Heliyon 9 (2023) e16110

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Review article

# Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review

Jinsun Jung [a,f], Hyungbok Lee [a,d], Hyunggu Jung [b,c], Hyeoneui Kim [a,e,*]

[a] College of Nursing, Seoul National University, Seoul, Republic of Korea
[b] Department of Computer Science and Engineering, University of Seoul, Seoul, Republic of Korea
[c] Department of Artificial Intelligence, University of Seoul, Seoul, Republic of Korea
[d] Emergency Nursing Department, Seoul National University Hospital, Seoul, Republic of Korea
[e] Research Institute of Nursing Science, College of Nursing, Seoul National University, Seoul, Republic of Korea
[f] Center for Human-Caring Nurse Leaders for the Future by Brain Korea 21 (BK 21) Four Project, College of Nursing, Seoul National University, Seoul, Republic of Korea

ARTICLE INFO

*Keywords:*
Explainable artificial intelligence
Interpretable artificial intelligence
Essential properties of XAI
Evaluation of explanation effectiveness

ABSTRACT

*Background:* Significant advancements in the field of information technology have influenced the creation of trustworthy explainable artificial intelligence (XAI) in healthcare. Despite improved performance of XAI, XAI techniques have not yet been integrated into real-time patient care.
*Objective:* The aim of this systematic review is to understand the trends and gaps in research on XAI through an assessment of the essential properties of XAI and an evaluation of explanation effectiveness in the healthcare field.
*Methods:* A search of PubMed and Embase databases for relevant peer-reviewed articles on development of an XAI model using clinical data and evaluating explanation effectiveness published between January 1, 2011, and April 30, 2022, was conducted. All retrieved papers were screened independently by the two authors. Relevant papers were also reviewed for identification of the essential properties of XAI (e.g., stakeholders and objectives of XAI, quality of personalized explanations) and the measures of explanation effectiveness (e.g., mental model, user satisfaction, trust assessment, task performance, and correctability).
*Results:* Six out of 882 articles met the criteria for eligibility. Artificial Intelligence (AI) users were the most frequently described stakeholders. XAI served various purposes, including evaluation, justification, improvement, and learning from AI. Evaluation of the quality of personalized explanations was based on fidelity, explanatory power, interpretability, and plausibility. User satisfaction was the most frequently used measure of explanation effectiveness, followed by trust assessment, correctability, and task performance. The methods of assessing these measures also varied.
*Conclusion:* XAI research should address the lack of a comprehensive and agreed-upon framework for explaining XAI and standardized approaches for evaluating the effectiveness of the explanation that XAI provides to diverse AI stakeholders.

[*] Corresponding author. 103 Daehak-ro, Jongno-gu, Seoul, 03080, Republic of Korea.
*E-mail address:* ifilgood@snu.ac.kr (H. Kim).

https://doi.org/10.1016/j.heliyon.2023.e16110
Received 6 November 2022; Received in revised form 26 March 2023; Accepted 5 May 2023
Available online 8 May 2023

## 1. Introduction

With the revolution of data-intensive information technologies, there has been increased interest in the healthcare field in obtaining meaningful insights from the massive amount of clinical data through use of artificial intelligence (AI). Despite attempts to make AI models more intelligible, they remain opaque and are seldom employed in clinical practice [1–3]. The expectations of users have not been fulfilled due to a lack of transparency in AI models [2]. The challenges in interpreting the inner process of AI algorithms can lead to biased outcomes or even dangerous conclusions regarding patient care. For example, IBM Watson for Oncology (WFO) has been criticized for suggesting incorrect or harmful medical treatments [4,5].

According to the High-Level Expert Group (HLEG) on AI principles, development of a trustworthy AI system that will have a beneficial impact on human life requires transparency, human values, governance, and accountability at all stages [6–9]. In addition, the General Data Protection Regulation (GDPR) is an important example of Europe's growing demands regarding explainable artificial intelligence (XAI). Data security and privacy laws for all European Union (EU) residents and organizations are enforced by the GDPR using personal information for automated decision-making [10,11]. According to its mandate, organizations must disclose how the AI algorithm reaches ultimate decisions, so that users can detect any potential bias [12–14].

According to the Defense Advanced Research Projects Agency (DARPA), XAI is defined as AI systems that communicate reasoning to users, with identification of benefits and drawbacks for prediction of future behavior [13]. XAI techniques are intrinsically suited for explaining an AI model developed from massive amounts of complex medical data [15]. This unique aspect of XAI is expected to lead to acceleration of data-driven care [5].

Ensuring the safety of patients is the primary concern in decision-making processes involving AI in the field of healthcare [16]. However, due to concerns regarding the safety of the decisions generated, numerous AI models have failed to instill sufficient confidence in healthcare users [17]. Given that clinical data can be highly distorted and noisy, in addition to its explainability, the robustness of an AI model is fundamental for building the trust and acceptance of users [17,18]. The robustness of a model refers to its capacity to produce consistent and reliable results, even when there are minor variations in the input data [17,19,20].

Although the significance of explainability has generally been acknowledged, only a few studies have proposed criteria for XAI that add value to XAI models for users [21–23,23]. In particular, one study suggested three essential properties of XAI: stakeholders, objectives, and quality of personalized explanations (see Table 1) [22,25].

The DARPA's XAI program placed emphasis on evaluation of explanation effectiveness [13]. The DARPA also announced specific

