Knowledge-Based Systems 283 (2024) 111107

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A unified and practical user-centric framework for explainable artificial intelligence

Sinan Kaplan [a,*], Hannu Uusitalo [b,c], Lasse Lensu [a]

[a] Department of Computational Engineering, LUT School of Engineering Sciences, Lappeenranta-Lahti University of Technology LUT, Yliopistonkatu 34, Lappeenranta, 53850, Finland
[b] Department of Ophthalmology, Faculty of Medicine and Health Technology, Tampere University, PO Box 100, Tampere, 33014, Finland
[c] Tays Eye Centre, Tampere University Hospital, Tampere, 33014, Finland

ARTICLE INFO

Keywords:
Artificial intelligence
Explainable artificial intelligence
Optical coherence tomography
Retinal image analysis

ABSTRACT

Adoption of artificial intelligence (AI) is causing a paradigm change in many fields. Its practical utilization, however, especially in safety-critical applications like medicine, remains limited, mainly due to the black-box nature of most advanced AI models, which creates difficulties understanding why and how a model produces a particular output or decision. To overcome this issue, various methods and techniques have been proposed within the emerging field of explainable artificial intelligence (XAI). In this paper, we introduce a user-centric and interactive framework that enables a holistic understanding of AI systems. The proposed framework is designed to aid the development of more explainable AI systems by promoting transparency and trust in their use and allow different stakeholders to better understand and evaluate AI decisions. To illustrate the effectiveness of the framework, we implement a case study of an AI system analyzing optical coherence tomography (OCT) images. The development of this example case is reported using the proposed framework.

1. Introduction

During the last past decade, there has been a significant growth in the number and usage of AI applications [1,2] and AI models have shown their effectiveness in many tasks, including image classification [2], product recommendation [3,4], object detection and recognition [5], in areas ranging from medicine to machine translation. For instance, applications of AI in retinal imaging [6] have shown considerable success in the early detection and recognition from retinal images of medical abnormalities such as diabetic retinopathy and age-related macular degeneration.

Despite its clear potential and a number of successful applications also in safety-critical tasks [7–9] like medical diagnosis, financial decision systems and self-driving cars, practical uses of AI in such tasks are currently limited [10]. The underlying reasons for the limited use of AI are the computational cost required to run large AI models in production and the black-box nature of AI models. The former issue seems to be being resolved with advances in hardware technology [1]; the latter issue, however, remains critical.

The term *black-box* refers to a concept or system in which there are measurable inputs and outputs, but the internal workings of the system remain unknowable, and users are thus unable to understand how and based on what grounds an AI algorithm makes a particular decision [11,12]. This inability to comprehend the workings of black-box AI is because the structure of AI models is complex, non-linear, and hard to interpret, and the amount of data used to train the models is typically large. The opaque nature of AI decision-making raises questions of safety, trustworthiness and transparency and hinders extensive use of AI. For instance, medical decision support systems need to be trustworthy and transparent, and the black-box nature of AI algorithms thus remains a bottleneck to widespread adaptation of AI in clinical use.

To address issues originating from the black-box nature of AI models, several solutions have been proposed under an emerging field called XAI [13,14]. XAI presents a set of techniques and methods to understand *why* AI makes a particular decision. Knowledge of the reasons for a decision is considered key to understanding model decisions and improving users' trust in AI-based solutions.

In the XAI literature, methods to explain AI models have mainly focused on post-hoc understanding [15]. For instance, typical XAI methods may try to explain each decision by an AI model trained to recognize objects of interest in images by highlighting relevant regions

[*] Corresponding author.
E-mail addresses: sinan.kaplan@student.lut.fi (S. Kaplan), hannu.uusitalo@tuni.fi (H. Uusitalo), lasse.lensu@lut.fi (L. Lensu).

https://doi.org/10.1016/j.knosys.2023.111107
Received 18 May 2023; Received in revised form 17 September 2023; Accepted 19 October 2023
Available online 31 October 2023

on the input image [16]. Another category of XAI models attempts to explain the AI model at a general level by giving information on how the model works [17,18]. Such XAI methods are primarily model-centric, meaning that they are designed for model audit and mainly targeted at model developers.

In addition to model-centric approaches, methods also exist that provide a holistic understanding of an AI system by focusing on specific building blocks such as data used and model trained. For example, one such approach involves the use of datasheets [19] and datacards [20] that provide detailed information about the data used to train the AI model, and another approach uses model cards [21] that provide an overview of the AI model itself. While these proposals offer valuable insights into specific parts of the AI system, they have important limitations. Firstly, they are restricted in their scope and do not take into account the diverse needs of different users of the AI systems. Additionally, they lack interactivity, which is crucial to allow users to fully explore and understand the system as it enables users to actively engage with and test the system's functionality and capabilities. Lastly, such methods are designed for a specific part of the AI system rather than providing a comprehensive understanding of the entire system. To fully understand the functionality of an AI system, it is essential to understand all its building blocks. International standards published by International Organization for Standardization (ISO), for example, promote transparency, ethics and trust related to AI systems by encouraging organizations to integrate transparency into all levels of the AI development cycle to foster trustworthiness [22].

To enable a comprehensive understanding of the entire process of an AI application from the collection of data and the development of models to the evaluation protocols and decision explorations, we propose a simple yet effective framework for explaining AI-based solutions. This framework incorporates user-centric design principles and enables interaction between users and the decisions made by AI models, thus allowing greater understanding and enabling debugging of AI models. Additionally, by embracing this framework, it is possible to introduce transparency already in the design of AI systems at the system level. This enables alignment with globally recognized standards for trustworthy AI, such as those established by ISO. Therefore, we consider the proposed framework a potentially valuable tool for promoting transparency and trust in the usage of AI systems. The framework provides a pathway towards the development of more ethical, explainable and trustworthy AI systems and allow stakeholders to understand better and evaluate the decision-making process of AI models.

2. Motivation

A typical AI model development includes three main stages [23–26]: pre-modeling, modeling, and post-modeling. The pre-modeling part involves mainly model search and data exploration; the modeling stage deals with data processing and model development in the form of training; and the post-modeling part is concerned with model testing, interpretation of model output and model deployment. This cycle is iterative and incremental.

In practical applications, it is important to understand every stage of AI application development from pre-modeling to post-modeling. In this way, it is easier to debug and isolate any part of the AI system should any issues arise. Consequently, it is critical to have explanations and details regarding the different stages: data, model, evaluation and post-hoc decisions.

In the XAI literature, more attention has generally been given to post-modeling explanations [27,28]. In an ideal situation, however, explanations should be generated for every stage of AI system development. Therefore, there exists a need for a standardized and generic framework that is able to highlight and present explanations for all the different stages of AI system development. This is also well stated in the guidelines proposed by ISO for trustworthy AI [22,26]. For instance, in [22] several critical aspects of AI systems are stated to be addressed while offering explanations: "(1) Information about algorithms, training data and user data, including how it was collected. (2) Disclosing the evaluation methods and metrics used to validate how a system works. (3) Explaining to stakeholders the inputs that were used to reach a decision. (4) Explaining to stakeholders, as much as is possible, how an AI system arrived at a decision. (5) Notifying stakeholders when a decision about them is made by an AI system. (6) Notifying stakeholders when they are interacting with an AI system. (7) Consider allowing stakeholders to submit test cases to see how the AI system and application reacts to different situations".

In a holistic framework, the explanations need to be generated in a user-centered fashion because different stakeholders are involved in each stage of the development of AI/ML applications [26,29–31]. Thus, the different stakeholders require explanations tailored to their needs to be able to understand the problem and solution. For instance, while a model developer might be interested in how a model is trained and which architecture is used, a product manager might be interested in what kind of data is used and how it is used. Policymakers, on the other hand, might wish to check any licensing and ethical considerations arising from use of the AI model and its data, and practitioners might be interested in understanding model decisions and may want to perform sensitivity analysis of the model. The different roles and responsibilities mean that user-centered explanations are crucial to support the needs of all stakeholders.

To address the above issues, we propose an interactive and user-centric [32,33] framework for AI applications. The framework enables transparency and understanding of AI systems, including the AI model, data used, performance evaluation schemes, and post-hoc decisions made by the AI model, via a unified simple framework. It also promotes alignment with the aforementioned ISO standards, thereby advancing the objective of fostering transparency throughout all stages of AI development. The framework is mainly intended for practical applications, however, it can also be used to report and present explanations for any AI system development from pre-modeling to post-modeling while taking into account different user groups.

3. Methodology

To be able to design a practical user-centric framework for interpretable and explainable AI system documentation, we first conducted a literature review with the aim of identifying the key elements and questions in the field of XAI. The literature review is based on peer-reviewed articles discussing and reviewing explainable AI, articles describing and reviewing the taxonomy of XAI methods, articles discussing human-centered design principles/approaches, and studies reviewing practical usage of XAI methods. Based on the literature review, we identify features/attributes of explainability that enable conclusions to be drawn and then propose a framework which is explainable by design. Finally, we use the framework to present a use case (application) in the medical domain. Within the context of this study, the use case is an implementation of a concrete application studied to validate the proposed framework.

Literature review for designing the framework. We constructed the user-centered [34–37] framework based on a literature review of XAI. It should be noted that defining the taxonomy, concepts, and methods of XAI [11,17,18,28,38] is out of the scope of this paper. Readers are referred to [17,27,28] for discussion of the theoretical foundations of XAI.

Before commencing the framework design, we defined five key questions for clarification: (1) Why is there a need for explainability? (2) What can be explained in AI systems? (3) What are the attributes of a good explanation in XAI? (4) For whom are the explanations in XAI generated? (Who are the users/stakeholders involved?) And finally, (5) how are explanations for AI constructed?

Case Study. To showcase the effectiveness and usefulness of the framework and give an example of a practical application, we examined a medical image analysis case. First, we trained an AI model to detect certain anomalies from medical images, and then the proposed framework was used to report development phases and results with explanations.

4. Framework design

This section presents the proposed framework for XAI. We start by discussing the concept of explainability, which is a central idea in AI research, and how it is defined in XAI studies. Next, we identify and address four key questions that guide the framework design: (1) why explain, (2) what to explain, (3) for whom to explain, and (4) how to explain. The answers to these questions provide the foundation for the framework design, which focuses on the question of how to explain the AI model and the model output. The framework is designed to be flexible and adaptable to different AI applications and scenarios while also addressing the challenges and limitations of explainable AI.

