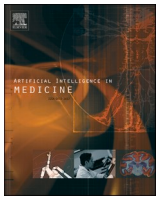
Artificial Intelligence In Medicine 143 (2023) 102616

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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



Research Paper

How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare

Johannes Allgaier [a,*], Lena Mulansky [a], Rachel Lea Draelos [b], Rüdiger Pryss [a]

[a] Institute of Clinical Epidemiology and Biometry, Julius-Maximilians-Universität Würzburg (JMU), Germany
[b] Cydoc, Durham, NC, United States of America

ARTICLE INFO

Keywords:
Explainable artificial intelligence
XAI
Interpretable machine learning
PRISMA
Medicine
Healthcare
Review

ABSTRACT

Background: Medical use cases for machine learning (ML) are growing exponentially. The first hospitals are already using ML systems as decision support systems in their daily routine. At the same time, most ML systems are still opaque and it is not clear how these systems arrive at their predictions.
Methods: In this paper, we provide a brief overview of the taxonomy of explainability methods and review popular methods. In addition, we conduct a systematic literature search on PubMed to investigate which explainable artificial intelligence (XAI) methods are used in 450 specific medical supervised ML use cases, how the use of XAI methods has emerged recently, and how the precision of describing ML pipelines has evolved over the past 20 years.
Results: A large fraction of publications with ML use cases do not use XAI methods at all to explain ML predictions. However, when XAI methods are used, open-source and model-agnostic explanation methods are more commonly used, with SHapley Additive exPlanations (SHAP) and Gradient Class Activation Mapping (Grad-CAM) for tabular and image data leading the way. ML pipelines have been described in increasing detail and uniformity in recent years. However, the willingness to share data and code has stagnated at about one-quarter.
Conclusions: XAI methods are mainly used when their application requires little effort. The homogenization of reports in ML use cases facilitates the comparability of work and should be advanced in the coming years. Experts who can mediate between the worlds of informatics and medicine will become more and more in demand when using ML systems due to the high complexity of the domain.

1. Introduction

Artificial intelligence (AI) in healthcare holds many opportunities and risks and has attracted great public interest. To date, however, experts involved in the development of machine learning (ML) systems come from diverse backgrounds, and the gap between ML engineers and healthcare providers, and often also other researchers, is wide. Methods that explain the predictions of complex algorithms in a user-friendly way can increase adoption and trust [1]. The use of ML systems for appropriate medical use cases has the potential to reduce costs, save time, increase treatment quality, and improve patient care.

ML systems can be categorized based on whether they can *replace* or *supplement* a healthcare provider. To date, there are no ML systems capable of replacing a healthcare provider; to the best of our knowledge, we did not find any system that appears to be sufficiently powerful or interpretable to operate safely without human supervision. In a few cases, ML systems are currently being used to *supplement* health care. Some of these are listed in an online database [2]. However, the companies using these AI systems in hospitals do not provide detailed information on their websites about whether these systems include explainability methods.

The medical specialties with the most ML activity are radiology and pathology, as they are both image-based and therefore ideally suited to recent advances in computer vision techniques [3]. ML systems applied to radiology images have the potential to reduce radiologist error rates by 3 to 5% [4] by alerting radiologists to potentially missed diagnoses, extend specialist expertise to under-supplied regions, where only one radiologist may be available for millions of patients [5], or improve triage by bringing scans with potentially urgent findings to the top of the physician's queue for earlier interpretation. In pathology, AI systems can

* Corresponding author.
E-mail addresses: johannes.allgaier@uni-wuerzburg.de (J. Allgaier), lena.mulansky@uni-wuerzburg.de (L. Mulansky), rachel.draelos@cydoc.ai (R.L. Draelos), ruediger.pryss@uni-wuerzburg.de (R. Pryss).

https://doi.org/10.1016/j.artmed.2023.102616
Received 26 September 2022; Received in revised form 22 February 2023; Accepted 15 May 2023
Available online 24 June 2023

speed up the interpretation of large slides by automatically identifying the most important areas for the pathologist to examine [6]. There is also interest in developing AI systems for dermatology [7], cardiology [8], genetics [9], intensive care [10], oncology [11], and gastroenterology [12]. In the future, ML systems focused on augmentation may influence administrative or research activities, such as chart review [13], in addition to clinical care. So we see that there are many approaches and good reasons to implement AI systems in the health care context. But is it also necessary to make AI systems explainable in this context? And if so, explainable to whom?

Is it necessary to make ML models explainable in medicine? Explainability of AI systems is not always necessary, or if the benefits outweigh the costs of explainability too much, then perhaps it can be dispensed with. For example, in logistics, if a package is occasionally misclassified, and therefore sent somewhere else, this need not be a major problem. However, the situation is different for decisions involving the health of patients. Explainability is therefore crucial for medical ML systems and benefits all parties involved: patients, physicians, governments, ML engineers, and other decision makers in the healthcare system. All these parties have a legitimate interest in fair, unbiased, reliable, and reasonable AI based on medical properties rather than spurious correlations [14]. Transparent AI that provides explanations for its predictions facilitates these goals by enabling users to better understand the factors that contributed to a prediction. Explanation methods also enable governments to more effectively regulate AI systems through audits, and machine learning engineers to more easily maintain and improve their models. Stakeholders expect decision support systems to be transparent and to fit seamlessly into existing workflows [1,15]. Above all, however, transparency includes being explainable.

Unfortunately, explainability methods are underutilized in medical ML research. It is already an immense amount of work to define a medical problem suitable for a ML solution, obtain the necessary data, clean the data so that it can be used for modeling, develop a model, and refine the model to achieve high performance. Therefore, once one is past this hurdle, in many cases the inclusion of explainability is no longer considered or was not planned for in the first place.

With this work, we hope to facilitate the incorporation of explainability methods into medical ML through two main contributions. First, we provide a representative overview of the major classes of ML interpretability methods and highlight the advantages and limitations of the various approaches. Second, we analyze the extent to which previously published papers in medical ML use explainability methods and quantify which methods are used and how they are presented. We hope that this will allow researchers who have not previously been familiar with explainable ML to select a method or class of methods that are appropriate for their area of research and to integrate explainability in the future.