**Table 1**
Synthesis criteria and definitions.

| Criterion | Definition |
|---|---|
| A. Essential Properties of XAI [22] | |
| a. Stakeholders of XAI | |
| AI regulators | Evaluate an AI system for certification under the legal requirements. |
| AI developers | Focus on performance optimization, debugging, and validation using a system engineering approach based on root-cause analysis. |
| AI managers | Control an AI system and supervise its compliance. |
| AI users | Understand explanations and compare the XAI decision-making process to know whether it is accurate, reliable, or trustworthy (e.g., physicians). |
| Individuals affected by AI-based decisions | Determine whether AI's decisions are acceptable with the help of explanations (e.g., patients). |
| b. Objectives of XAI | |
| Explainability to evaluate AI | Verifies if the system's behavior is sufficiently understood to uncover potential vulnerabilities. |
| Explainability to justify AI | Acquires essential knowledge necessary to justify AI, while making illogical judgments. |
| Explainability to improve AI | Pursues to improve AI by gaining more excellent knowledge of its inner workings and enhancing the system's accuracy and utility. |
| Explainability to learn from AI | Enables learning from AI when explanations acquire a lot of information on how AI works and what it achieves. |
| Explainability to manage AI | Facilitates the implementation of AI into organizational processes and its use in work routines. |
| c. Quality of Personalized Explanations | |
| Fidelity | Describes how closely the explanations adhere to the models' input-output mapping. |
| Generalizability | Indicates the range of a model that an XAI approach can be explained or applied to. |
| Explanatory Power | Demonstrates a scope of possible queries that can be responded to. |
| Interpretability | Explains the degree to which an explanation is understandable to users. |
| Comprehensibility | Refers to the objective ability of an explanation to help a user complete a task. |
| Plausibility | Refers to understanding as a subjective measure to accept the explained information. |
| Effort | Indicates the work necessary to understand an explanation. |
| Privacy | The extent to which information is obtained, kept, and used. |
| Fairness | The measure of how equitable an explanation can be delivered. |
| B. Measures of Explanation Effectiveness [13] | |
| Mental model | Known as the user's understanding to examine how users understand an XAI model. |
| User satisfaction | Identifies the explanations' clarity and utility. |
| Trust assessment | Assesses reliability that impacts the user's performance when hard-edge scenarios are present. |
| Task performance | Enhances the user's decisions and task accomplishment. |
| Correctability | Detects errors and evaluates explanations' correctness, completeness, and consistency. |

criteria for measurement of explanation effectiveness: mental model, user satisfaction, trust assessment, task performance, and correctability (see Table 1) [13]. However, these criteria have rarely been applied to the healthcare sector. The lack of thorough evaluations of the explanation effectiveness of XAI models has delayed the broad adoption of XAI techniques [26,27].

A systematic review of XAI studies in the field of healthcare that evaluated explanation effectiveness was conducted. The primary purpose of this study was to examine the essential properties of XAI and the methods used for evaluation of explanation effectiveness, based on criteria suggested in previous studies (see Table 1) [13,22]. This study was conducted in order to provide a broad overview of the trends and gaps in development of a successful XAI model in a clinical setting.

## 2. Material and methods

### 2.1. Data sources and search strategy

This systematic review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [28]. A search of the PubMed and Embase databases for potentially relevant studies published between January 1, 2011, and April 30, 2022, was conducted. A broad range of keyword-based queries were created for searching titles and abstracts (see Appendix A). Duplicates were eliminated after the retrieved articles had been exported into the EndNote reference manager.

### 2.2. Inclusion and exclusion criteria

All primary source studies found to be in accordance with the inclusion criteria were initially included: (1) year of publication; (2) XAI studies in healthcare; (3) studies on development of an XAI model using clinical data; and (4) studies that evaluated explanation effectiveness. Studies that met at least one of the following exclusion criteria were eliminated: (1) duplicate publication; (2) non-English-language; (3) non-peer-reviewed primary articles; and (4) review papers.

### 2.3. Screening

Screening of the title/abstract and the full-text article was performed independently by two reviewers (JJ and HL), and discrepancies were resolved by reaching a consensus between the two reviewers.

### 2.4. Data synthesis and analysis

A review of the papers that passed the screening process was performed for identification of the essential properties of XAI and the measures of explanation effectiveness adopted in the studies based on the criteria shown in Table 1. In addition, we attempted to determine whether a particular property of XAI was associated with a specific measure of explanation effectiveness.

## 3. Results

### 3.1. Overview of search results

The literature selection process is depicted in Fig. 1. After conducting a full-text screening based on the inclusion and exclusion criteria, only six articles met the purpose of this review as shown in Table 2, out of the initial 586 unique articles. The 340 articles that did not mention XAI models or clinical use in the title or abstract were excluded during the title/abstract screening process. In five studies, the aim was to develop an explanatory model to assist healthcare providers in diagnosing a patient's disease [29–33]. The aim of the remaining study was to examine the explainability of the model in order to identify variables that have an important influence on the prediction results [34]. Four of the six studies utilized publicly available clinical data [29–32], and the other two studies utilized electronic health record (EHR) data [33,34]. The studies mainly used image and multimedia data, such as computed tomography (CT) scans [33], pathogen images [32], electrocardiogram (ECG) [29,31], ultrasound videos [30], and operation videos [34].

Various types of AI algorithms, including the rule-based algorithm [29], K-Nearest Neighbors (KNN) [31], Convolutional Neural Network (CNN) [30–32], conventional image processing algorithm [33], and Support Vector Machine (SVM) [34] were used in the studies. The performance of the algorithms was presented using accuracy, balanced accuracy, sensitivity or recall, specificity, Positive Predictive Value (PPV) or precision, Negative Predictive Value (NPV), F1-score, Area Under the Receiver Operating Characteristic Curve (AUROC), Area Under the Precision-Recall Curve (AUPRC), Matthew's Correlation Coefficient (MCC), and error rate.

The studies adopted the following explanation approaches: the Permutation Feature Importance (PFI) [31], Local Interpretable Model-agnostic Explanation (LIME) [31], SHappley Additive exPlanation (SHAP) [31], faster Region with Convolutional Neural Network (R-CNN) [33], Class Activation Map (CAM) [30], Cumulative Fuzzy Class Membership Criterion (CFCMC) [32], pseudo-coloring methodology [29], and value permutation and feature-object semantics [34]. Brief definitions of these explanation approaches are included in Appendix B.

### 3.2. *Essential properties of XAI*

#### 3.2.1. *Stakeholders of XAI*

All six studies included an evaluation of the explainability primarily for physicians as the target AI users (Table 3) [30–34]. Only one study included comments from patients, *individuals affected by AI-based decisions*, by presenting the explanations for the XAI model [29].