Explainability in AI. Before making any claims about explainability in AI, it is worthwhile reviewing what explainability means and how it is translated to AI as XAI. The concept of explainability has its roots in philosophy, and many different theories of explanation are discussed in the literature [39,40]. Since explanations are predominantly defined in the context of a specific domain, multiple definitions of explainability can be found in the XAI community. For instance, while explainability is defined in [41] as "the capacity to provide or bring out the meaning of an abstract concept and understandability as the capacity to make the model understandable by end-users", it is defined in [30] as "numerous ways of exchanging information about a phenomenon, in this case, the functionality of a model or the rationale and criteria for a decision, to different stakeholders". Furthermore, in [42] the term "to explain" is defined as "a process which is substantially akin to the process of describing something, of giving some information about a portion of reality or of a theoretical framework". Additionally, there are arguments stating that explainability in AI should be defined from the perspective of social sciences. For instance, Miller et al. [43] define explainability as "the degree to which a human can understand the cause of a decision".

There is no standard way of how to define explainability in XAI. Therefore, in this paper we propose the following definition inspired by the diverse definitions found in the XAI literature [11,17,18,28] and also the definition of explainable AI by ISO [26] (in which explanation is defined as making an AI system understandable to different stakeholders/users beyond providing post-hoc explanations): an AI-based system is explainable if it can be described, understood and deconstructed to its main components by design while considering the needs of different user groups interacting with the system. Consequently, the system should be designed in a way that allows its users to understand how it works, why it works, and why it fails. Thus, explainability in AI should aim to present AI solutions and their components in an understandable format. Such explanations of solutions and components will then allow users to understand the behavior of the AI system and how it works at every stage of AI system development. To achieve this aim, transparency must be built into every stage of AI development, from pre-modeling to the post-modeling stages, and this transparency can help to build trust and confidence in AI.

Why explain AI. The goal of XAI is to make AI systems transparent, trustworthy, ethical, safe, secure and robust by allowing users to understand why AI works, how it works and why it reaches the outcomes that it does. Consequently, the explanations should be aimed at [33,44] (Fig. 1): increasing transparency [44], understanding decisions made by AI models [34], enabling exploration [45], understanding and inspecting data [11,46], reproducibility [45], justifying decisions, improving and inspecting models [45], ensuring compliance with regulations [30, 44], understanding the problem context [47], improving user's mental models [34], and ensuring trust in the system, and enabling performance verification [44,45]. In particular, enabling justification on what grounds an AI system makes a decision improves trust towards the system improving trustworthiness [48].

What to explain. Certain attributes of an AI system need to be explained to ensure that the system is fully explainable. In the XAI literature, these attributes to be explained are defined as the AI system in overview [29,49], decisions made by any AI model (global and local) [11,18], performance evaluation [41], details of AI model functionality and how the model works [34], data (summary, stats, issues) [46,50], and ethical and legal considerations [51–53]. Knowledge of these attributes is important for understanding of the workings of an AI system and ensuring its transparency and accountability. The attributes to be explained are illustrated in Fig. 2.

For whom to explain (Users). Different stakeholders require different explanations about an AI system [33,37] and these explanations are triggered by specific user requests. Therefore, it is important to understand who the users of such systems are and why they need explanations in the first place [54]. To do so, one can embrace user-centric approaches to generate explanations for specific user groups. XAI literature has focused on algorithms/methods for explainability [35] but has not given much attention to user-centric approaches.

Most studies consider the issue of explaining model decisions from the perspective of explaining decisions to domain experts or developers who design and implement AI models, and user-centric approaches have drawn little attention from the community. Nevertheless, a small number of studies [32–35,42,54] can be found that highlight the importance of user-centric approaches in XAI. A common point highlighted in these studies is that user-centric approaches can provide more personalized and targeted explanations than conventional approaches and are more useful for a wider range of stakeholders. Additionally, user-centric approaches can increase the trust and transparency of AI systems, leading to more effective adoption and deployment. Since user-centric approaches are still at an early stage of development, there is no clearly defined set of stakeholders. Hence, by considering the aforementioned studies, we construct profiles of initial user groups in XAI based on their roles and their expectations [42] from the system, see Fig. 3. In the next section, we consider these user groups when answering the question of how to explain and explore possible options to define user groups in XAI.

What makes a good explanation: Attributes of a good explanation. There is a lot of discussion in the literature about what makes a good explanation in the context of XAI, and what the attributes of such an explanation should be [31,33,35,38,41,55]. One characteristic of a good explanation is being experiential, meaning that it should make the decision understandable and enable decision manipulation for robustness tests [26]. It should also be explanatory, meaning that it should be accompanied by other things such as instructions, tutorials, etc. In addition, two types of uncertainty appear in every AI system: epistemic uncertainty and aleatoric uncertainty [56,57], and these uncertainties need to be addressed in explanations. Epistemic uncertainty arises due to a lack of data and can be reduced by providing more data and details about the data used to train the AI model [58]. Users can thus make more informed decisions. Aleatoric uncertainty, on the other hand, exhibits in the data as noise and it cannot be reduced. In addition to the data, a good explanation should help understand the behavior of the model. The attributes of a good explanation of an AI system are summarized in Fig. 4

4.1. How to explain: Proposed framework in detail

Frameworks for unified reporting, such as data sheets, model cards and data boards, can be found in the XAI literature. These frameworks are designed for specific parts of the AI development cycle. For instance, datasheet [19], datacard [20], FactSheets [59] and Data Nutrition Label [60] are used for data monitoring, while model cards [21, 36] are used mainly for model experiment reporting.

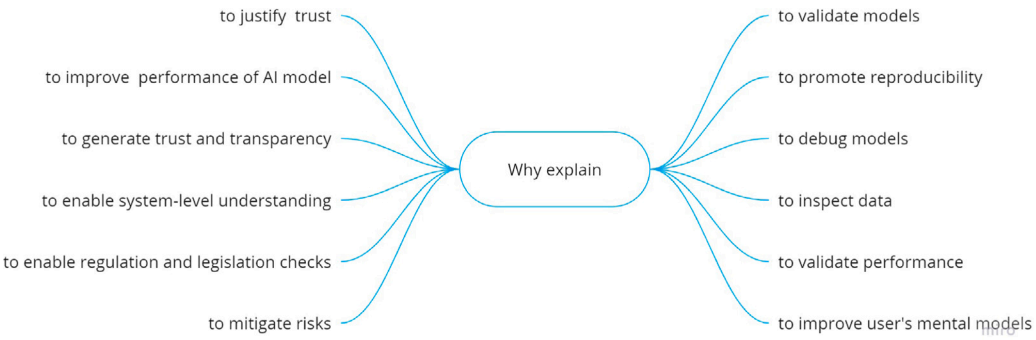

Why explain
to justify trust
to improve performance of AI model
to generate trust and transparency
to enable system-level understanding
to enable regulation and legislation checks
to mitigate risks
to validate models
to promote reproducibility
to debug models
to inspect data
to validate performance
to improve user's mental models
miro

Fig. 1. Overview of reasons for explaining an AI system.

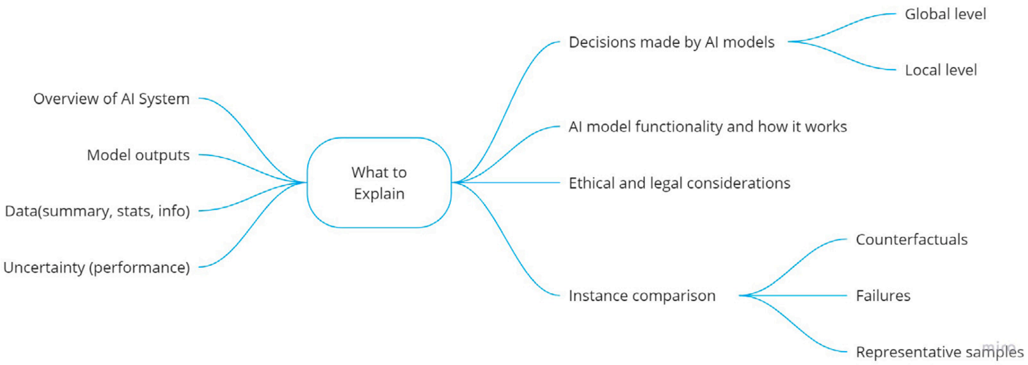

What to Explain
Overview of AI System
Model outputs
Data(summary, stats, info)
Uncertainty (performance)
Decisions made by AI models
Global level
Local level
AI model functionality and how it works
Ethical and legal considerations
Instance comparison
Counterfactuals
Failures
Representative samples
miro

Fig. 2. Overview of the aspects or components of an AI system that can be explained.

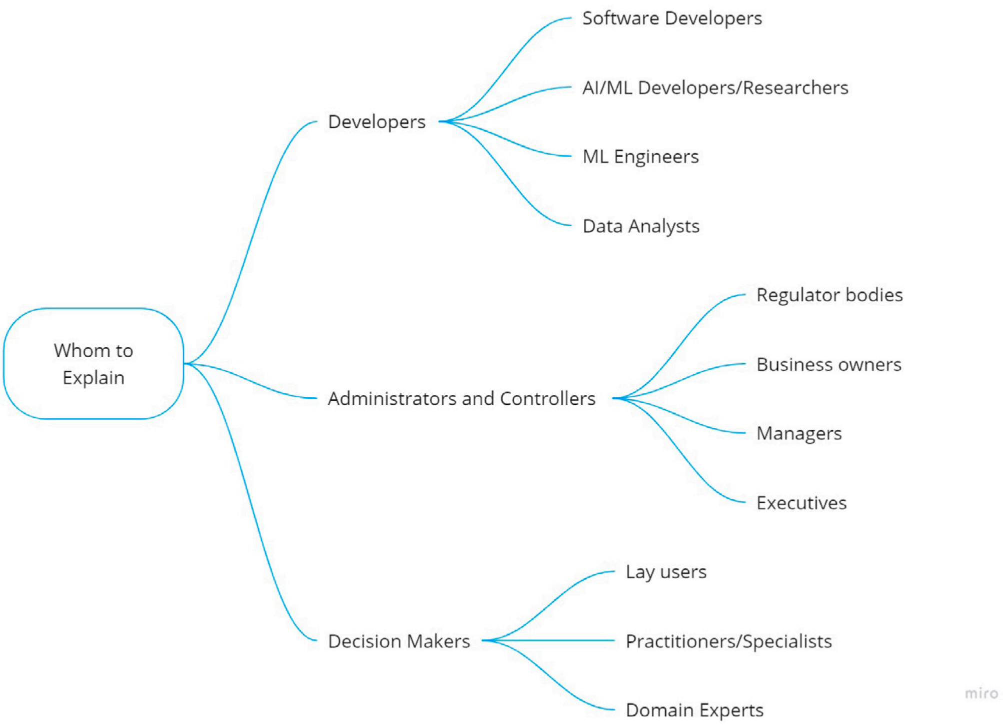


Fig. 3. Target user groups categorized in terms of the group characteristics and requirements from an AI system.