To achieve our goal, we conducted a comprehensive and systematic literature search. Inspired by recent analyses (e.g. [16]) and several discussions with medical professionals, we systematically searched PubMed using PRISMA guidelines. We have paid particular attention to the following aspects, the combination of which has received little attention to date:

- Is there a concrete medical supervised ML use case that uses interpretability methods?
- Who is potentially able to understand the XAI methods explained in the paper?
- Which kind of data is used, and wow well is ML pipeline described?
- Do authors provide their source code and data?

With these aspects in mind, our PubMed search found 2568 papers, of which 450 remained after applying exclusion criteria.

In the following, we present our approach and show that the field is changing dynamically.

2. Related work

This section is divided into three subsections. The first section deals with other work on explainability methods and the description of the XAI taxonomy. The second section deals with reviews from similar or related XAI areas that follow slightly different naming conventions. The third section deals with cutting edge topics such as causal ML. Although the topic of XAI is still young, even in medicine, we consider this subdivision already important and discuss the related works along the categories. There are other reviews of explainability methods, i. e., Ward prepared a summary table of explainability methods, available here. Although this paper mentions essential XAI methods, the methods are not classified according to their application to tabular data or image data. In addition, some more recent methods are missing. This GitHub repository contains a large Markdown table with hyperlinks to source code for explainability methods organized by year. However, this repository is mainly focused on image classification rather than medical data. Tjoa and Guan provide an overview of some interpretability methods related to medicine. However, they follow a different taxonomy for ML interpretability methods that we have not found in other works, which makes comparability and classification difficult [17]. Another systematic review considers XAI systems in the medical field [18]. The authors found that post-hoc methods were more common than intrinsic methods in the papers reviewed, and they discuss human-in-the-loop and inclusion of domain experts[1].However, they did not examine why other papers did not use XAI. Linardatos et al. [19] published a comprehensive collection of existing methods of ML interpretation methods. They propose an alternative taxonomy to allow a multi-perspective comparison between techniques. Methods are categorized into four main groups by intended use: *methods for explaining complex black-box models*, *methods for building white-box models*, *methods for limiting discrimination and improving fairness in models*, and *methods for analyzing the sensitivity of model predictions*. In the group of methods that explain black-box models, the authors further distinguish between black box deep learning models and arbitrary black box models. The second category contains methods that create easy-to-understand models, while the third class includes techniques that focus exclusively on the discrimination, inequality, and impartiality of an ML algorithm and evaluate it with respect to these properties. The methods in the last group are applied to evaluate ML algorithms in terms of reliability and sensitivity to ensure that their predictions are credible and consistent [19]. Linardatos et al. do not address the medical application of the XAI methods presented, nor do they discuss real-world use cases of the approaches. In the recent papers [20–24], the authors focus on XAI in a medical context. However, they each consider a specific medical subspecialty rather than a general view of medicine. In these works, the authors also did not analyze whether source code is provided, what stakeholders benefit from the XAI, and for which data format the presented methods are suitable.

There are also reviews with different wordings, i. e., Antoniadi et al. performed a systematic literature review for clinical decision support systems (CDSS). The main finding was the absence of XAI in CDSS for tabular and image data [25]. Quinn et al. provide an overview of the current state of machine learning in healthcare and provide an optimistic and pessimistic scenario for future diagnosis of AI systems in healthcare. However, they do not go into detail about current explanatory methods, but rather trace the historical development of ML in healthcare [26]. Other related work in the area of XAI research include i. e. Holzinger et al., who argue beyond explainability by saying that the domain expert understands an ML system better because s/he knows the causality of the relationships, but the system only knows the data [27]. In this context, Holzinger et al. emphasize the importance of causality

[1] By domain experts, we mean experts in the field to which the ML algorithm is applied. For example, in healthcare use cases, this could be physicians

relations in XAI, but also mention that so far these cannot be given by the algorithm but require domain knowledge. Adida and Berrada provide an overview of XAI in general and classify common ML interpretation methods using the taxonomy explained in the Methods section of this paper. However, they do not do so in terms of medicine, nor do they rank the methods in terms of their applicability to tabular data or neural networks [28]. Longo et al. address the challenges and emphasize the relevance of XAI in sensitive sectors such as medicine or law, but the focus is on XAI in general rather than medicine in particular [29].

While existing work has mostly examined the existing literature for XAI applications in specific medical subspecialties, to our knowledge, there has been no general literature review of a similar scope to our work for XAI applications in the medical field overall.

3. Explainability methods

By the word *model*, we generally mean the model learned by the system after performing some learning algorithm. Some machine learning models are inherently explainable, including linear regression, logistic regression, generalized linear models, or decision trees. Other models, such as neural networks, are black box by default, but can be augmented with explainability methods. These "add-on" explainability methods for otherwise non-interpretable models are the focus of this section. For a more detailed overview of the explainability methods we consider, see our supplementary material at GitHub. Explainability is also referred to in the literature as interpretability, intelligibility [30,31], causability [27], or understandability [32]. There is a tendency for "explainability" to refer to model-specific methods and "interpretability" to refer to inherently interpretable ML models or model-agnostic methods, but there is no consensus in the research community. We thus use the terms explainability and interpretability interchangeably in this paper, under the following definition:

Explainability method (synonym *interpretability method*): A method that enables humans to understand why a model makes certain predictions.

3.1. Trustworthy vs. untrustworthy explainability methods

Some explainability methods are trustworthy, meaning that their explanations are provably guaranteed to reflect the model's computations. A trustworthy explainability method can be used to assess how a model performs and help distinguish between performing and non-performing models. Models that do not perform should be tested further, while models that do perform could be moved into a deployment process. Whether a model performs or not depends on the use case and the chosen metrics to optimize, such as Mean Squared Error for regression tasks or weighted F1-scores for multi-class classification tasks. When a trustworthy explainability method is applied to a **non-performing model**, the explanations may seem strange or unexpected - for example, a neural network that uses metal tokens and postprocessing artefacts to predict pneumonia from chest x-rays [14] or a neural network that highlights snow to explain its classification of a *wolf*. The key is that the explanations can be used to conclude that the model is non-performing, because the explanation method is trustworthy. When a trustworthy explainability method is applied to a **performing model**, then the explanations will make sense even under maximum scrutiny by a human domain expert. In the ideal case, the explanation of a performing model will match the explanation of a group of domain experts, and the performing model could then be considered for deployment in a real-world setting. Some explainability methods are not trustworthy because they do not come with mathematical guarantees that they reflect the model's computations [33]. For example, the explainability method Grad-CAM is popular and highly cited, but has recently been shown to sometimes produce misleading explanations that do not represent how the model makes predictions [34]. Grad-CAM is thus not a trustworthy explainability method and cannot be used to draw conclusions about whether a model is non-performing or good. Because of the flaw in Grad-CAM, a new method called HiResCAM [34] was developed which does come with mathematical guarantees that it accurately reflects the underlying model, and thus HiResCAM is a trustworthy explainability method.