#### 3.2.2. *Objectives of XAI*

An illustration of four of the five objectives of XAI is shown in Table 1 – *explainability to evaluate AI, explainability to justify AI, explainability to improve AI, and explainability to learn from AI* – were relevant to all six studies [29–34]. In two studies, explanations of the XAI model were used for evaluation of the AI decision-making process, determining how the AI algorithm distinguishes between normal and abnormal ECG rhythms [29,31]. In the other studies, analysis of images and video clips was performed for diagnosis of a specific disease or to predict a particular outcome in order to explain the rationales [30,32–34]. The aim of all of the studies was to improve AI by providing more convincing explanations; the intention was to attain additional knowledge through discovery of hidden information [29–34]. None of the studies attempted to assess the *explainability to manage AI* through integration of XAI models into actual clinical practice.

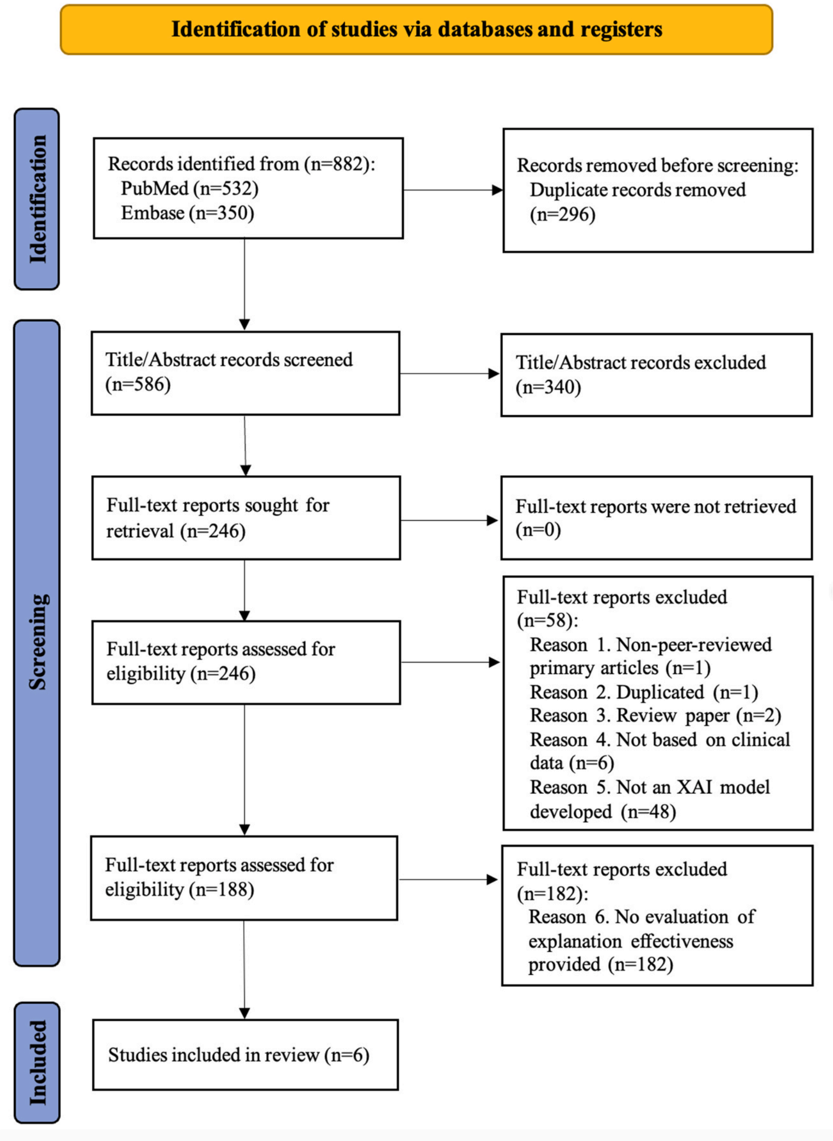
Identification of studies via databases and registers
Identification
Screening
Included
Records identified from (n=882):
PubMed (n=532)
Embase (n=350)
Records removed before screening:
Duplicate records removed
(n=296)
Title/Abstract records screened
(n=586)
Title/Abstract records excluded
(n=340)
Full-text reports sought for retrieval (n=246)
Full-text reports were not retrieved
(n=0)
Full-text reports assessed for eligibility (n=246)
Full-text reports excluded
(n=58):
Reason 1. Non-peer-reviewed primary articles (n=1)
Reason 2. Duplicated (n=1)
Reason 3. Review paper (n=2)
Reason 4. Not based on clinical data (n=6)
Reason 5. Not an XAI model developed (n=48)
Full-text reports assessed for eligibility (n=188)
Full-text reports excluded
(n=182):
Reason 6. No evaluation of explanation effectiveness provided (n=182)
Studies included in review (n=6)

**Fig. 1.** The PRISMA flow presents identification, screening, and inclusion.

**Table 2**
Summary of final articles.

| Authors, Year | Research aims | Data resources | Data types | Input | Output | AI methods | AI Performance metrics | Explainable techniques |
|---|---|---|---|---|---|---|---|---|
| Alahmadi, A. et al., 2021 [29] | To develop an explainable rule-based decision tree classification model to automate the detection of QT-prolongation at risk of Torsades de Pointes (TdP) | Public dataset, clinical trial approved by Food and Drug Administration (FDA) in 2014 | ECG image data | ECG | Classification of Torsade de Pointes (TdP) | Rule-based algorithm | Accuracy<br>Balance<br>Sensitivity<br>Specificity<br>PPV<br>F1-score<br>ROC (AUC)<br>Precision-Recall (AUC)<br>MCC<br>Error rate | Pseudo-coloring methodology |
| Born, J. et al., 2021 [30] | To develop an explainable classification model for differential COVID-19 diagnosis | Public dataset, Lung Point-Of-Care Ultrasound (POCUS) | Ultrasound video data | Ultrasound | Classification of COVID-19 | CNN | Precision<br>Recall<br>F1-score<br>Specificity<br>MCC | CAM |
| Neves, I. et al., 2021 [31] | To develop an explainable ECG classification model on time series | Public dataset, Massachusetts Institute of Technology-Beth Israel Hospital (MIT-BIH) Arrhythmia | ECG image data | ECG | Classification of arrhythmia | KNN<br>CNN | F1-score<br>Precision<br>Recall<br>AUC | PFI<br>LIME<br>SHAP |
| Sabol, P. et al., 2020 [32] | To develop an explainable classification model for colorectal cancer diagnosis | Public dataset, Colorectal cancer pathology image | Histopatho-logical image data | Colorectal cancer pathology image data | Classification of colorectal cancer | CNN | Accuracy<br>Precision<br>Recall<br>F1-score | CFCMC |
| Tan, W. et al., 2021 [33] | To develop an explainable deep learning model for the automatic diagnosis of fenestral OS | EHR data, the Fudan University | CT scan image data | Temporal bone high-resolution computed tomography (HRCT) | Classification of fenestral otosclerosis | Conventional image processing algorithm | Accuracy<br>Sensitivity<br>Specificity<br>PPV<br>NPV | Faster-RCNN |
| Derathé, A. et al., 2021 [34] | To explain the previously developed prediction model for surgical practice quality | EHR data, the CHU Grenoble Alpes Hospital | Laparoscopic sleeve gastrectomy (LSG) operation video data | Laparoscopic operation videos | Extraction of the most important variables to predict the quality of surgical practice | SVM | Accuracy<br>Sensitivity<br>Specificity | Value-permutation and Feature-object semantics |