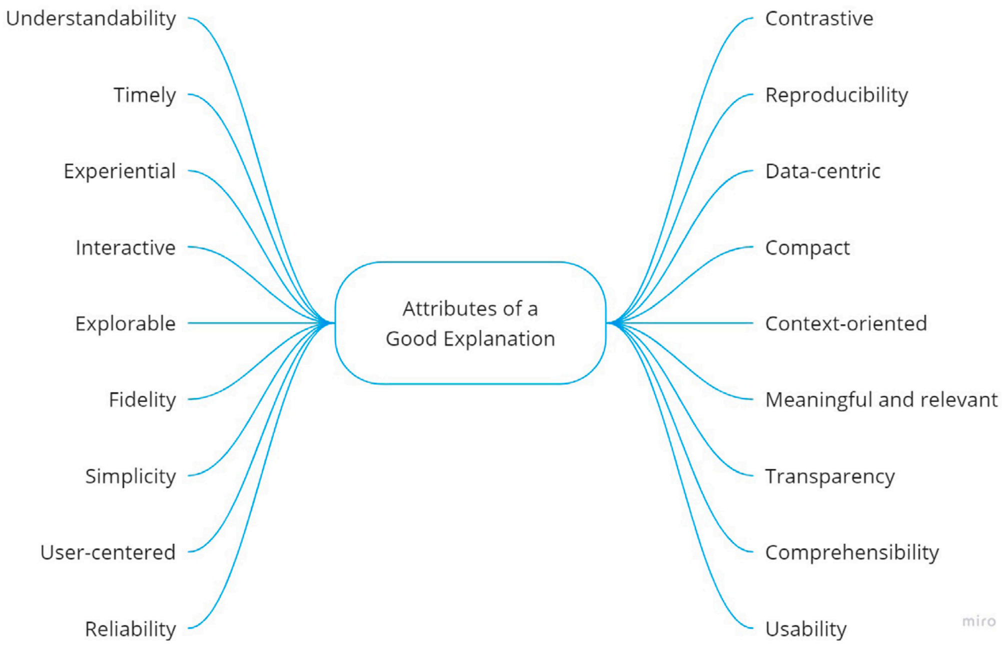


Fig. 4. Summary of attributes that characterize good explanations in AI.

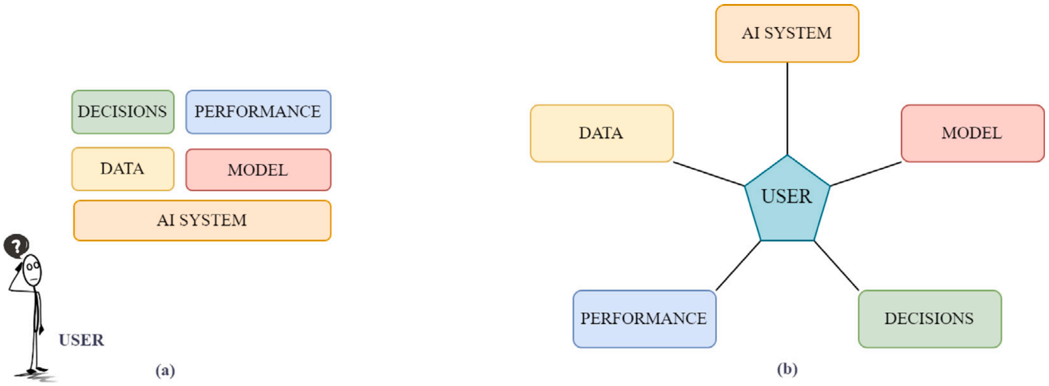
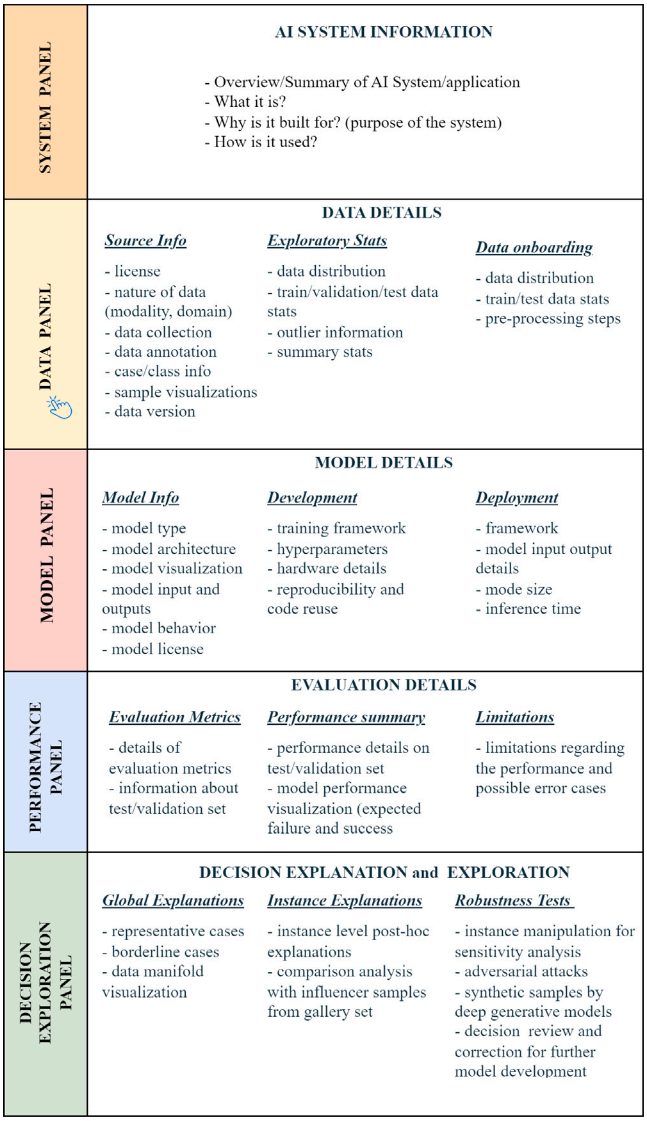


Fig. 5. Design of two explainability approaches in AI: (a) a conventional approach that does not consider the different user groups when presenting explanations, and (b) a user-centric approach that provides explanations tailored to the requirements of different user groups.

There are shortcomings in the existing methods. Firstly, current explanation proposals do not take into account different user groups. However, an ideal solution should include the users in the loop while generating explanations, since there exist different user groups of the AI system and each of them requires different explanations. Additionally, the static nature of current methods makes it difficult for users to interact with the explanations. As noted earlier, interactivity is one of the key elements of good explanations, which is a means of communicating the components of AIsystem to the user, thus building trust. Through an interactive interface, the user can remain in control and can test/verify the robustness of the system by challenging the decisions made by AI [61]. Lastly, both model cards and data cards focus on a specific part of the model development cycle. However, as all AI system components are interdependent, it is important to deconstruct the AI system into its components and generate explanations for each part [22]. Therefore, the framework needs to be both user-centric and interactive in its exploration of AI systems. Fig. 5 illustrates traditional versus desired explanation approaches in XAI.

To address the shortcomings of current approaches and to answer the question of how to explain, we propose a simple but powerful framework that enables explainability at every stage of the AI development cycle. This framework is user-centric, interactive, and deconstructs the AI system into its components. It also meets the characteristics of the questions outlined above. The framework deconstructs the AI system into the following parts and provides explanations for each part under the name of: (1) system overview panel, (2) model panel, (3) data panel, (4) performance and evaluation panel, and (5) decision exploration panel. In this way, transparency is introduced at every stage of AI model development, from pre-modeling to post-modeling. It should be noted that transparency is important to enable alignment with the regulatory compliance demands stated in the EU AI Act [51] and GDPR [52].

The framework is interactive and user-centric and provides explanations tailored to different user groups-[33,35,42]. To create user groups, various methods exist that can be applied to construct well-grounded user profiling and they can be summarized as follows. (1) Dynamically learning user-profile in real-time while the user is engaging with the system and tailoring the explanations based on the learnt profile. (2) Collecting information about possible users of the system via questionnaires or surveys and then constructing user-profiles based on answers. (3) Applying predefined personas for AI, for example, via Personas for AI toolbox [62]. (4) Constructing user groups via given initial user groups studied in XAI as illustrated in Fig. 3. Depending on the adopted method of creating user profiles, at each panel of the



SYSTEM PANEL

AI SYSTEM INFORMATION

- Overview/Summary of AI System/application
- What it is?
- Why is it built for? (purpose of the system)
- How is it used?

DATA PANEL

DATA DETAILS

Source Info
- license
- nature of data (modality, domain)
- data collection
- data annotation
- case/class info
- sample visualizations
- data version

Exploratory Stats
- data distribution
- train/validation/test data stats
- outlier information
- summary stats

Data onboarding
- data distribution
- train/test data stats
- pre-processing steps

MODEL PANEL

MODEL DETAILS

Model Info
- model type
- model architecture
- model visualization
- model input and outputs
- model behavior
- model license

Development
- training framework
- hyperparameters
- hardware details
- reproducibility and code reuse

Deployment
- framework
- model input output details
- mode size
- inference time

PERFORMANCE PANEL

EVALUATION DETAILS

Evaluation Metrics
- details of evaluation metrics
- information about test/validation set

Performance summary
- performance details on test/validation set
- model performance visualization (expected failure and success

Limitations
- limitations regarding the performance and possible error cases

DECISION EXPLORATION PANEL

DECISION EXPLANATION and EXPLORATION

Global Explanations
- representative cases
- borderline cases
- data manifold visualization

Instance Explanations
- instance level post-hoc explanations
- comparison analysis with influencer samples from gallery set

Robustness Tests
- instance manipulation for sensitivity analysis
- adversarial attacks
- synthetic samples by deep generative models
- decision review and correction for further model development

Fig. 6. The proposed framework for XAI. The framework comprises five panels, each of which is dedicated to providing explanations related to the specific aspects highlighted in the respective panel.

proposed framework, the target group user information can be provided in the form of user tags (user groups). The framework and its panels are shown in Fig. 6, and the details of each panel are discussed in the following subsections.

4.2. Definition of framework components

This section covers what can be included and reported in every module of the framework. The modules are flexible and open to extension.

System Panel. This module provides a summary of the AI system and key details about its purpose and use [27]. We propose that details are provided answering for the following questions about the system: (1) what does the system do? (2) For what purpose has it been designed and implemented? And (3) how is it used?

- *What is the problem that the system aims to solve?* The answer to this question should clarify the motivation for building the AI system and explain the problem or need that the system is intended to address.
- *What does the system do?* The answer to this question should provide a high-level overview of the system's main functions. For example, if the system is a recommendation engine, this section could describe the types of recommendations it makes and how it uses user data and other information to generate the recommendations.
- *How is it used?* The answer to this question should present information on how to use the system and its various components. It could include details on the user interface, input and output formats, and any other relevant information that users need to know to use the system effectively.