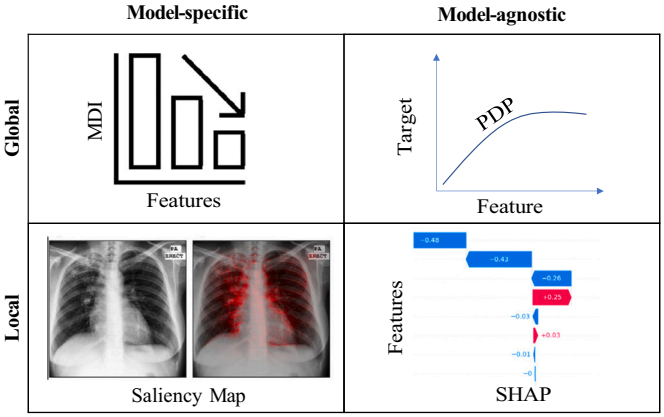


Fig. 1. Taxonomy concept used to classify the XAI approaches, following [19,28,32,39]. The methods outlined in the quadrants are exemplary representatives of this category. MDI = Mean Decrease Impurity, PDP = Partial Dependence Plots, SHAP = Shapley Additive Explanations.

Is there a difference between the terms explainability and interpretability? We review some definitions of interpretability and explainability in the literature. Sometimes the definitions agree, and sometimes they contradict each other. We use these terms interchangeably. Interpretability answers the question of how the models work, while explainability answers the question of what else the model says according to [32] [35]. defines interpretability as the ability to explain to a human in terms that can be understood [33]. in turn states that post hoc methods can be considered examples of explainability, while intrinsic methods can be considered examples of interpretability [36]. says that interpretability is the degree to which a human can understand the cause of a decision. Explainability in the technical sense highlights decision-relevant parts of the used representations of the algorithms and active parts in the algorithmic model that […] contribute to model accuracy […] [27]. Interpretability, in turn, is the extent to which a human can consistently predict the outcome of the model [37]. Gilpin et al. argue that explainable models are interpretable by default, but not vice versa. They describe explainability as models that can summarize the reasons for neural network behavior [38]. Glassboxmedicine states that interpretability means that the algorithm is intrinsically designed to establish a relationship between input and output that is understandable to humans, such as in linear regression, and counters that explainability means that the algorithm's decision making can be understood, even if it is abstractly detached from a human logic. For example, a deep-learning algorithm can explain a wolf by highlighting snow.

Explainability is not causality, algorithmic transparency, or simple input variables. Explainability methods typically specify which parts of the input contribute to the output of a model, but do not specify causal relationships or indicate *how* particular parts of the input affect a prediction. Explainability is also distinct from algorithmic transparency, which is a clear description of the algorithm's implementation and training process. In our opinion, understanding what an input variable means (e.g., the dictionary definition of the variable "age") does not make the model explainable.

3.2. Types of explainability methods

Explainability methods can be classified as model-agnostic or model-specific and as global or local, following the taxonomy of [19,28,32,39],

and as shown in Fig. 1. Model-agnostic methods are independent of the structure of the ML algorithm, while model-specific methods can only be applied to specific classes of models. Global methods explain the model as a whole, while local methods explain a single prediction. In the following subsections, representative examples of different classes of interpretability methods are described. Following our filter criteria of our literature review, the following described machine learning methods have the purpose to primarily explain supervised machine learning methods. More technical explanations are aimed at ML engineers, context dependent methods like saliency maps are rather aimed at domain experts, simpler methods like partial dependence plots or decision trees are aimed primarily but not exclusively at non-ML experts and non-domain experts like, i. e., patients.

3.2.1. Global model-agnostic methods

Global, model-agnostic methods describe the overall average behavior of models and are applicable to a wide range of machine learning models. Typical examples of methods in this category are Partial Dependence Plots (PDPs), Permutation Feature Importance, Leave One Covariate Out (LOCO) and Maximum Mean Discrepancy Critic (MMD-Critic), and Permutation Importance.

A partial dependence plot (PDP) illustrates the marginal effect of a feature on the target [40]. A PDP can be constructed for categorical and continuous features as well as for classification and regression problems. However, the PDP assumes that the features are uncorrelated, which can be problematic for multidimensional prediction problems.

Permutation feature importance [41] estimates the importance of a particular feature to a trained model. It is the absolute difference in performance score when a real feature is replaced by a dummy feature; the more performance degrades, the more important that feature is to the model. The importance of the permutation depends on the model and the performance score chosen; any change in the performance score can change the ranking of the features. This method also cannot account for covariances between features.

Leave One Covariate Out (LOCO) [42] is a model-agnostic global and local feature importance method, similar to feature importance in Random Forests. However, unlike feature importance in Random Forests, the feature under consideration is not replaced by a dummy variable, but simply omitted. Both methods have in common that they ask *How good is the model without this feature?* The assumption behind this is that a feature is important for the model if the performance is significantly worse without this feature.

Maximum Mean Discrepancy (MMD-Critic) [37] distinguishes between representative samples of a class and outliers. Typical representative samples are called prototypes, and the outliers are called criticisms. The distinction between prototypes and outliers is intended to provide additional insight into the model.

Other methods for global model-agnostic diagnostics include accumulated local effects plots, H-statistics, and functional decomposition.

3.2.2. Local model-agnostic methods

Local model-agnostic methods explain individual predictions and are applicable to a wide range of machine learning models. Popular examples of methods in this category include Individual Conditional Expectation (ICE), Locally Interpretable Model-agnostic Explanation (LIME), anchors (scaled rules), SHapley Additive exPlanations (SHAP), and influence functions.

Individual conditional expectation (ICE) [43] is a refinement of PDP and accounts for the heterogeneity of individual data points. ICE disaggregates PDPs to illuminate individual conditional expectations from supervised models.

Local Interpretable Model-agnostic Explanation (LIME) [44] trains an interpretable model to approximate the predictions of the real model. LIME can locally explain text models from tree-based algorithms as well as computer vision models, such as deep neural networks. Later work has shown that random noise leads to instability in LIME-generated explanations [45,46], leading to the development of LIME variants, including S-LIME [47] and DLIME [46].