**Table 3**
Assessments of the essential properties of XAI and the measures of explanation effectiveness.

| | | | [29] | [30] | [31] | [32] | [33] | [34] |
|---|---|---|---|---|---|---|---|---|
| Essential properties of XAI | Stakeholders of XAI | AI regulators | | | | | | |
| | | AI developers | | | | | | |
| | | AI managers | | | | | | |
| | | AI users | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | | Individuals affected by AI-based decisions | ✓ | | | | | |
| | Objectives of XAI | Explainability to evaluate AI | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | | Explainability to justify AI | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | | Explainability to improve AI | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | | Explainability to learn from AI | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | | Explainability to manage AI | | | | | | |
| | Quality of personalized explanations | Fidelity | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | | Generalizability | N/A | N/A | N/A | N/A | N/A | N/A |
| | | Explanatory power | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | | Interpretability | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | | Comprehensibility | | | | | ✓ | |
| | | Plausibility | ✓ | ✓ | ✓ | ✓ | | ✓ |
| | | Effort | | | ✓ | | | |
| | | Privacy | | | | | | |
| | | Fairness | | | | | | |
| Measures of explanation effectiveness | | Mental model | | | | | | |
| | | User satisfaction | ✓ | ✓ | ✓ | ✓ | | ✓ |
| | | Trust assessment | ✓ | | | ✓ | | |
| | | Task performance | | | ✓ | ✓ | ✓ | |
| | | Correctability | | ✓ | | ✓ | | ✓ |

#### 3.2.3. *Quality of personalized explanations*

The quality of personalized explanations can be measured using *fidelity, generalizability, explanatory power, interpretability, comprehensibility, plausibility, effort, privacy, and fairness* (Table 1). The accuracy, precision, and recall of AI algorithms were examined for assessment of *fidelity* [29–34]. Determination of *interpretability* and *explanatory power* was based on how well an AI model answered questions in a human-understandable way [29–34]. *Plausibility* was confirmed mainly by scoring of the acceptance of XAI explanations by users [29–32,34]. By contrast, determination of *comprehensibility* was based on the average accuracy of the user's decision performance using the AI model with and without descriptions of XAI [33]. The time required to perform a task is an indication of the user's *effort* to understand explanations of XAI [31]. None of the XAI studies included discussion of *privacy* or *fairness*. Because the six studies were designed for application of known explainable techniques rather than development of new ones, *generalizability* was non-applicable.

### 3.3. *Evaluation of explanation effectiveness*

The five criteria were used in measurement of explanation effectiveness, as shown in Table 3. The most examined measure was *user satisfaction* [29–32,34]. Three articles evaluated *task performance* [31,32,34] and *correctability* [30,32,34], and two articles evaluated *trust assessment* [29,32]. None of the six studies included consideration of a *mental model*. A summary of the details regarding the approaches to evaluation of explanation effectiveness is shown in Table 4. A questionnaire survey was used for assessment of the task completion time, performance accuracy, usefulness, and typicalness of explanations [30–32,34]. One study compared the diagnostic performance of the AI model, clinicians, and clinicians with AI assistance [33]. Opinions on explanations of XAI were discussed openly in focus groups or using open-ended questions in four studies [29–31,34].

### 3.4. *Association between properties of XAI and measures of explanation effectiveness*

The association between an XAI property of interest and a measure of explanation effectiveness that was applied is shown in Table 5. The most common measure for *user satisfaction* and *trust assessment* was *plausibility* of participants' satisfaction and reliability with explanations of XAI [29–32,34]. However, examination of multiple qualities of personalized explanations, including *fidelity, comprehensibility, plausibility,* and *effort,* was performed for measurement of *task performance* [31–33]. Users were asked to identify errors in XAI explanations that need correcting for evaluation of *correctability* through *fidelity* [30,32,34].

## 4. Discussion

XAI has captured the attention of professionals in the field of healthcare, resulting in the creation of numerous XAI models. XAI studies have focused on improving state-of-the-art explainable techniques [35]. However, there is an urgent need for XAI models that can be applied to real-world practice [36–38]. This systematic review examined the essential properties of XAI and included an evaluation of explanation effectiveness in order to inform efforts to facilitate the use of XAI models in clinical practice.