Data Panel. The data panel module utilizes a data-centric approach to explainability, which enables a detailed understanding of the data used by the AI system. Providing detailed information regarding the data is one of the ways to reduce and mitigate bias in AI systems as stated in [26]. To enable understanding of AI system data and adhering to FAIR principles (findable, accessible, interoperable and reusable) [61], information about the data characteristics should be reported in three subcategories: (1) source information, (2) exploratory data statistics, and (3) data onboarding. The data to be presented is shown in Fig. 6.

- *Raw data details*: This section provides a summary of the data source, data collection and annotation procedure, and describes the data characteristics including its modality (text, image, audio, etc.), format (CSV, JSON, etc.), and domain (healthcare, finance, etc.). It also includes information on ethical considerations and regulatory checks, such as any relevant licenses or permissions to use the data. Additionally, data version information is given to help users track the evolution of the data over time.

- *Exploratory data statistics*: This section includes information on the train, cross-validation and test data division and provides summary statistics and other relevant information from exploratory data analysis. Examples of outliers, if applicable, can also be included in this section to help users understand the data in detail.
- *Data onboarding*: This section provides details on the data preprocessing and post-processing steps, including any data cleaning, normalization, or other transformations that were performed. Also, if any data augmentation techniques and synthetic data are used to enlarge the data, these procedures are described here.

By providing the details of the data in each part, we can introduce transparency and enable regulatory compliance, such as EU AI Act [51] and GDPR [52].

Model Panel. The model panel provides details on the development of the AI model, from the initial development stage to deployment. It enables users to understand what the model does, how it was developed, and how it is used in production. The panel is divided into three subsections: model generic information, model development information, and model deployment information. These subsections provide a user-centric view of the model and its development as shown in Fig. 6:

- *Model generic information*: This section provides a consistent summary of the model's details, including its main features, capabilities, and intended use. It also describes the model's behavior, such as the type of data inputs (image, feature etc.) it can handle and the types of outputs it produces.
- *Model development information*: This section provides details on the model's development, including information on the hyperparameters, development frameworks such as Tensorflow or PyTorch, and other technical details. It also includes a complete analysis of the model's inference performance, such as the model size, hardware-specific (GPU and CPU) inference time, and speed, as well as any information relevant for reproducibility [63].
- *Model deployment information*: This section provides information on how the model is used in production, including details on the inference process and any relevant deployment-specific details. It also includes information on how users can access and interact with the model in production.

Overall, this module provides a transparent and user-centric view of the AI model and its development, enabling users to understand its capabilities and performance and how it is used in practice.

Performance Panel. The performance and evaluation panel provides information on the evaluation of the AI model's performance, including details on the metrics used and the results of the evaluation. It also includes information on any issues that were discovered during the evaluation, as well as visualizations of the performance metrics and examples of observed failure and success cases.

- *Evaluation Metrics*: This section provides a detailed description of the performance metrics used to evaluate the model, such as accuracy, precision, recall, and other task-relevant metrics. It explains how each metric is calculated and provides a rationale for its use in evaluating the model. It should also present the characteristics of the evaluation data set, including its size, composition, and any relevant details on the data used for evaluation. It also includes information on how the evaluation data set was selected and prepared for use in evaluating the model.
- *Performance Summary*: This section provides the results of the model's performance evaluation on the evaluation set, including the values of the performance metrics and any relevant details on the evaluation process. It also includes a comparison of the model's performance on the test set to other relevant benchmarks or baselines if such benchmark data sets exist. It can also include visualizations of the performance metrics, such as graphs, charts, or other relevant visual representations. These visualizations help users to understand the model's performance and its strengths and weaknesses, and they provide a clear and concise view of the evaluation results.
- *Limitations*: This section provides examples of observed failure and success cases, along with visualizations

Decision Exploration Panel. The decision exploration panel is an essential part of understanding the decision-making process of AI models. It provides both local and global explanations, and is user-centric, enabling human interaction [64]. Global explanations can be provided through representative samples and task-specific borderline samples from the training data, such as representative samples from each category in a classification task. These samples enable users to understand which inputs contribute to the model's decisions. On the other hand, borderline cases can be used to highlight potential failure cases.

The instance exploration part of this panel is intended for model sensitivity analysis and robustness tests, decision correction, and decision highlighting using both local explanation methods and presenting similar samples to the target sample from the representative samples of the training set. In this way, a user is given a chance to compare decisions made by AI models across similar cases in the gallery set. There are various methods [65,66] in the literature for highlighting features that contribute to a decision, and these methods can be used to provide explanations.

An important facet of the panel is to provide robustness tests for trustworthy AI. It is crucial to provide the users with a set of options to perform robustness checks. This can be achieved via several approaches found in the literature [48,67]. The notable techniques are (1) augmenting input samples by varying transformations and distortions, (2) testing the robustness via adversarial samples, and (3) utilizing generative models to synthesize new samples to test the reliability of the model. By embracing such methods, the users can challenge the model's decisions by manipulating input instances to perform sensitivity analysis via interactive user interfaces for this purpose. In addition to robustness tests, this panel introduces alignment with human-in-the-loop approaches for verifying the system, enabling domain experts to provide feedback for further improvement of the system via the chance to review and correct the decision of AI's model [26,48,61].

In conclusion, this panel is a powerful tool for providing comprehensive explanations of AI models' decision-making process. It allows users to gain a deeper understanding of the model's behavior and provides a user-centric, interactive platform for analyzing and manipulating input instances, and for decision correction. The implementation of this panel is key to promoting transparency and trust inAI systems by providing users with an accessible and understandable explanation of the model's behavior.

5. Case study: Optical coherence tomography image analysis by deep learning

To demonstrate the framework in use, AI-based OCT image analysis is used as an example. Given the challenges and regulatory landscape within the medical field, there is a demand for human-AI interfaces that utilize XAI techniques to support the decision-making process making use of AI systems [61]. Therefore, this aligns well with the purpose of the framework, making it suitable for demonstrating the framework's capabilities. The case is explained in detail in this section, as well as an example that illustrates how the proposed framework can be used. The framework presents details about the problem, data, model, evaluations, and decision explanations, improving understanding and transparency throughout the entire pipeline of this case study

OCT is a technique that captures detailed cross-sectional images of the retina [68–70]. It has made a significant contribution to the diagnosis of retinal diseases such as age-related macular degeneration (AMD) and diabetic retinopathy [71]. AI based solutions have received a great deal of interest in the field of medical image analysis [72,73] as they are considered a potentially efficient and effective way to identify such disorders from OCT images.

The primary objective of the case study is to detect choroidal neovascularization (CNV), diabetic macular edema (DME) and DRUSEN in the OCT images of the retina using AI. CNV and DME are commonly associated with AMD, which is one of the leading causes of blindness, and early detection and diagnosis are thus critical. An OCT image set utilized in previous studies [74,75] is used in the case study. A demonstration developed for this example case applying the framework can be accessed via private link.[1]

5.1. System panel

This panel presents the purpose of our case study, describes each part of the framework and explains how to use the framework. Fig. 7 contains a screenshot from the system panel serving as a visual aid for further understanding. This comprehensive overview is essential for users to gain insight into the framework's functionality.

5.2. Data panel

This part of the framework provides comprehensive details about the data. The data panel is designed to provide insight into the source of the data, data collection method, licensing and case explanations. Additionally, descriptive data statistics related to the train and test sets are provided along with any post-processing techniques applied before using the data for model training. In the case study, an important issue was identified related to the class imbalance problem. A solution proposed in Algorithm 1 in the Appendix is highlighted to address this challenge effectively. Fig. 8 presents a screenshot of this panel.

5.3. Model panel

The model panel highlights the model details in three parts: (1) general, (2) development and (3) production. Each part is designed for a specific user group and provides important insights into the development of the model for the case study. The panel format allows users to track every stage of modeling from development to deployment so that any issues related to the model can be easily identified and fixed. Furthermore, this structure enables users to have an overview of all aspects involved in building this case study such as architecture design choices, hyperparameter tuning strategies or deploying techniques. A screenshot of this panel is shown in Fig. 9.

5.4. Performance evaluation panel

This panel covers the details of the OCT model evaluation by answering questions about how the model performance is measured, which metrics are used for the evaluation, and which results were obtained using the test data. Additionally, it provides an explanation of any limitations or failure cases that may have occurred during testing. The insights gained from this panel can help users and inform further development and refinement of the model to maximize performance and reliability. Fig. 10 presents a screenshot of the performance evaluation panel.

5.5. Decision exploration panel

Our case study employs a hybrid approach to generate decision explanations that encompass both global and local explanation methods. Global explanations are derived from the training set, while local explanations are produced through an interactive interface that allows the user to manipulate inputs and observe how the model makes a decision in a specific case, including which part of the image it considers relevant for the prediction.

5.5.1. Global explanations

To provide global explanations, we selected representative and adversarial (borderline) cases from the training set. These global explanations are crucial for comparing the similarity between instances and the actual data used for training. Therefore, the representative and adversarial samples are used as the gallery set in the local instance explanations. The procedures for selecting representative and borderline cases are outlined in Algorithms 2 and 3 in the Appendix. Fig. 11 depicts how the global explanations are presented in the framework.