Anchors (scoped rules) is another method developed by LIME authors [48]. In this method, IF-THEN rules are created to indicate which feature values anchor a prediction. Rules for rare classes or near the boundary of decision functions can become complex and sometimes ambiguous.

SHapley Additive ExPlanations (SHAP) is a model-agnostic method that allows for both global and local explanations and considers both structured and unstructured data [49]. SHAP indicates the contribution of a feature value to the difference between the actual prediction and the mean prediction. SHAP is based on Shapley values from game theory [50], dispersion activation features [51], and model intrinsic approaches from tree-based methods.

Influence functions [52] trace a prediction through the model and back to the training data to identify training points that are most responsible for a particular prediction. The influence function method can be applied to any model for which a second derivative exists. A derivative (synonym differentiation) exists, i. e., for neural networks. However, it is computationally intensive because the model must be retrained when the training data changes.

Other local model-agnostic methods include individual conditional expectation curves (which can be used to generate partial dependence diagrams) and counterfactual explanations.

3.2.3. Global model-specific methods

Global model-specific methods describe the overall average behavior of a model for a given class of models. Methods in this category include Mean Decrease Impurity (MDI), Testing Concept Activation Vectors (TCAV), Soft Decision Trees, and TabNet. Mean Decrease Impurity (MDI) [53] explains the importance of features for tree ensembles. Testing Concept Activation Vectors (TCAV) is a global and local explanation method for computer vision models and tabular discrete data [54]. Soft decision trees use a decision tree to mimic the input-output function of a neural network [55]. In a soft decision tree, all leaf nodes contribute to the final decision with different probabilities [56]. For some leaf nodes, the soft decision tree allows a visual interpretation of the neural network. However, not all learned filters are interpretable to the human eye. TabNet [57] uses sequential neural networks to mimic the logic of a decision tree on tabular data. Feature meanings provide global explanations, while heat maps provide local explanations. Instance-based feature selection can lead to confusion when local and global feature meanings contradict each other. Other global model-specific methods include Automatic Concept-Based Explanations (ACE) [58] and Deep Lattice Networks (DLN) [59]. Related to Decision Trees, but bringing in the aspect of symbolic AI, is the *Trepan Reloaded* method [60]. It uses ontologies to represent a network of information with logical relations and thus brings Domain Expert knowledge directly into the XAI system.

3.2.4. Local model-specific methods: gradient-based explanations for neural networks

Local model-specific methods explain a particular prediction of a particular class of models. The most popular local model- specific methods are gradient-based neural network explanations, which we consider in this section as an example of this class.

Gradient-based neural network explanation methods use the gradient of a model to produce an explanation for a given input example and output class [61]. They are most applied to neural networks for image classification, for which they provide a visualization to highlight which regions of an input image were used to make a prediction.

Input-Level Gradient-Based Methods. Gradient-based methods at the input level involve gradients or gradient-like calculations that lead from the output layer back to the input layer. So, the input layer refers to the level at input. Non-technical readers might want to know that the gradient is a derivative vector for a multivariate function, and the

Table 1
Overview of interpretability methods relevant to tabular and computer vision tasks, ordered by year of publication. Method relevance to neural networks, computer vision, and tabular data is indicated in the respective columns. The number of citations was derived from Google Scholar as of December 23, 2021. Links to the source code are provided via hyperlinks. We included methods that had more than 100 citations on Google Scholar, whose source code was publicly available, and that were optionally used in the review articles. An explanation of each method with advantages and limitations can be found in the supplementary material on GitHub. Regr = Regression, Classif = Classification.

Method / taxonomy	Specific (S) or agnostic (A)	Local (L) or global (G)	Neural networks	Computer vision	Tabular data	Year	No. citations	Regr. (R) or classif. (C)	Source code available
Partial Dependence Plots (PDP) [40]	A	G	No	No	Yes	2001	15,545	R and C	Yes
Permutation Importance [41]	A	G	No	No	Yes	2010	15,545	R and C	Yes
Mean Decrease Impurity [53]	S	G	No	No	Yes	2013	823	R and C	Yes
Individual Conditional Expectation [43]	A	L	Yes	No	Yes	2013	571	R and C	Yes
DeepLIFT (Deep Learning Important FeaTures) [51]	S	L	Yes	Yes	No	2016	1629	C	Yes
Layer-Wise Relevance Propagation [69]	S	L	Yes	Yes	No	2016	2160	C	Yes
Maximum Mean Discrepancy - Critic [37]	A	G	Yes	Yes	No	2016	445	C	Yes
Gradient-weighted Class Activation Mapping [74]	S	L	Yes	Yes	No	2016	6758	C	Yes
Integrated Gradients [71]	S	L	Yes	Yes	No	2017	2017	C	Yes
Local Interpretable Model-agnostic Explanation (LIME) [44]	A	L	Yes	Yes	Yes	2017	5020	R and C	Yes
SHapely Additive exPlanations (SHAP) [49]	A	L and G	Yes	Yes	Yes	2017	5020	R and C	Yes
Leave One Covariate Out [42]	A	L	No	No	Yes	2017	274	R	Yes
Influence Functions [52]	A	L	Yes	Yes	No	2017	1377	C	Yes
Soft Decision Trees [55]	S	G	Yes	No	No	2017	357	C	Yes
SmoothGrad [68]	S	L	Yes	Yes	No	2017	867	C	Yes
Testing Concept Activation Vectors [54]	S	L and G	Yes	Yes	No	2018	583	C	Yes
Anchors [48]	A	L	Yes	Yes	Yes	2018	922	R and C	Yes
Representer Point Selection [76]	S	L	Yes	Yes	No	2018	105	C	Yes
Automatic Concept-based Explanations [58]	S	G	Yes	Yes	No	2019	157	C	Yes