**Table 4**
Summary of evaluating explanation effectiveness.

| | Participant | Methods | Measurement/Instruments | Limitation/Future work |
|---|---|---|---|---|
| Alahmadi, A. et al., 2021 [29] | 7 cancer patients,<br>2 female nurses,<br>1 male physician | Two focus group discussions | Understandability,<br>Usefulness,<br>Reliability<br>: Guidance for Reporting Involvement of Patients and Public (GRIPP) reporting checklists | The focus groups included a limited number of participants mentioning the need for a more diversified clinical population. Evaluating explanation effectiveness is essential to demonstrate the technique's applicability in clinical practice. |
| Born, J. et al., 2021 [30] | 2 physicians | A user study<br>: a questionnaire and comments | Scoring of −3 (the heatmap is only distracting) to 3 (the heatmap is very helpful for diagnosis),<br>The average ratio of correctly explained patterns | Explanation parts of the incorrectly highlighted visible patterns were detected. |
| Neves, I. et al., 2021 [31] | 1 expert cardiologist,<br>1 graduate medical student,<br>1 resident | A user study<br>: an online questionnaire in random order of 20 ECGs and free-text comments | Performance accuracy,<br>Task completion time,<br>Usefulness levels: a 5-point scale,<br>Typicalness levels: a 5-point scale | There was a lack of agreement on evaluating the quality of explanations and usefulness of model outputs. |
| Sabol, P. et al., 2020 [32] | 14 pathologists | A clinical trial<br>: a questionnaire | Objectivity,<br>Details,<br>Reliability,<br>Quality<br>: average score | The broad experiment to include other pathologists from varied domains was necessary. |
| Tan, W. et al., 2021 [33] | 2 chief physicians,<br>3 associate chief physicians,<br>1 attending physician,<br>1 resident | An experiment for the diagnostic performance assisted by the LNN model comparing to otosclerosis-LNN, otologists, and XAI-assisted otologists | Average accuracy,<br>Sensitivity,<br>Specificity | During the experiment, otologists often combine clinical diagnoses utilizing diverse patient information, including CT scans, clinical complaints, medical records, and audiological examinations. |
| Derathé, A. et al., 2021 [34] | 6 experienced digestive surgeons | A survey<br>: a questionnaire and comments | Level of agreement with the statement for each surgeon<br>: a 5-level Likert scale | Due to the ambiguity of the survey questions, the respondents provided responses that were inconsistent with the question's intent. |

### 4.1. Accommodating more diverse stakeholders of XAI

We found that the reported stakeholders of XAI in the field of healthcare were not a monolithic group, rather each group had different expectations for an XAI model [39]. According to the findings of this review, *AI users* (e.g., physicians, nurses, and pharmacists) were the primary focus, although fulfillment of several legal, legislative, ethical, social, and technical prerequisites will be required before XAI models can be employed in clinical practice [40–42]. Based on these entry requirements, close cooperation between *AI managers, AI regulators,* and *AI developers* is required. In addition, contemporary *individuals affected by AI-based decisions* (e.g., patients, families, and caregivers) can accept innovative medical technologies while requiring adequate explanations for the treatments [43]. Nonetheless, providing user-friendly explanations of XAI that can be easily understood by non-experts remains challenging [39,44].

### 4.2. Extended objectives of XAI

Although application of XAI models within the hospital system has an impact on real-time EHR systems, no studies included in this review examined the effects of XAI integration on a clinical pathway. Uncertainty regarding the infrastructure for clinical deployment is a significant obstacle to the *explainability to manage AI*, along with a lack of access to big data or a lack of engagement with clinical workflows [45,46]. Furthermore, there are no agreed-upon objectives of XAI, and the criteria presented in this study do not include all critical objectives regarding use of XAI in the healthcare system. For example, explanations of XAI can be focused on prevention of harmful consequences. Acquiring additional information from explanations of XAI ensures transparency so that errors can be avoided [47]. Thus, adding the objective of *explainability to control AI* can prevent erroneous outcomes and enable debugging, which reduces the risk of patient harm [24].

### 4.3. Including data privacy and AI bias in the quality of XAI explanations

The two requirements – *privacy and fairness* – were outlined as complimentary notions in the HLEG presented by the Ethics Guidelines for Trustworthy AI in 2019 [48,49]. In this review, personal information obtained from clinical data was used for development of XAI models. However, none of the studies included an attempt to determine how easily sensitive data could be inferred or to

**Table 5**
Association between XAI properties and explanation effectiveness measures.

| Essential properties of XAI | | Measures of explanation effectiveness | | | | |
|---|---|---|---|---|---|---|
| | | Mental model | User satisfaction | Trust assessment | Task performance | Correctability |
| Stakeholders of XAI | AI regulators | | | | | |
| | AI developers | | | | | |
| | AI managers | | | | | |
| | AI users | | [29]<br>[30]<br>[31]<br>[32]<br>[33] | [32] | [31]<br>[32]<br>[33] | [30]<br>[32]<br>[34] |
| | Individuals affected by AI-based decisions | | [29] | [29] | | |
| Objectives of XAI | Explainability to evaluate AI | | | | | [30]<br>[32]<br>[34] |
| | Explainability to justify AI | | | [29]<br>[32] | | |
| | Explainability to improve AI | | [29]<br>[30]<br>[31]<br>[32]<br>[34] | | [31]<br>[32]<br>[33] | |
| | Explainability to learn from AI | | | | | |
| | Explainability to manage AI | | | | | |
| Quality of personalized explanations | Fidelity | | | | [33] | [30]<br>[32]<br>[34] |
| | Generalizability | N/A | N/A | N/A | N/A | N/A |
| | Explanatory power | | | | | |
| | Interpretability | | | | | |
| | Comprehensibility | | | | [33] | |
| | Plausibility | | [29]<br>[30]<br>[31]<br>[32]<br>[34] | [29]<br>[32] | [32] | |
| | Effort | | | | [31] | |
| | Privacy | | | | | |
| | Fairness | | | | | |

what degree XAI explanations were secured. In addition, discovery of health inequalities resulting from a biased dataset that affects minorities and for application of equitable treatments is difficult due to the complex nature of health data [50,51]. A previous study demonstrated that XAI techniques can be used for detection of bias [52]. For example, a visual analytic system known as FariSight was developed for determination of fairness in AI decisions and to explain implicit bias [53]. In other words, measurement of *fairness* can be performed using XAI techniques to safeguard against potential discrimination in the healthcare context [54]. Therefore, *fairness* of the AI model should be regarded as an important quality of the explanation generated by an XAI model in pursuit of health equity.

### *4.4. Requiring a standardized approach to the evaluation of explanation effectiveness of XAI*

Non-standardized, heterogeneous methods were used for evaluation of explanation effectiveness in all screened studies. A different measurement was utilized in each study for evaluation of explanation effectiveness. The findings of this study, for example, demonstrated that in evaluation of the *trust assessments*, a questionnaire with scoring from 1 to 6 was used in one study [32], and another study asked for open comments [29]. Furthermore, although the *mental model* is one of the most important criteria for measurement of explanation effectiveness, particularly when using XAI techniques on users who were unfamiliar with an XAI model, it was not considered in any studies [52,55].