5.5.2. Instance explanations

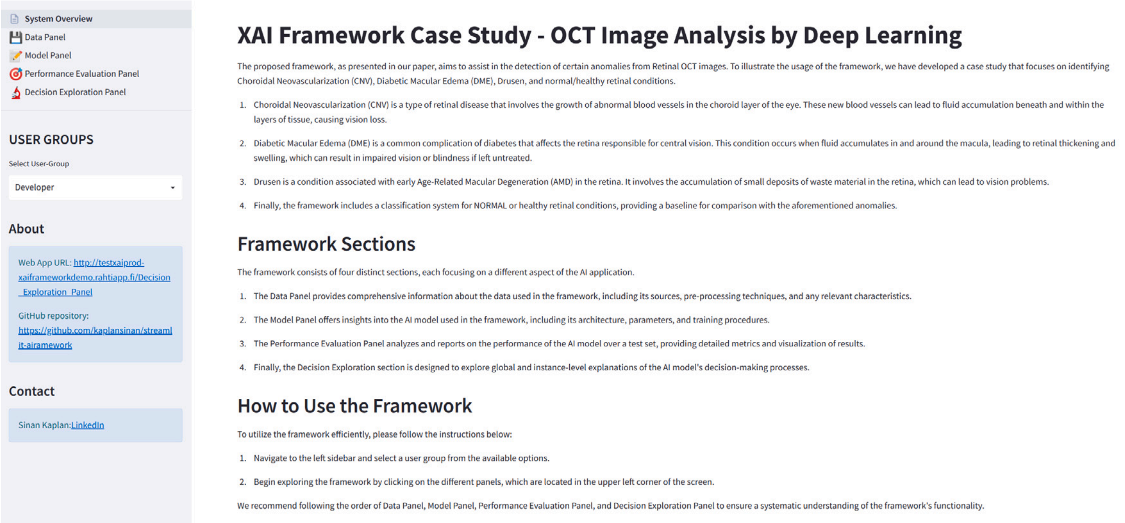
This section describes how the framework presents and explains the prediction of a single instance. The chosen approach allows the end-user to manipulate the input and observe how the prediction changes accordingly. The framework presents the following information for each decision made by the model: (1) the probability distribution of each class prediction, (2) the region of interest (ROI) in the input image that contributes to the decision and highlights where the model focuses while making the prediction, and (3) top-k instances similar to the input image presented from the gallery set. The similarity between an instance and the samples from the gallery set is crucial to understand the model decision. To highlight the ROI, the framework uses the GradCAM [76] method, which is applied with post-processing steps to enhance the visualization. Various local explanation methods are available, for example, Integrated Gradients and SHAP [77], and GradCAM was chosen as an example approach working well with OCT images. Additionally, the framework allows the correction of decisions made by the model, and these corrected samples are added to the next training bucket to allow human-in-the-loop and active learning approaches into the framework. A screenshot of the designed instance explanation module is presented in Fig. 12.

6. Conclusions

In this study, we propose a user-centric and interactive framework for XAI that enables a holistic understanding of AI systems by providing comprehensive explanations at different stages. To achieve this goal, we deconstruct the AI system into its fundamental components, namely data, model, performance evaluation, and post-hoc decisions. Concise explanations are then provided for each stage, which promotes transparency throughout the stages of AI development. The framework is designed to take into account the diverse needs of different users and provide explanations tailored to their needs. In addition, interactivity is a key design principle of the framework, which encourages engagement between users and AI systems.

A case study from the medical domain is presented to demonstrate the usability of our framework. For this purpose, we developed an OCT image analysis system and utilize our framework to report on its development. The case study demonstrates how the framework can be used to provide a clear understanding of the data and the model, how the performance of the AI system was evaluated, and the limitations of the developed system. Additionally, an interactive decision exploration panel allows users to engage with the AI model and obtain explanations for the model decisions at the instance level and also the global level. Overall, the framework was seen to be useful in promoting transparency and trust in AI systems. The use case presented in this paper serves

[1] https://huggingface.co/spaces/hodorfi/xai_framework



USER GROUPS

Select User-Group: Developer

About

Web App URL: http://testxaiprod-xaiframeworkdemo.rahtiapp.fi/Decision_Exploration_Panel

GitHub repository: https://github.com/kaplansinan/streamlit-airamework

Contact

Sinan Kaplan:LinkedIn

XAI Framework Case Study - OCT Image Analysis by Deep Learning

The proposed framework, as presented in our paper, aims to assist in the detection of certain anomalies from Retinal OCT images. To illustrate the usage of the framework, we have developed a case study that focuses on identifying Choroidal Neovascularization (CNV), Diabetic Macular Edema (DME), Drusen, and normal/healthy retinal conditions.

1. Choroidal Neovascularization (CNV) is a type of retinal disease that involves the growth of abnormal blood vessels in the choroid layer of the eye. These new blood vessels can lead to fluid accumulation beneath and within the layers of tissue, causing vision loss.
2. Diabetic Macular Edema (DME) is a common complication of diabetes that affects the retina responsible for central vision. This condition occurs when fluid accumulates in and around the macula, leading to retinal thickening and swelling, which can result in impaired vision or blindness if left untreated.
3. Drusen is a condition associated with early Age-Related Macular Degeneration (AMD) in the retina. It involves the accumulation of small deposits of waste material in the retina, which can lead to vision problems.
4. Finally, the framework includes a classification system for NORMAL or healthy retinal conditions, providing a baseline for comparison with the aforementioned anomalies.

Framework Sections

The framework consists of four distinct sections, each focusing on a different aspect of the AI application.

1. The Data Panel provides comprehensive information about the data used in the framework, including its sources, pre-processing techniques, and any relevant characteristics.
2. The Model Panel offers insights into the AI model used in the framework, including its architecture, parameters, and training procedures.
3. The Performance Evaluation Panel analyzes and reports on the performance of the AI model over a test set, providing detailed metrics and visualization of results.
4. Finally, the Decision Exploration section is designed to explore global and instance-level explanations of the AI model's decision-making processes.

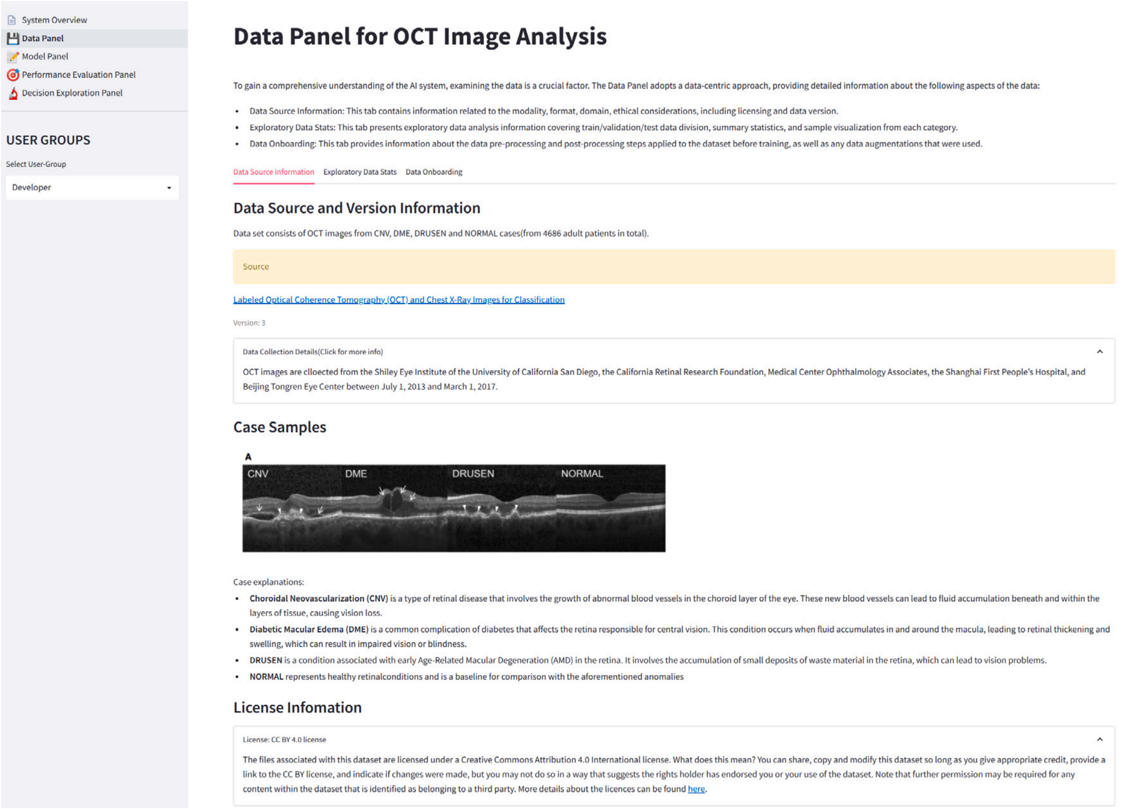
How to Use the Framework

To utilize the framework efficiently, please follow the instructions below:

1. Navigate to the left sidebar and select a user group from the available options.
2. Begin exploring the framework by clicking on the different panels, which are located in the upper left corner of the screen.

We recommend following the order of Data Panel, Model Panel, Performance Evaluation Panel, and Decision Exploration Panel to ensure a systematic understanding of the framework's functionality.

Fig. 7. The screenshot of the framework's system panel from the case study: it offers an overview of the system design and instructions on how to use it. By selecting a user group and a framework panel from the left sidebar, users can interact with the system.



USER GROUPS

Select User-Group: Developer

Data Panel for OCT Image Analysis

To gain a comprehensive understanding of the AI system, examining the data is a crucial factor. The Data Panel adopts a data-centric approach, providing detailed information about the following aspects of the data:

- Data Source Information: This tab contains information related to the modality, format, domain, ethical considerations, including licensing and data version.
- Exploratory Data Stats: This tab presents exploratory data analysis information covering train/validation/test data division, summary statistics, and sample visualization from each category.
- Data Onboarding: This tab provides information about the data pre-processing and post-processing steps applied to the dataset before training, as well as any data augmentations that were used.

Data Source and Version Information

Data set consists of OCT images from CNV, DME, DRUSEN and NORMAL cases(from 4686 adult patients in total).

Source

Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification

Version: 3

Data Collection Details(Click for more info)

OCT images are clloected from the Shiley Eye Institute of the University of California San Diego, the California Retinal Research Foundation, Medical Center Ophthalmology Associates, the Shanghai First People's Hospital, and Beijing Tongren Eye Center between July 1, 2013 and March 1, 2017.