derivative of a function is the change of the function for a given input. Gradients are used, i. e., to fit neural networks to a dataset. The resulting explanation has the same number of pixels as the input image. Input layer approaches include saliency mapping, Guided Backpropagation, Deconvolutional Networks, SmoothGrad, Gradient × Input, Layer-Wise Relevance Propagation, and DeepLIFT. All these approaches are computationally efficient but suffer from white noise caused by shattered gradients [62], which sometimes prevents the resulting explanations from appearing class-specific in practice. Saliency mapping is the original gradient-based explanation method for neural networks. Saliency mapping computes the gradient of the class score with respect to the input image [63]. DeconvNets [64] and Guided Backpropagation [65] are explanation methods developed independently that happen to be identical to saliency mapping except for handling of the ReLU nonlinearities [66]. Saliency mapping passes explanation method sanity checks, while Guided Backpropagation does not, and may in fact function more like an edge detector than a model explanation [67]. SmoothGrad is another variant of saliency mapping that aims to reduce noise in explanations [68], but is not demonstrably more faithful to the model. The Gradient × Input method is equivalent to Saliency Mapping, except that the saliency map is multiplied element by element with the input image to create the final visualization. It has been shown later that the Gradient × Input method fails sanity checks [67]. Layer-Wise Relevance Propagation (LRP) [69] generates relevance values for the input pixels by iteratively distributing the final value across the layers of the neural network, starting with the output layer and working backwards to the input layer. Values greater than zero indicate that a particular pixel is relevant to the selected class. There are several variants of LRP. While LRP was not originally described as a gradient-based explanation method, it was later demonstrated [61] that ε-LRP is a variant of the gradient-* input method, in which the gradient calculation is changed based on the ratio of output to input at each nonlinearity. Finally, Deep Learning Important FeaTures (DeepLIFT) [51] provides explanations by estimating how much each neuron in a neural network is activated for an individual input compared to a reference input. The reference input is neutral (*foil*), while the individual input can be described as *fact* [70]. After the development of DeepLIFT, it was proved [61] that DeepLIFT computes backpropagation for a modified gradient function. Other input-level gradient-based methods include integrated gradients [71] and EXplanation Ranked Area Integrals (XRAI) [72]. For all these gradient-based methods one should keep in mind that the shattered gradients problem negatively affects the quality of the pixel importance values.

Output-Level Gradient-Based Methods. In gradient-based explanations at the output layer, a gradient is computed that runs backwards from the output layer for only one or a few layers of the neural network without going all the way back to the input layer. Thus, the bare explanation has a smaller dimension than the input and must be upsampled before it is overlaid with the input to create the final explanation. Such an upsampling step is permissible because in a typical neural network the spatial relationship between output and input is preserved. Output-level approaches include Class Activation Mapping (CAM), Grad-CAM, and HiResCAM. Class Activation Mapping (CAM) is the fundamental method in this class [73]. CAM is based on a particular convolutional neural network architecture, where convolutional layers are followed by a global average pooling and a single fully connected layer, which provide the final predictions. A CAM explanation is obtained by multiplying the class-specific weights of the final fully connected layer by the corresponding feature maps before the global average pooling step. CAM is a gradient-based method because these final weights represent the gradient of the class score with respect to the feature maps. The CAM method is trustworthy and guaranteed to

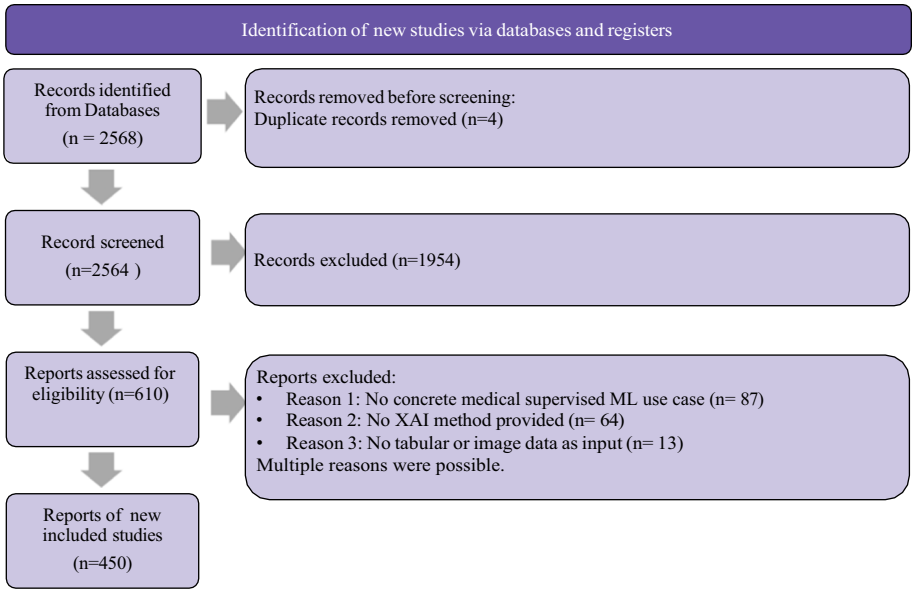


Fig. 2. Flowchart of our PRISMA literature research in the PubMed database on 2022-03-07 based on the template of [77]. Out of 2568 papers, 450 have been finally included.

highlight only regions the model used, but it has architecture restrictions. Gradient-weighted Class Activation Mapping (Grad-CAM) [74] aims to generalize CAM to other architectures. In Grad-CAM, the gradient of the class score is computed with respect to a given set of feature maps. Then, the gradient is averaged per feature and the averaged gradient is multiplied by the corresponding feature map. The aggregation of these weighted feature maps is the Grad-CAM explanation. The paper presenting Grad-CAM has been cited over 9000 times, but unfortunately it has recently been shown that Grad-CAM is not faithful to the underlying model due to the gradient averaging step [34]. Grad-CAM's explanations highlight irrelevant regions of the input image that were not used for prediction, which can lead to misleading explanations that deviate significantly from the true behavior of the model [75]. HiResCAM [34] is a newer method that eliminates the inaccuracy of Grad-CAM. HiResCAM eliminates the gradient averaging step in Grad-CAM. By retaining the detailed gradient information and multiplying the gradients element by element with the corresponding feature maps, the relationship between the model explanation and the class evaluation is provably maintained, resulting in trustworthy class-specific explanations. The source code for HiResCAM is publicly available here and as part of this package.

3.3. Summary

Describing all the machine learning interpretability methods ever developed would require an entire textbook. Therefore, this section is not comprehensive, but rather is intended to provide representative examples of the major classes of interpretability methods. Table 1 gives a brief overview of common interpretability methods and summarizes the characteristics of each method.

4. Materials and methods

This section describes the search term, search results, inclusion and exclusion criteria, and research questions answered for each of the papers in this literature review.