### *4.5. Limitations*

Our systematic review has several limitations. First, because the majority of relevant research on XAI has neglected to evaluate the effectiveness of the explanation, this review included only six studies. While we were able to assess the essential properties of XAI and the measures of explanation effectiveness, six studies were not sufficient for generalization of its characteristics in research on XAI. Second, because there was no consensus regarding the essential properties of XAI in the field of healthcare, the screened papers were categorized according to the suggested properties of XAI based on the previous research.

## 5. Conclusion

This study included a critical review of existing literature on creation of XAI models using clinical data as well as an evaluation of explanation effectiveness. The screened papers were evaluated for the essential properties of XAI and the measures of explanation effectiveness. Understanding of the complex aspects of XAI is required for implementation of XAI models in healthcare services, which may go beyond the properties presented in this review. In addition, ensuring explanation effectiveness is particularly important for implementation of XAI in the field of healthcare. Additional efforts are needed for development of a comprehensive and agreed-upon framework that explains the core properties of XAI, along with standardized approaches to measurement of the effectiveness of explanation produced by an XAI model.

## Author contribution statement

JJ and HL conducted the primary literature search and reviewed the literature. JJ, HL, HJ, and HK analyzed screened papers. JJ wrote the first draft of the manuscript, and HJ and HK provided guidance and amendments. HK edited the manuscript. All authors reviewed the content.

## Data availability statement

Data will be made available on request.

## Declaration of competing interest

The authors have no interests to declare.

## Acknowledgments

This research was supported by the BK21 four project (Center for Human-Caring Nurse Leaders for the Future) funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF).

## Appendix A. Keyword-based queries.

| Database | Query | Results |
|---|---|---|
| PubMed | “explainable artificial intelligence" [Title/Abstract] OR “xai" [Title/Abstract] OR “explainable ai" [Title/Abstract] OR “interpretable ai" [Title/Abstract] OR “interpretable artificial intelligence" [Title/Abstract] | 532 |
| Embase | ‘explainable artificial intelligence’:ab,ti OR xai:ab,ti OR ‘explainable ai’:ab,ti OR ‘interpretable ai’:ab,ti OR ‘interpretable artificial intelligence’:ab,ti | 350 |

## ***Appendix B.*** *Definitions of explainable techniques.*

| Explainable techniques | Mechanism |
|---|---|
| Permutation Feature Importance (PFI) | The PFI is a technique for overall interpretability by examining the model score after shuffling a single feature value [31]. |
| Local Interpretable Model-agnostic Explanation (LIME) | The LIME is a perturbation-based strategy that uses a surrogate interpretable model to substitute the complex model locally, providing local interpretability [31]. |
| SHappley Additive exPlanation (SHAP) | The SHAP is a method for determining how each feature contributes to a specific outcome [31]. |
| Faster Region with Convolutional Neural Network (R-CNN) | The faster R-CNN presented the Region Proposal Network [RPN], which speeds up the selective search. RPN adheres to the last convolution layer of CNN. Proposals from RPN are given to a region of interest pooling (RoI pooling), then classification and bounding-box regression [56]. |
| Pseudo-coloring methodology | The pseudo-coloring methodology employs a range of colors to represent continuously changing values [29]. |
| Class Activation Map (CAM) | The CAM uses global average pooling to generate class-specific heatmaps that indicate discriminative regions [30]. |
| Value permutation and Feature-object semantics | The permutation of values is analyzed for their impact on predictions, and the most significant variables are then translated into statements using feature-object semantics [34]. |
| Cumulative Fuzzy Class Membership Criterion (CFCMC) | The CFCMC offers a confidence measure for a test image’s classification, followed by a representation of the training image and the most similar images [32]. |

## References

[1] L.H. Gilpin, D. Bau, B.Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: an overview of interpretability of machine learning, in: IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2018, pp. 80–89, https://doi.org/10.1109/DSAA.2018.00018.

[2] R. Kocielnik, S. Amershi, P.N. Bennett, Will you accept an imperfect AI? Exploring designs for adjusting end-user expectations of AI systems, Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, J. ACM (2019) 1–14, https://doi.org/10.1145/3290605.3300641.

[3] D. Shin, The effects of explainability and causability on perception, trust, and acceptance: implications for explainable AI, Int. J. Hum. Comput. Stud. 146 (2021), 102551, https://doi.org/10.1016/j.ijhcs.2020.102551.

[4] V. Harish, F. Morgado, A.D. Stern, S. Das, Artificial intelligence and clinical decision making: the new nature of medical uncertainty, Acad. Med. 96 (2020) 31–36, https://doi.org/10.1097/ACM.0000000000003707.

[5] G. Yang, Q. Ye, J. Xia, 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 (2022) 29–52, https://doi.org/10.1016/j.inffus.2021.07.016.

[6] European Commission. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Artificial Intelligence for Europe, Brussels, 2018. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237&from=EN/ (Accessed 23 June 2022).

[7] S. Larsson, On the governance of artificial intelligence through ethics guidelines, Asian J. Law Soc. 7 (2020) 437–451, https://doi.org/10.1017/als.2020.19.

[8] N. Palladino, The role of epistemic communities in the "constitutionalization" of internet governance: the example of the European Commission High-Level Expert Group on Artificial Intelligence, Telecommun. Pol. 45 (2021) 102–149, https://doi.org/10.1016/j.telpol.2021.102149.

[9] M. Veale, A critical take on the policy recommendations of the EU high-level expert group on artificial intelligence, Eur. J. Risk Regul. 11 (2020) 1–10, https://doi.org/10.1017/err.2019.65.

[10] G. Bodea, K. Karanikolova, D.K. Mulligan, J. Makagon. Automated Decision-Making on the Basis of Personal Data that Has Been Transferred from the EU to Companies Certified under the EU-US Privacy Shield: Fact-Finding and Assessment of Safeguards provided by US Law, European Commission, 2018. https://ec.europa.eu/info/sites/default/files/independent_study_on_automated_decision-making.pdf/ (Accessed 15 July 2022).