Case Samples

Case explanations:

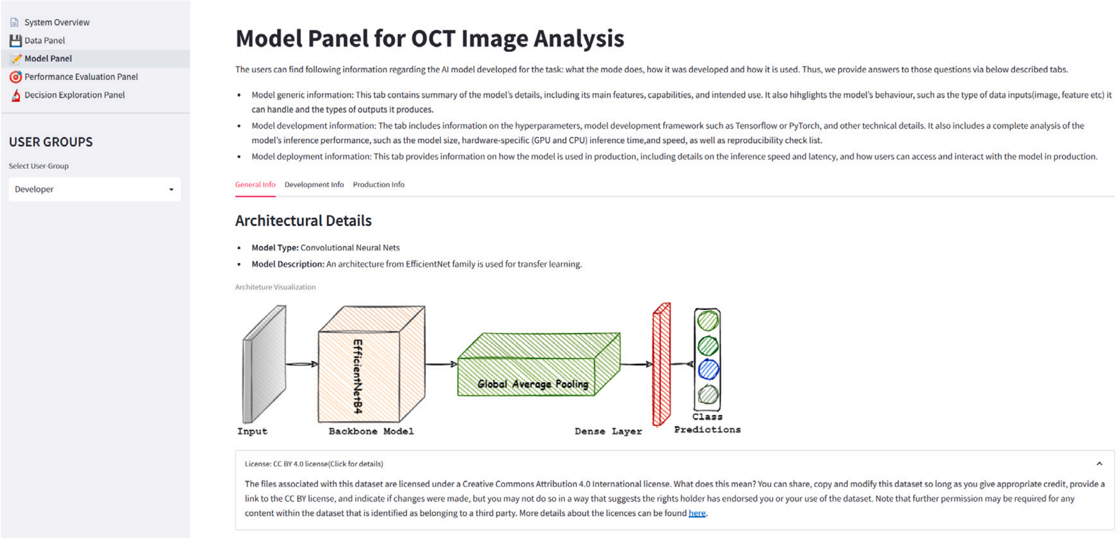
- **Choroidal Neovascularization (CNV)** is a type of retinal disease that involves the growth of abnormal blood vessels in the choroid layer of the eye. These new blood vessels can lead to fluid accumulation beneath and within the layers of tissue, causing vision loss.
- **Diabetic Macular Edema (DME)** is a common complication of diabetes that affects the retina responsible for central vision. This condition occurs when fluid accumulates in and around the macula, leading to retinal thickening and swelling, which can result in impaired vision or blindness.
- **DRUSEN** is a condition associated with early Age-Related Macular Degeneration (AMD) in the retina. It involves the accumulation of small deposits of waste material in the retina, which can lead to vision problems.
- **NORMAL** represents healthy retinalconditions and is a baseline for comparison with the aforementioned anomalies

License Infomation

License: CC BY 4.0 license

The files associated with this dataset are licensed under a Creative Commons Attribution 4.0 International license. What does this mean? You can share, copy and modify this dataset so long as you give appropriate credit, provide a link to the CC BY license, and indicate if changes were made, but you may not do so in a way that suggests the rights holder has endorsed you or your use of the dataset. Note that further permission may be required for any content within the dataset that is identified as belonging to a third party. More details about the licences can be found here.

Fig. 8. The screenshot of the framework's data panel from the case study: The panel has three different tabs, each offering distinct explanations that focus on different aspects of the data utilized for model training. The first tab, which contains general information about the data, is presented in this figure, while the remaining tabs can be accessed via the link provided above.



USER GROUPS

Select User-Group: Developer

Model Panel for OCT Image Analysis

The users can find following information regarding the AI model developed for the task: what the mode does, how it was developed and how it is used. Thus, we provide answers to those questions via below described tabs.

- Model generic information: This tab contains summary of the model's details, including its main features, capabilities, and intended use. It also hihglights the model's behaviour, such as the type of data inputs(image, feature etc) it can handle and the types of outputs it produces.
- Model development information: The tab includes information on the hyperparameters, model development framework such as Tensorflow or PyTorch, and other technical details. It also includes a complete analysis of the model's inference performance, such as the model size, hardware-specific (GPU and CPU) inference time,and speed, as well as reproducibility check list.
- Model deployment information: This tab provides information on how the model is used in production, including details on the inference speed and latency, and how users can access and interact with the model in production.

Architectural Details

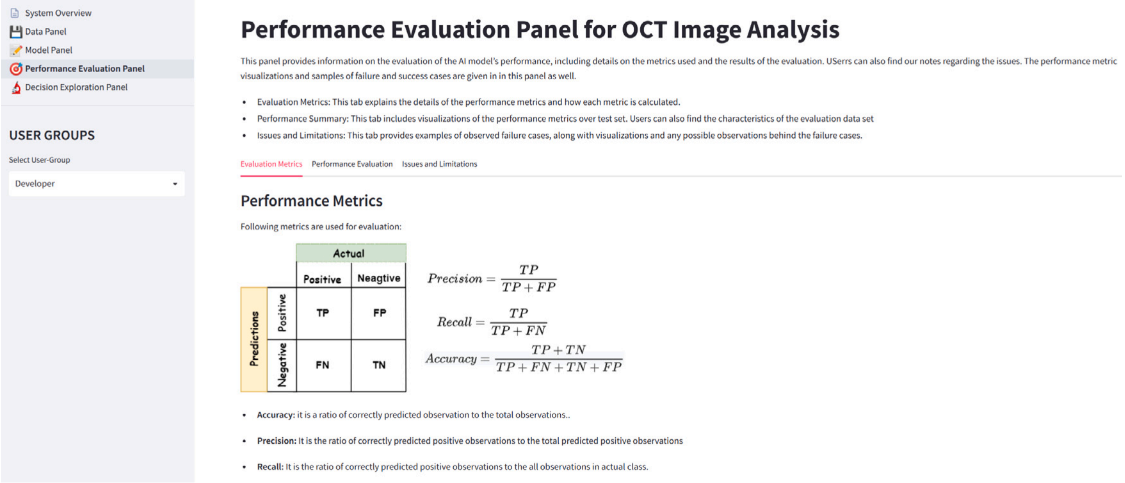
- **Model Type:** Convolutional Neural Nets
- **Model Description:** An architecture from EfficientNet family is used for transfer learning.

Architeture Visualization

License: CC BY 4.0 license(Click for details)

The files associated with this dataset are licensed under a Creative Commons Attribution 4.0 International license. What does this mean? You can share, copy and modify this dataset so long as you give appropriate credit, provide a link to the CC BY license, and indicate if changes were made, but you may not do so in a way that suggests the rights holder has endorsed you or your use of the dataset. Note that further permission may be required for any content within the dataset that is identified as belonging to a third party. More details about the licences can be found here.

Fig. 9. The screenshot of the framework's model panel from the case study: each tab presents distinct explanations related to different aspects of the model used in experiments. This figure displays the first tab, which offers general information about the model.



USER GROUPS

Select User-Group: Developer

Performance Evaluation Panel for OCT Image Analysis

This panel provides information on the evaluation of the AI model's performance, including details on the metrics used and the results of the evaluation. USerrs can also find our notes regarding the issues. The performance metric visualizations and samples of failure and success cases are given in in this panel as well.

- Evaluation Metrics: This tab explains the details of the performance metrics and how each metric is calculated.
- Performance Summary: This tab includes visualizations of the performance metrics over test set. Users can also find the characteristics of the evaluation data set
- Issues and Limitations: This tab provides examples of observed failure cases, along with visualizations and any possible observations behind the failure cases.

Performance Metrics

Following metrics are used for evaluation:

		Actual	
		Positive	Neagtive
Predictions	Positive	TP	FP
	Negative	FN	TN

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

- **Accuracy:** it is a ratio of correctly predicted observation to the total observations..
- **Precision:** It is the ratio of correctly predicted positive observations to the total predicted positive observations
- **Recall:** It is the ratio of correctly predicted positive observations to the all observations in actual class.

Fig. 10. The screenshot of the framework's performance evaluation panel from the case study: three tabs are visible, each providing distinct explanations related to performance evaluation. This figure shows the first tab, which highlights the performance metrics used in the evaluation.

as a baseline example of how the framework can be used to provide explanations for similar systems.

To continue the research, different options to create or learn user groups to provide explanations can be adopted to the framework. However, a more in-depth study of identifying and testing user groups is required for fine-grained user-centric explanations. A further area of future research is assessing the effectiveness of the framework on medical image analysis and diagnosis via an empirical study, which should involve medical experts utilizing the framework followed by a comprehensive questionnaire. This methodological approach can yield valuable insight into its effectiveness and areas for improvement and enhancement. Additionally, the quality assessment of the explanations provided by the framework may be conducted by applying the System Causability Scale introduced in [78]. Lastly, cross-domain applicability and generalizability of the framework across different domains can be tested further to examine its usability both within and beyond the medical field.

CRediT authorship contribution statement

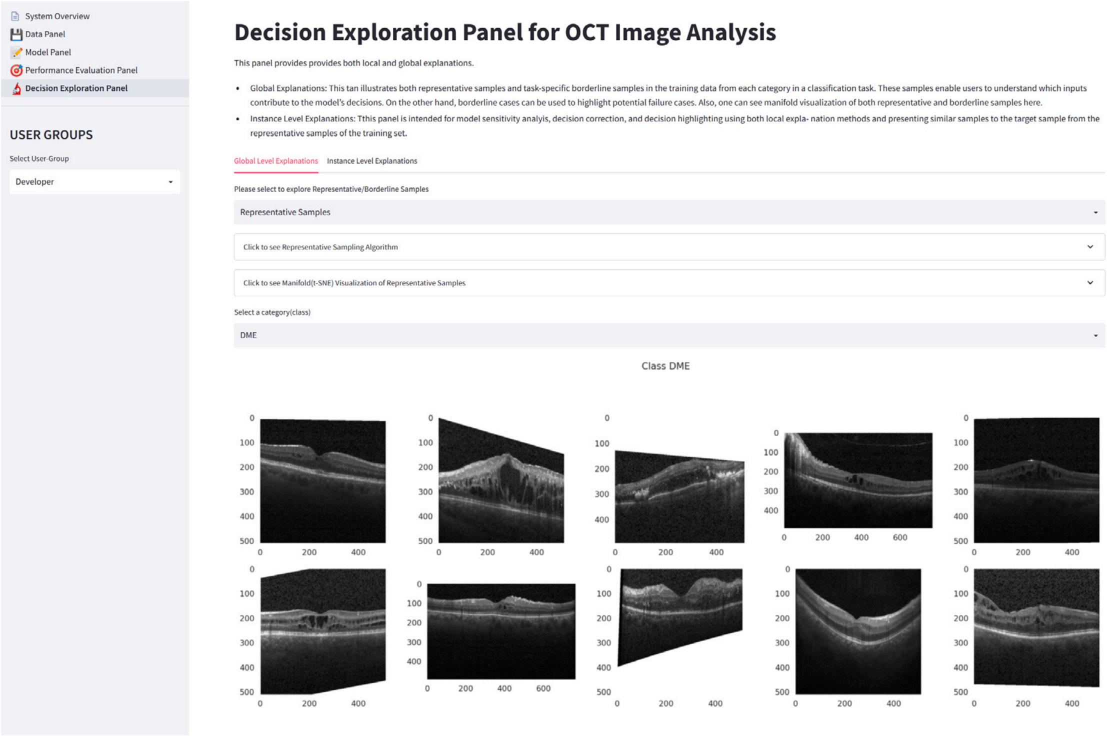
Sinan Kaplan: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Hannu Uusitalo:** Writing – review & editing. **Lasse Lensu:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The code is shared as a case study and the link is given in the paper. The details regarding the data source can be also found in the demonstration of the case study.



Decision Exploration Panel for OCT Image Analysis

This panel provides provides both local and global explanations.