4.1. Literature selection

Since we focus on medical data, we deliberately chose PubMed as our search database. Technical papers in this field, such as from the journal Artificial Intelligence in Medicine, are also archived in PubMed, so we consider the database to be representative for our study purpose. We chose the search query (i. e., search term) that we applied to PubMed as follows: Within the title or abstract, we searched for terms associated with explainability, intelligibility or interpretability, combined with general terms that are related with machine learning and the medical domain. The search period covers the years from 2020 to 2022 and was conducted on March 7, 2022. We further applied the Preferred Reporting Items for Systematic Reviews [77] (PRISMA) guidelines. The aim of the literature search was to identify all papers from the last 20 years that met the following inclusion criteria:

- Used a supervised machine learning method,
- for a medical use case,
- with tabular or image data as input,
- and incorporating at least one explainability method.

Papers that focused on unsupervised learning (e.g., clustering), did not include a medical use case, or did not explicitly consider explainability were excluded. Borderline cases where it was not clear whether a paper should be included or not were discussed by the reviewers in separate meetings and concordantly accepted or rejected.

Using a search term that conjuncts the fields of machine learning, explainability and medicine (for details, please refer to the supplementary material), 2568 references were initially identified in the PubMed database on 2022-03-07. The search was limited to the title and abstract of the paper, meaning that machine learning and explainability had to be explicitly mentioned in these sections. After removing 4 duplicates, 2564 references remained. Then, an author (J.A. or L.M.) applied the inclusion and exclusion criteria described above to each title and abstract, resulting in 610 references that were eligible for full-text screening. A common reason for exclusion at this stage was the lack of a true explainability method. For example, papers were excluded that used a black-box model but claimed that their model was explainable because a human could understand the dictionary definition of the input variable (e.g., “age”). We reviewed papers on time series data from electrocardiograms (ECGs) or electroencephalograms (EEGs), but unfortunately had to exclude all of these time series papers because none of them included an explainability method. Although our initial search

Table 2
The research questions used were related to the 450 papers with a specific use case of supervised machine learning in medicine that used table or image data as input and applied at least one XAI method to explain ML predictions.

No.	Question	Answer options	Question type
1	Is there a concrete medical supervised ML use case?	Yes | No	Single Choice
2	Which XAI method is used?	*Several options inlcuding an open text field*	Multiple Choice
3	From which data format is the input?	Tabular | Image | Audio | Text	Single Choice
4	Who is potentially able to understand the XAI method?	Developers | Medical Professionals | Patients | Other	Multiple Choice
5	How well is the ML pipeline described?	Not described | Described | Elaborately described	Single Choice
6	Is the source code provided?	Yes | No | Upon request	Single Choice
7	Is the data publicly available?	Yes | No | Upon request	Single Choice

We analyzed the Microsoft Forms data for all papers using Python 3.9.[2] All source code and raw data is available on the supplementary material on GitHub.

The ML pipeline was rated 1, 2, or 3 according to the following criteria, which were agreed upon in advance:

- 1 = not described;
- 2 = described = the ML pipeline was mentioned and described briefly;
- 3 = elaborately described = the ML pipeline was described in detail, illustrated with a figure, or provided as publicly available code.

5. Results

The results presented here are not always direct answers to the research questions but are primarily a combination of the information we obtained from the research questions. Therefore, a quick overview of the non-combined results is given in Fig. 3.

What is the publication rate of medical ML papers over time?

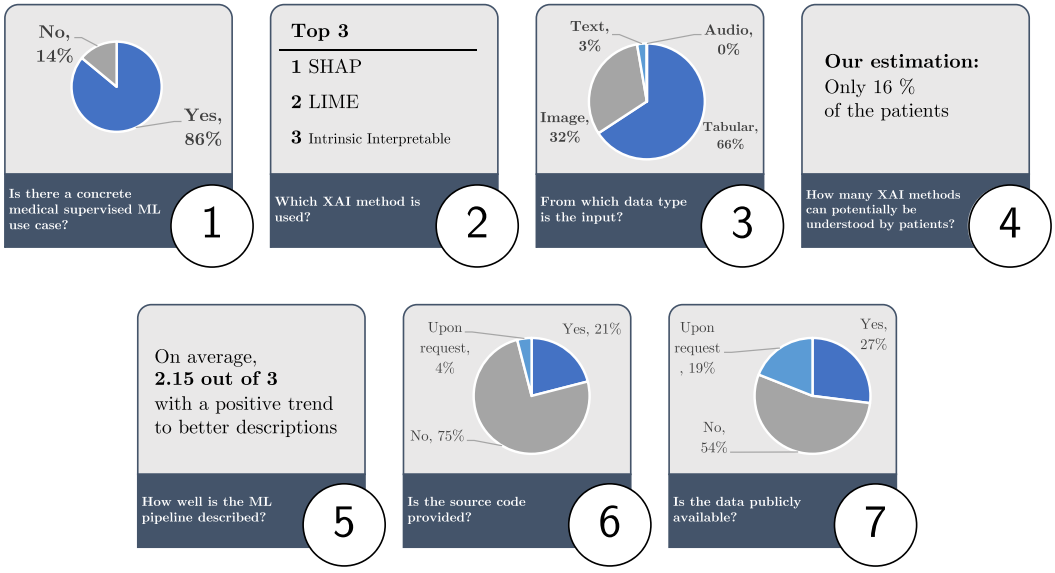


Fig. 3. Brief overview of the results of the 7 research questions.

term included papers from 2002 onward, the oldest paper that met the inclusion criteria was from 2008. The 610 references approved based on title and abstract were then subjected to full-text screening. At this stage, 160 references were excluded for at least one of the following reasons: no specific medical supervised ML use case (87), no XAI method (64), or no image or tabular data as input (13). There were 450 references left for data analysis with our 7 research questions. The Results section contains the analyses performed on these 450 references. The full PRISMA flowchart is shown in Fig. 2.

4.2. Literature review

We evaluated each of the 450 final papers individually. For each paper, we determined which XAI methods were used and how they were described using 7 screening questions listed in Table 2. We did not consider model confidence estimation or model uncertainty as explanatory methods because they do not provide insight into how a model arrives at a prediction.