[11] D. Pedreschi, F. Giannotti, R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, Meaningful explanations of black box AI decision systems, Proc. AAAI Conf. Artif. Intell. 33 (2019) 9780–9784, https://doi.org/10.1609/aaai.v33i01.33019780.

[12] A. Asatiani, P. Malo, P.R. Nagbøl, E. Penttinen, T. Rinta-Kahila, A. Salovaara, Challenges of explaining the behavior of black-box AI systems, MIS Q. Exec. 19 (2020) 259–278, https://doi.org/10.17705/2msqe.00037.

[13] D. Gunning, D. Aha, DARPA's explainable artificial intelligence (XAI) program, AI Mag. 40 (2019) 44–58, https://doi.org/10.1609/aimag.v40i2.2850.

[14] A.J. Wulf, O. Seizov, Please Understand We Cannot Provide Further Information": Evaluating Content and Transparency of GDPR-Mandated AI Disclosures, AI and Society, 2022, pp. 1–22, https://doi.org/10.1007/s00146-022-01424-z.

[15] K.W. Johnson, J. Torres Soto, B.S. Glicksberg, K. Shameer, R. Miotto, M. Ali, E. Ashley, J.T. Dudley, Artificial intelligence in cardiology, J. Am. Coll. Cardiol. 71 (2018) 2668–2679, https://doi.org/10.1016/j.jacc.2018.03.521.

[16] A. Čartolovni, A. Tomičić, EL. Mosler Ethical, legal, and social considerations of AI-based medical decision-support tools: a scoping review, 104737, Int. J. Med. Inf. 161 (2022) 104738.

[17] A. Holzinger, M. Dehmer, F. Emmert-Streib, R. Cucchiara, I. Augenstein, J. Del Ser, et al., Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence, Inf. Fusion 79 (2022) 263–278, https://doi.org/10.1016/j.inffus.2021.10.007.

[18] A. Holzinger, The Next Frontier: AI We Can Really Trust. Machine Learning and Principles and Practice of Knowledge Discovery in Databases, International Workshops of ECML PKDD 2021, Virtual Event, Springer, 2022, pp. 427–440, https://doi.org/10.1007/978-3-030-93736-2_33.

[19] Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang, J. Jorge. Counterfactuals and causability in explainable artificial intelligence: theory, algorithms, and applications, Inf. Fusion 81 (2022) 59–83, https://doi.org/10.1016/j.inffus.2021.11.003.

[20] N. Gozzi, L. Malandri, F. Mercorio, A. Pedrocchi, XAI for myo-controlled prosthesis: explaining EMG data for hand gesture classification, Knowl. Base Syst. 240 (2022), 108053, https://doi.org/10.1016/j.knosys.2021.108053.

[21] A.F. Markus, J.A. Kors, P.R. Rijnbeek, The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies, J. Biomed. Inf. 113 (2021), 103655, https://doi.org/10.1016/j.jbi.2020.103655.

[22] C. Meske, E. Meske, J. Schneider, M. Gersch, Explainable artificial intelligence: objectives, stakeholders, and future research opportunities, Inf. Syst. Manag. 39 (2022) 53–63, https://doi.org/10.1080/10580530.2020.1849465.

[23] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160, https://doi.org/10.1109/access.2018.2870052.

[24] G. Vilone, L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, Inf. Fusion 76 (2021) 89–106, https://doi.org/10.1016/j.inffus.2021.05.009.

[25] J. Schneider, J. Handali, Personalized Explanation in Machine Learning: A Conceptualization, 27th European Conference on Information Systems, 2019, pp. 1–17, https://doi.org/10.48550/arXiv.1901.00770.

[26] S. Mohseni, N. Zarei, E.D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable AI systems, ACM Trans. Interact. Intell. Syst. 11 (2021) 1–45, https://doi.org/10.1145/3387166.

[27] C. Mal, Understanding Explainable AI: Role in IoT-Based Disease Prediction and Diagnosis, Medical Internet of Things, Chapman and Hall/CRC, 2021, pp. 205–218, https://doi.org/10.1201/9780429318078.

[28] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, et al., The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, Syst. Rev. 89 (2021) 1–11, https://doi.org/10.1186/s13643-021-01626-4.

[29] A. Alahmadi, A. Davies, J. Royle, L. Goodwin, K. Cresswell, Z. Arain, et al., An explainable algorithm for detecting drug-induced QT-prolongation at risk of torsades de pointes (TdP) regardless of heart rate and T-wave morphology, Comput. Biol. Med. 131 (2021), 104281, https://doi.org/10.1016/j.compbiomed.2021.104281.

[30] J. Born, N. Wiedemann, M. Cossio, C. Buhre, G. Brändle, K. Leidermann, Accelerating detection of lung pathologies with explainable ultrasound image analysis, Appl. Sci. 11 (2021) 672, https://doi.org/10.3390/app11020672.

[31] I. Neves, D. Folgado, S. Santos, M. Barandas, A. Campagner, L. Ronzio, F. Cabitza, H. Gamboa, Interpretable heartbeat classification using local model-agnostic explanations on ECGs, Comput. Biol. Med. 133 (2021), 104393, https://doi.org/10.1016/j.compbiomed.2021.104393.

[32] P. Sabol, P. Sinčák, P. Hartono, P. Kočan, Z. Benetinová, A. Blichárová, et al., Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images, J. Biomed. Inf. 109 (2020), 103523, https://doi.org/10.1016/j.jbi.2020.103523.

[33] W. Tan, P. Guan, L. Wu, H. Chen, J. Li, Y. Ling, et al., The use of explainable artificial intelligence to explore types of fenestral otosclerosis misdiagnosed when using temporal bone high-resolution computed tomography, Ann. Transl. Med. 9 (2021) 969, https://doi.org/10.21037/atm-21-1171.

[34] A. Derathé, F. Reche, P. Jannin, A. Moreau-Gaudry, B. Gibaud, S. Voros, Explaining a model predicting quality of surgical practice: a first presentation to and review by clinical experts, Int. J. Comput. Assist. Radiol. Surg. 16 (2021) 2009–2019, https://doi.org/10.1007/s11548-021-02422-0.