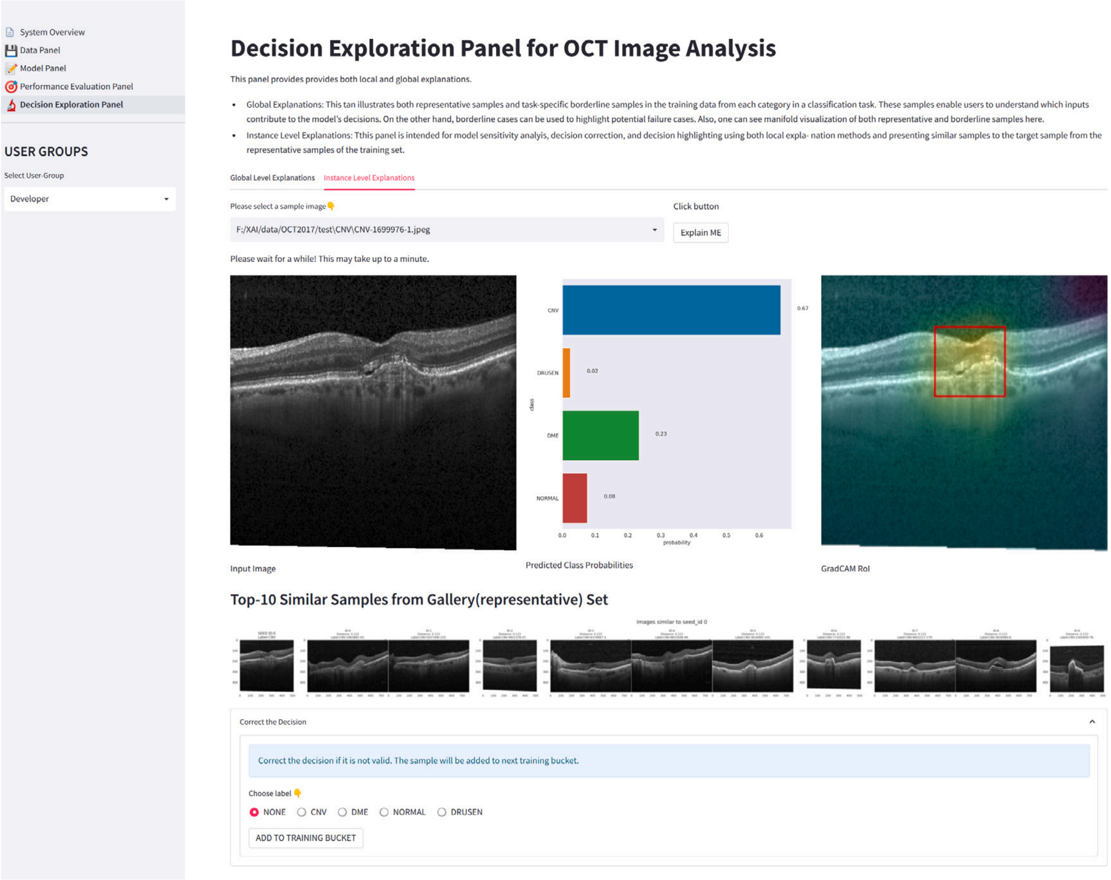
- Global Explanations: This tan illustrates both representative samples and task-specific borderline samples in the training data from each category in a classification task. These samples enable users to understand which inputs contribute to the model's decisions. On the other hand, borderline cases can be used to highlight potential failure cases. Also, one can see manifold visualization of both representative and borderline samples here.
- Instance Level Explanations: Tthis panel is intended for model sensitivity analyis, decision correction, and decision highlighting using both local expla- nation methods and presenting similar samples to the target sample from the representative samples of the training set.

Please select to explore Representative/Borderline Samples: Representative Samples

Click to see Representative Sampling Algorithm
Click to see Manifold(t-SNE) Visualization of Representative Samples

Select a category(class): DME

Fig. 11. The decision exploration panel of the framework, as seen in the case study screenshot, presents the global explanation tab, which enables users to explore representative and borderline samples from each category.



Decision Exploration Panel for OCT Image Analysis

This panel provides provides both local and global explanations.

- Global Explanations: This tan illustrates both representative samples and task-specific borderline samples in the training data from each category in a classification task. These samples enable users to understand which inputs contribute to the model's decisions. On the other hand, borderline cases can be used to highlight potential failure cases. Also, one can see manifold visualization of both representative and borderline samples here.
- Instance Level Explanations: Tthis panel is intended for model sensitivity analyis, decision correction, and decision highlighting using both local expla- nation methods and presenting similar samples to the target sample from the representative samples of the training set.

Please select a sample image 👇: F:/XAI/data/OCT2017/test\CNV\CNV-1699976-1.jpeg

Please wait for a while! This may take up to a minute.

Top-10 Similar Samples from Gallery(representative) Set

Correct the Decision

Correct the decision if it is not valid. The sample will be added to next training bucket.

Choose label 👇

● NONE ○ CNV ○ DME ○ NORMAL ○ DRUSEN

Fig. 12. The screenshot of the framework's decision exploration panel from the case study: the instance explanation tab is displayed, which allows users to obtain predictions and explanations for a given input or sample.

Algorithm 1 Class Imbalance Solver (Representative Sampling)

```
DATASET = D
CLASS_LIST = ['NORMAL', 'DME', 'DRUSEN', 'CNV']
# resnet50 model trained on Imagenet
EMBEDDING_MODEL = RESNET50
#number of instances to be sampled from each class
DESIRED_NUMBER_OF_SAMPLES_PER_CLASS = N

def class_imbalance_solver():
  SAMPLE_LIST = []
  for class in CLASS_LIST:
    # get all samples of a class in dataset D
    data = get_class_samples(class, D)
    # extract embedding features from the data
    data_emb_features = get_embedding_features(data, EMBEDDING_MODEL)
    # run hierarchical clustering over the embedding features
    clusters = run_hierarchical_clustering(data_emb_features)
    # sample representative samples from each cluster
    repr_samples = get_n_representative_samples(clusters,N)
    # add representative samples to the sampling bucket
    SAMPLE_LIST.append(repr_samples)

  return SAMPLE_LIST
```

Algorithm 2 Representative Sampling

```
TRAINING_DATASET = T
TRAINED_OCT_MODEL = OCT_MODEL_V1
# number of instances to be sampled for gallery set
DESIRED_NUMBER_OF_SAMPLES_FOR_GALLERY_SET = N

def representative_gallery_set_sampler():
  data = get_correctly_classified_samples(T)
  data_emb = get_embedding_features(data, OCT_MODEL_V1)
  ann_tree = build_approximate_nearest_neighbor_tree(data_emb)
  gallery_set = get_n_dissimilar_samples(annoy_tree)
  return gallery_set
```

Algorithm 3 Borderline Sampling

```
TRAINING_DATASET = T
TRAINED_OCT_MODEL = OCT_MODEL_V1
# number of instances to be sampled for gallery set
DESIRED_NUMBER_OF_SAMPLES_FOR_GALLERY_SET = N

def borderline_gallery_set_sampler():
  data = get_missclassified_samples(T)
  data_emb = get_embedding_features(data, OCT_MODEL_V1)
  ann_tree = build_approximate_nearest_neighbor_tree(data_emb)
  gallery_set = get_n_dissimilar_samples(annoy_tree)
  return gallery_set
```

Appendix. Algorithms

See Algorithms 1–3.

References

[1] S. Dong, P. Wang, K. Abbas, A survey on deep learning and its applications, Comp. Sci. Rev. 40 (2021) 100379.

[2] S. Dargan, M. Kumar, M.R. Ayyagari, G. Kumar, A survey of deep learning and its applications: A new paradigm to machine learning, Arch. Comput. Methods Eng. 27 (4) (2020) 1071–1092.

[3] S. Zhang, L. Yao, A. Sun, Y. Tay, Deep learning based recommender system: A survey and new perspectives, ACM Comput. Surv. 52 (1) (2019) 1–38.

[4] F. Fessahaye, L. Perez, T. Zhan, R. Zhang, C. Fossier, R. Markarian, C. Chiu, J. Zhan, L. Gewali, P. Oh, T-recsys: A novel music recommendation system using deep learning, in: 2019 IEEE International Conference on Consumer Electronics, ICCE, IEEE, 2019, pp. 1–6.

[5] X. Zhou, W. Gong, W. Fu, F. Du, Application of deep learning in object detection, in: 2017 IEEE/ACIS 16th International Conference on Computer and Information Science, ICIS, IEEE, 2017, pp. 631–634.

[6] M. Badar, M. Haris, A. Fatima, Application of deep learning for retinal image analysis: A review, Comp. Sci. Rev. 35 (2020) 100203.

[7] O. Willers, S. Sudholt, S. Raafatnia, S. Abrecht, Safety concerns and mitigation approaches regarding the use of deep learning in safety-critical perception

tasks, in: International Conference on Computer Safety, Reliability, and Security, Springer, 2020, pp. 336–350.

[8] J. Linnosmaa, P. Tikka, J. Suomalainen, N. Papakonstantinou, Machine learning in safety critical industry domains, 2020.

[9] S. Grigorescu, B. Trasnea, T. Cocias, G. Macesanu, A survey of deep learning techniques for autonomous driving, J. Field Robotics 37 (3) (2020) 362–386.

[10] B. Sahoh, A. Choksuriwong, The role of explainable artificial intelligence in high-stakes decision-making systems: A systematic review, J. Ambient Intell. Humaniz. Comput. (2023) 1–17.

[11] Y. Zhang, P. Tiňo, A. Leonardis, K. Tang, A survey on neural network interpretability, IEEE Trans. Emerg. Top. Comput. Intell. (2021).

[12] W. Jin, X. Li, M. Fatehi, G. Hamarneh, Guidelines and evaluation of clinical explainable AI in medical image analysis, Med. Image Anal. 84 (2023) 102684.

[13] W. Samek, G. Montavon, A. Vedaldi, L.K. Hansen, K.-R. Müller, Explainable AI: interpreting, Explaining and Visualizing Deep Learning, vol. 11700, Springer Nature, 2019.

[14] W. Saeed, C. Omlin, Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities, Knowl.-Based Syst. (2023) 110273.

[15] X. Bai, X. Wang, X. Liu, Q. Liu, J. Song, N. Sebe, B. Kim, Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments, Pattern Recognit. 120 (2021) 108102.

[16] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, et al., Explainable AI (XAI): Core ideas, techniques, and solutions, ACM Comput. Surv. 55 (9) (2023) 1–33.

[17] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis, Explainable AI: A review of machine learning interpretability methods, Entropy 23 (1) (2021) 18.

[18] P.P. Angelov, E.A. Soares, R. Jiang, N.I. Arnold, P.M. Atkinson, Explainable artificial intelligence: An analytical review, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 11 (5) (2021) e1424.

[19] T. Gebru, J. Morgenstern, B. Vecchione, J.W. Vaughan, H. Wallach, H.D. Iii, K. Crawford, Datasheets for datasets, Commun. ACM 64 (12) (2021) 86–92.

