Thaler AM, Schmid U (2021) Explaining machine learned relational concepts in visual domains-effects of perceived accuracy on joint performance and trust. In: Proceedings of the annual meeting of the cognitive science society, vol 43
52. Tizhoosh HR, Pantanowitz L (2018) Artificial intelligence and digital pathology: challenges and opportunities. J Pathol Inform 9
53. Uner O, Roediger Henry LI (2022) Do recall and recognition lead to different retrieval experiences? Am J Psychol 135(1):33–43. https://doi.org/10.5406/19398298.135.1.03
54. Vellido A (2020) The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Comput Appl 32(24):18069–18083
55. Wang Y, Song W, Tao W, Liotta A, Yang D, Li X, Gao S, Sun Y, Ge W, Zhang W, Zhang W (2022) A systematic review on affective computing: emotion models, databases, and recent advances. Inf Fusion 83–84:19–52. https://doi.org/10.1016/j.inffus.2022.03.009
56. Wittekind C, Bootz F, Meyer HJ (2004) Tumoren des Verdauungstraktes. In: Wittekind C, Bootz F, Meyer HJ (eds) TNM Klassifikation maligner Tumoren, International Union against cancer. Springer, pp 53–88
57. Yang G, Ye Q, Xia J (2022) Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: a mini-review, two showcases and beyond. Inf Fusion 77:29–52. https://doi.org/10.1016/j.inffus.2021.07.016