[35] S.N. Payrovnaziri, Z. Chen, P. Rengifo-Moreno, T. Miller, J. Bian, J. H Chen, et al., Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review, J. Am. Med. Inf. Assoc. 27 (2020) 1173–1185, https://doi.org/10.1093/jamia/ocaa053.

[36] A. Adadi, M. Berrada, Explainable AI for Healthcare: from Black Box to Interpretable Models, Embedded Systems and Artificial Intelligence, vol. 1076, Springer, 2020, pp. 327–337, https://doi.org/10.1007/978-981-15-0947-6_31.

[37] O.I. Dauda, J.B. Awotunde, M. AbdulRaheem, S.A. Salihu, Basic Issues and Challenges on Explainable Artificial Intelligence (XAI) in Healthcare Systems, Principles and Methods of Explainable Artificial Intelligence in Healthcare, 2022, pp. 248–271, https://doi.org/10.4018/978-1-6684-3791-9.ch011.

[38] J. Gerlings, M.S. Jensen, A. Shollo, Explainable AI, but explainable to whom? An exploratory case study of XAI in healthcare, Handb. Artif. Intell. Healthcare: Springer 212 (2022) 169–198, https://doi.org/10.1007/978-3-030-83620-7_7.

[39] J. Souza, C.K. Leung, Explainable Artificial Intelligence for Predictive Analytics on Customer Turnover: A User-Friendly Interface for Non-expert Users, Explainable AI within the Digital Transformation and Cyber Physical Systems, Springer, 2021, pp. 47–67, https://doi.org/10.1007/978-3-030-76409-8_4.

[40] J. Amann, A. Blasimme, E. Vayena, D. Frey, V.I. Madai, Explainability for artificial intelligence in healthcare: a multidisciplinary perspective, BMC Med. Inf. Decis. Making 20 (2020) 1–9, https://doi.org/10.1186/s12911-020-01332-6.

[41] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, et al., What do we want from Explainable Artificial Intelligence (XAI) –A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research, Artif. Intell. 296 (2021), 103473, https://doi.org/10.1016/j.artint.2021.103473.

[42] C. Panigutti, A. Monreale, G. Comandé, D. Pedreschi, Ethical, Societal and Legal Issues in Deep Learning for Healthcare, Deep Learning in Biology and Medicine, World Scientific, 2022, pp. 265–313, https://doi.org/10.1142/9781800610941_0009.

[43] S. Timmermans, M. Berg, The practice of medical technology, Sociol. Health Illness 25 (2003) 97–114, https://doi.org/10.1111/1467-9566.00342.

[44] C. Bove, J. Aigrain, M.-J. Lesot, C. Tijus, M. Detyniecki, Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users, 27th International Conference on Intelligent User Interfaces (2022) 807–819, https://doi.org/10.1145/3490099.3511139.

[45] S. Reddy, S. Allan, S. Allan, P. Cooper, A governance model for the application of AI in health care, J. Am. Med. Inf. Assoc. 27 (2020) 491–497, https://doi.org/10.1093/jamia/ocz192.

[46] L. Rundo, R. Pirrone, S. Vitabile, E. Sala, O. Gambino, Recent advances of HCI in decision-making tasks for optimized clinical workflows and precision medicine, J. Biomed. Inf. 108 (2020), 103479, https://doi.org/10.1016/j.jbi.2020.103479.

[47] E.M. Kenny, C. Ford, M. Quinn, M.T. Keane, Explaining black-box classifiers using post-hoc explanations-by-example: the effect of explanations and error-rates in XAI user studies, Artif. Intell. 294 (2021), 103459, https://doi.org/10.1016/j.artint.2021.103459.

[48] P. Ala-Pietilä, Y. Bonnet, U. Bergmann, M. Bielikova, C. Bonefeld-Dahl, W. Bauer, et al.. The Assessment List for Trustworthy Artificial Intelligence (ALTAI), European Commission, 2020. https://op.europa.eu/en/publication-detail/-/publication/73552fcd-f7c2-11ea-991b-01aa75ed71a1 (Accessed 10 March 2022).

[49] N.A. Smuha, The EU approach to ethics guidelines for trustworthy artificial intelligence, Comput. Law Rev. Int. 20 (2019) 97–106, https://doi.org/10.9785/cri-2019-200402.

[50] P. Ivaturi, M. Gadaleta, A.C. Pandey, M. Pazzani, S. R Steinhubl, G. Quer, A comprehensive explanation framework for biomedical time series classification, IEEE J. Biomed. Health Inform.c 25 (2021) 2398–2408, https://doi.org/10.1109/JBHI.2021.3060997.

[51] Y.S. Jeon, K. Yoshino, S. Hagiwara, A. Watanabe, S.T. Quek, H. Yoshioka, et al., Interpretable and lightweight 3-D deep learning model for automated ACL diagnosis, IEEE J. Biomed. Health Inform.c 25 (2021) 2388–2397, https://doi.org/10.1109/JBHI.2021.3081355.

[52] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, et al., Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115, https://doi.org/10.1016/j.inffus.2019.12.012.

[53] Y. Ahn, Y.-R. Lin, Fairsight: visual analytics for fairness in decision making, IEEE Trans. Visual. Comput. Graph. 26 (2019) 1086–1095, https://doi.org/10.1109/TVCG.2019.2934262.

[54] I. Palatnik de Sousa, M.M.B.R. Vellasco, E. Costa da Silva, Explainable artificial intelligence for bias detection in COVID CT-scan classifiers, Sensors 21 (2021) 5657, https://doi.org/10.3390/s21165657.

[55] A. Anderson, J. Dodge, A. Sadarangani, Z. Juozapaitis, E. Newman, J. Irvine, et al., Mental models of mere mortals with explanations of reinforcement learning, ACM Trans. Interact. Intell. Syst. 10 (2020) 1–37, https://doi.org/10.1145/3366485.

[56] D. Alamsyah, M. Fachrurrozi, Faster R-CNN with inception v2 for fingertip detection in homogenous background image, J. Phys. Conf. 1196 (2019), 122017, https://doi.org/10.1088/1742-6596/1196/1/012017.