[20] M. Pushkarna, A. Zaldivar, O. Kjartansson, Data cards: Purposeful and transparent dataset documentation for responsible AI, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 1776–1826, http://dx.doi.org/10.1145/3531146.3533231.

[21] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I.D. Raji, T. Gebru, Model cards for model reporting, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019, pp. 220–229.

[22] ISO/IEC, Information Technology — Artificial Intelligence — Overview of Ethical and Societal Concerns, Tech. Rep. ISO/IEC TR 24368:2022, 2022.

[23] D. Kreuzberger, N. Kühl, S. Hirschl, Machine learning operations (MLOps): Overview, definition, and architecture, 2022, arXiv preprint arXiv:2205.02302.

[24] I.S. Di Laurea, MLOps-Standardizing the Machine Learning Workflow (Ph.D. thesis), University of Bologna Bologna, Italy, 2021.

[25] U. Paschen, C. Pitt, J. Kietzmann, Artificial intelligence: Building blocks and an innovation typology, Bus. Horiz. 63 (2) (2020) 147–155.

[26] ISO/IEC, Information Technology — Artificial Intelligence — Overview of Trustworthiness in Artificial Intelligence, Tech. Rep. ISO/IEC TR 24028:2020, 2020.

[27] A. Chaddad, J. Peng, J. Xu, A. Bouridane, Survey of explainable AI techniques in healthcare, Sensors 23 (2) (2023) 634.

[28] S.T. Mueller, R.R. Hoffman, W. Clancey, A. Emrey, G. Klein, Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI, 2019, arXiv preprint arXiv:1902.01876.

[29] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J.M.F. Moura, P. Eckersley, Explainable machine learning in deployment, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 648–657.

[30] B. Mittelstadt, C. Russell, S. Wachter, Explaining explanations in AI, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019, pp. 279–288.

[31] V. Belle, I. Papantonis, Principles and practice of explainable machine learning, Front. Big Data (2021) 39.

[32] A. Kirsch, Explain to whom? Putting the user in the center of explainable AI, in: Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 Co-Located with 16th International Conference of the Italian Association for Artificial Intelligence, AI* IA 2017, 2017.

[33] S. Laato, M. Tiainen, A.K.M.N. Islam, M. Mäntymäki, How to explain AI systems to end users: A systematic literature review and research agenda, Internet Res. 32 (7) (2022) 1–31.

[34] M. Ribera, A. Lapedriza, Can we do better explanations? A proposal of user-centered explainable AI, in: IUI Workshops, Vol. 2327, 2019, p. 38.

[35] T.A.J. Schoonderwoerd, W. Jorritsma, M.A. Neerincx, K. Van Den Bosch, Human-centered XAI: Developing design patterns for explanations of clinical decision support systems, Int. J. Hum.-Comput. Stud. 154 (2021) 102684.

[36] A. Crisan, M. Drouhard, J. Vig, N. Rajani, Interactive model cards: A human-centered approach to model documentation, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 427–439, http://dx.doi.org/10.1145/3531146.3533108.

[37] A. Abdul, J. Vermeulen, D. Wang, B.Y. Lim, M. Kankanhalli, Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 2018, pp. 1–18.

[38] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (5) (2018) 1–42.

[39] J. Díez, K. Khalifa, B. Leuridan, General theories of explanation: Buyer beware, Synthese 190 (3) (2013) 379–396.

[40] F.C. Keil, Explanation and understanding, Annu. Rev. Psychol. 57 (2006) 227.

[41] G. Vilone, L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, Inf. Fusion 76 (2021) 89–106.

[42] F. Cabitza, A. Campagner, G. Malgieri, C. Natali, D. Schneeberger, K. Stoeger, A. Holzinger, Quod erat demonstrandum?-towards a typology of the concept of explanation for the design of explainable AI, Expert Syst. Appl. 213 (2023) 118888.

[43] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artif. Intell. 267 (2019) 1–38.

[44] A. Holzinger, C. Biemann, C.S. Pattichis, D.B. Kell, What do we need to build explainable AI systems for the medical domain? 2017, arXiv preprint arXiv:1712.09923.

[45] W. Samek, T. Wiegand, K.-R. Müller, Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models, 2017, arXiv preprint arXiv:1708.08296.

[46] L. Bertossi, F. Geerts, Data quality and explainable AI, J. Data Inf. Qual. 12 (2) (2020) 1–9.

[47] Z. Wang, P.A. Keane, M. Chiang, C.Y. Cheung, T.Y. Wong, D.S.W. Ting, Artificial intelligence and deep learning in ophthalmology, Artif. Intell. Med. (2020) 1–34.

[48] A. Holzinger, The next frontier: AI we can really trust, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2021, pp. 427–440.

[49] J. Klaise, A. Van Looveren, C. Cox, G. Vacanti, A. Coca, Monitoring and explainability of models in production, 2020, arXiv preprint arXiv:2007.06299.

[50] S. Holland, A. Hosny, S. Newman, J. Joseph, K. Chmielinski, The dataset nutrition label: A framework to drive higher data quality standards, 2018, arXiv preprint arXiv:1805.03677.

[51] M. Kop, EU artificial intelligence act: The European approach to AI, Transatl. Antitrust IPR Dev. (2) (2021) URL https://papers.ssrn.com/abstract=3930959.

[52] P. Voigt, A. Von dem Bussche, The EU general data protection regulation (GDPR), in: A Practical Guide, 1st Ed., 10, (3152676) Springer International Publishing, Cham, 2017, pp. 10–5555.

[53] H. Muller, M.T. Mayrhofer, E.-B. Van Veen, A. Holzinger, The ten commandments of ethical medical AI, Computer 54 (07) (2021) 119–123.

[54] U. Ehsan, M.O. Riedl, Human-centered explainable AI: Towards a reflective sociotechnical approach, in: International Conference on Human-Computer Interaction, Springer, 2020, pp. 449–466.

[55] S.T. Mueller, E.S. Veinott, R.R. Hoffman, G. Klein, L. Alam, T. Mamun, W.J. Clancey, Principles of explanation in human-AI systems, 2021, arXiv preprint arXiv:2102.04972.

[56] A. Der Kiureghian, O. Ditlevsen, Aleatory or epistemic? Does it matter? Struct. Saf. 31 (2) (2009) 105–112.

[57] M. Kläs, A.M. Vollmer, Uncertainty in machine learning applications: A practice-driven classification of uncertainty, in: International Conference on Computer Safety, Reliability, and Security, Springer, 2018, pp. 431–438.

[58] U. Bhatt, J. Antorán, Y. Zhang, Q.V. Liao, P. Sattigeri, R. Fogliato, G. Melançon, R. Krishnan, J. Stanley, O. Tickoo, et al., Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021, pp. 401–413.

[59] M. Arnold, R.K.E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K.N. Ramamurthy, A. Olteanu, D. Piorkowski, et al., FactSheets: Increasing trust in AI services through supplier's declarations of conformity, IBM J. Res. Dev. 63 (4/5) (2019) 6–1.

[60] S. Holland, A. Hosny, S. Newman, J. Joseph, K. Chmielinski, The dataset nutrition label, Volume 12: Data Protection and Democracy, Data Protect. Priv. 12 (2020) 1.

[61] H. Müller, A. Holzinger, M. Plass, L. Brcic, C. Stumptner, K. Zatloukal, Explainability and causability for artificial intelligence-supported medical image analysis in the context of the European in vitro diagnostic regulation, New Biotechnol. 70 (2022) 67–72.

[62] A. Holzinger, M. Kargl, B. Kipperer, P. Regitnig, M. Plass, H. Müller, Personas for artificial intelligence (AI) an open source toolbox, IEEE Access 10 (2022) 23732–23747.

[63] J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché Buc, E. Fox, H. Larochelle, Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program), J. Mach. Learn. Res. 22 (1) (2021) 7459–7478.

[64] B. Leichtmann, C. Humer, A. Hinterreiter, M. Streit, M. Mara, Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task, Comput. Hum. Behav. 139 (2023) 107539.

[65] T. Fel, L. Hervier, D. Vigouroux, A. Poche, J. Plakoo, R. Cadene, M. Chalvidal, J. Colin, T. Boissin, L. Bethune, et al., Xplique: A deep learning explainability toolbox, 2022, arXiv preprint arXiv:2206.04394.

[66] L.K. Gupta, D. Koundal, S. Mongia, Explainable methods for image-based deep learning: A review, Arch. Comput. Methods Eng. (2023) 1–16.

[67] A. Holzinger, M. Dehmer, F. Emmert-Streib, R. Cucchiara, I. Augenstein, J. Del Ser, W. Samek, I. Jurisica, N. Díaz-Rodríguez, Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence, Inf. Fusion 79 (2022) 263–278.

[68] D. Huang, E.A. Swanson, C.P. Lin, J.S. Schuman, W.G. Stinson, W. Chang, M.R. Hee, T. Flotte, K. Gregory, C.A. Puliafito, et al., Optical coherence tomography, science 254 (5035) (1991) 1178–1181.

[69] M.D. Abràmoff, M.K. Garvin, M. Sonka, Retinal imaging and image analysis, IEEE Rev. Biomed. Eng. 3 (2010) 169–208.

[70] D. Hillmann, OCT on a chip aims at high-quality retinal imaging, Light Sci. Appl. 10 (2021).

[71] J. De Fauw, J.R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, et al., Clinically applicable deep learning for diagnosis and referral in retinal disease, Nature Med. 24 (9) (2018) 1342–1350.

[72] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, J. Dean, A guide to deep learning in healthcare, Nature Med. 25 (1) (2019) 24–29.

[73] S. Sengupta, A. Singh, H.A. Leopold, T. Gulati, V. Lakshminarayanan, Ophthalmic diagnosis using deep learning with fundus images—A critical review, Artif. Intell. Med. 102 (2020) 101758.

[74] D.S. Kermany, M. Goldbaum, W. Cai, C.C.S. Valentim, H. Liang, S.L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, et al., Identifying medical diagnoses and treatable diseases by image-based deep learning, Cell 172 (5) (2018) 1122–1131.

[75] D. Kermany, K. Zhang, M. Goldbaum, et al., Labeled optical coherence tomography (OCT) and chest X-ray images for classification, Mendeley Data 2 (2) (2018).

[76] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[77] A. Singh, A.R. Mohammed, J. Zelek, V. Lakshminarayanan, Interpretation of deep learning using attributions: Application to ophthalmic diagnosis, in: Applications of Machine Learning 2020, vol. 11511, SPIE, 2020, pp. 39–49.

[78] A. Holzinger, A. Carrington, H. Müller, Measuring the quality of explanations: The system causability scale (SCS) comparing human and machine explanations, KI-Künstliche Intelligenz 34 (2) (2020) 193–198.