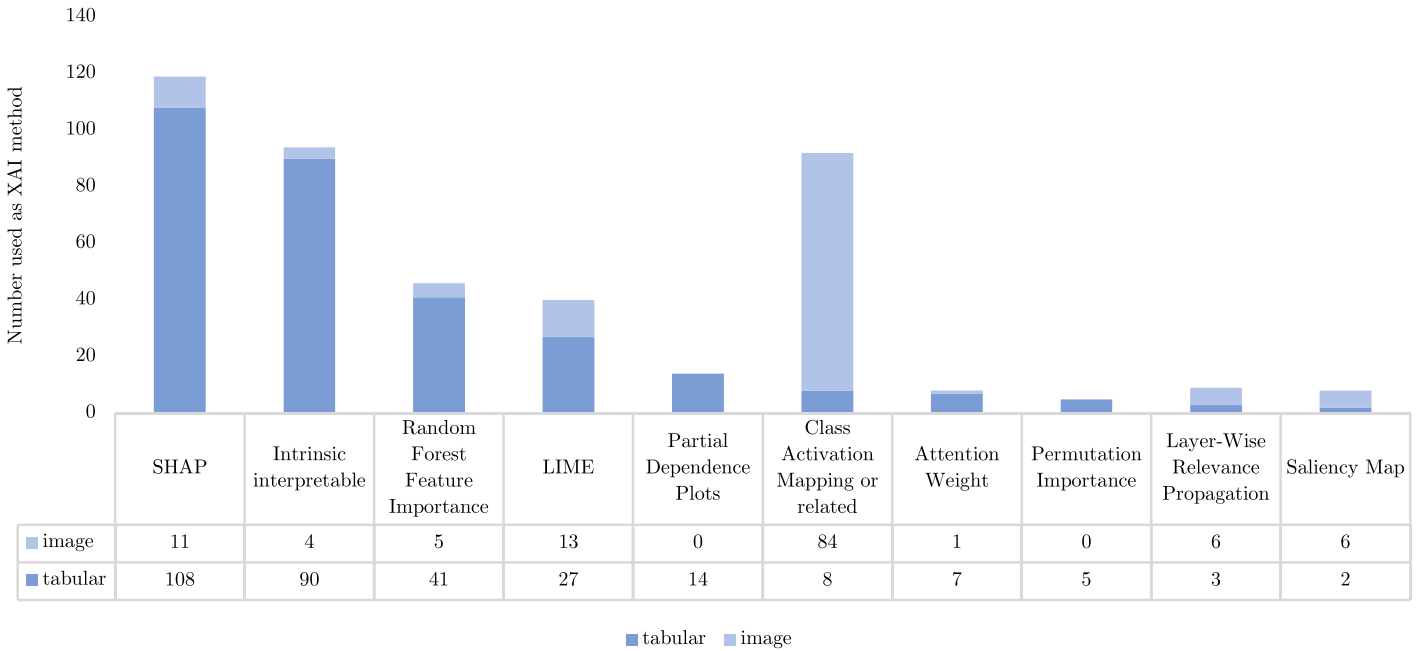
Data Synthesis. We used Microsoft Forms to collect responses to our research questions and Python to aggregate the data according to our research questions. Each paper was reviewed by an author (J.A. or L.M.).

The publication rate of medical ML papers increased over time, with 79 papers published between 2008 and 2019, 108 more in 2020 only, and 200 in 2021. The 63 papers for 2022 were published within the first 67 days of the year. If we extrapolate this to 365 days, we expect about 343 publications in 2022.

What XAI methods are most used? Of the total 535 XAI methods used, 45 were developed by the authors of the use cases themselves and 490 were based on previously published methods such as SHAP, LIME or Grad-CAM. About 1.2 XAI methods were applied per paper. Fig. 4 shows all XAI methods used in at least 3 papers, grouped by tabular data and image data. For tabular data, the most common XAI methods are SHAP, Random Forest Feature Importance, and intrinsic methods. For image data, the most used methods include class activation methods, SHAP, and LIME. Although we grouped gradient-based explanatory methods in the same category, we found that Grad-CAM was the most used method in this category.

Is tabular or image data more frequently used in medical ML papers? A total of 307 papers (68 %, (95 % CI [63.7 %, 72.5 %])) dealt

[2] https://www.python.org/downloads/release/python-390/.



	SHAP	Intrinsic interpretable	Random Forest Feature Importance	LIME	Partial Dependence Plots	Class Activation Mapping or related	Attention Weight	Permutation Importance	Layer-Wise Relevance Propagation	Saliency Map
image	11	4	5	13	0	84	1	0	6	6
tabular	108	90	41	27	14	8	7	5	3	2

Fig. 4. Number of XAI methods used for image and tabular data. At least one XAI method was considered for each paper. For tabular data, SHAP was by far the most popular XAI method; for image data, it was Grad-CAM. Methods that were used fewer than three times are not listed here.

XAI Method	2008-2019	2020	2021	2022	Sum
SHAP	1	20	73	25	**119**
Intrinsic interpretable	33	25	34	2	**94**
Class Activation Mapping or related	7	23	40	22	**92**
Random Forest Feature Importance	9	13	19	5	**46**
LIME	5	6	24	5	**40**
Partial Dependence Plots	2		10	2	**14**
Layer-Wise Relevance Propagation	1	3	4	1	**9**
Attention Weight	2	3	2	1	**8**
Saliency Map	0	2	5	1	**8**
Permutation Importance	1	1	3	0	**5**
DeepLift	2	0	1	0	**3**
Sum (inlc. all methods)	**75**	**106**	**244**	**65**	**490**

Fig. 5. XAI methods used by year. Due to the small number of papers per year between 2008 and 2019, we have combined these years as one group. Note that for 2022, only papers published through 2022-03-07 were included.

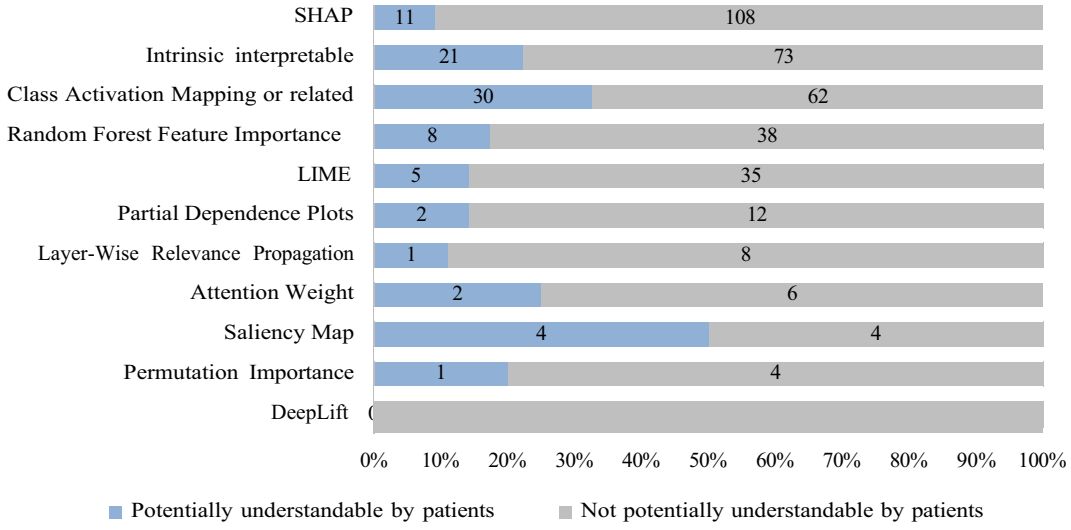


Fig. 6. Which method of explanation is potentially understandable to patients? Each row sums to 100 %. The numbers within the bars indicate the number of papers that used this method in our review. For example, for the bar belonging to permutation meaning, we think that 1 in 5 papers (20 %) could be understood by patients.

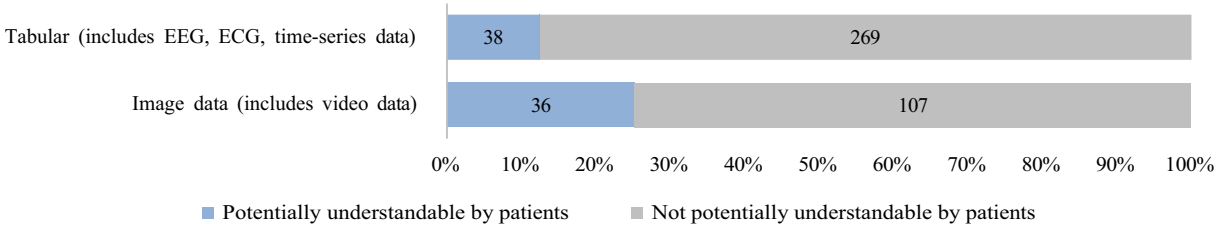


Fig. 7. Which explanation method is potentially be understood by patients? Each row sums to 100 %. The numbers within the bars indicate the number of papers that fall into each category.

with tabular data and the remaining 143 with image data. When grouping the type of input data by year, there is a clear trend towards greater use of image data over time. The ratio of tabular to image data was 20/80 in 2008–2019, while it increased to 36/64 in subsequent years. We think that this is likely due to the increasing availability of labeled image data and the increased computing power and availability of GPU clusters. Fig. 5 shows the use of XAI methods by year. While the use of intrinsic interpretable methods like decision trees and linear regression stays relatively constant, there is a large increase in the use of methods available in Python like SHAP, LIME, and Class Activation Mapping.

Who is potentially be able to understand the XAI method? On our opinion, patients are an important stakeholder group for machine learning applications in medicine and, on average, also have the greatest barriers to understanding due to their unfamiliarity with machine learning or medicine. To assess patient understanding of explainable ML models, we would ideally conduct a direct survey of patients. However, such a survey is beyond the scope of this work, so we instead made a subjective assessment of whether patients might be able to understand an XAI method. We considered a paper to be potentially understandable

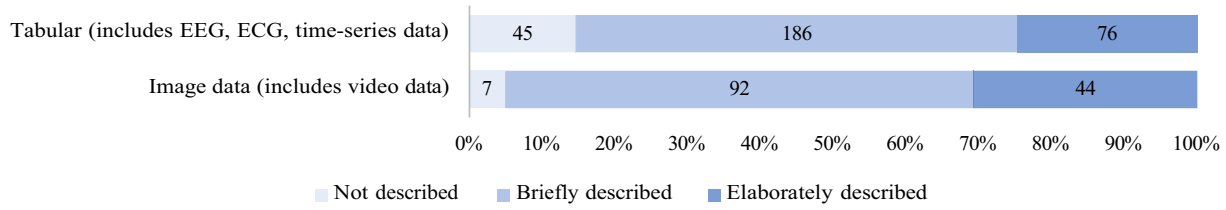


Fig. 8. Granularity of machine learning description grouped by input type of data. We defined three categories: not described, briefly described, and elaborately described.

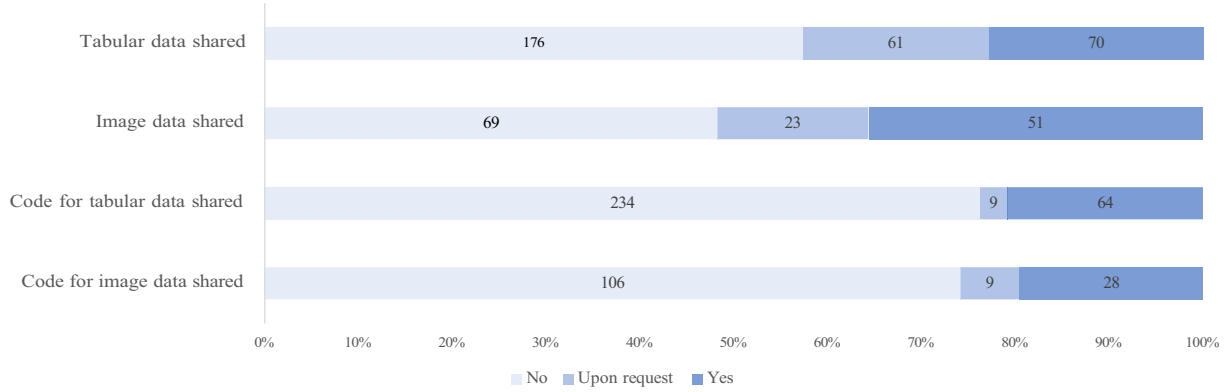


Fig. 9. Availability of data and code to reproduce ML results. Each row sums to 100 %. Note that the sum of the line *tabular data shared* is equal to *code for tabular data shared*. The same is true for the lines related to image data.

Table 3
Assessment of ML pipeline and ratio of code and data sharing over time. For 2022, only publications through 2022-03-07 are considered. The column *ML pipeline description* refers to the research questions *How well is the ML pipeline described?* and is a mean score. The higher the score between 1 and 3, the more detailed the ML pipeline description.

Year	# Papers	ML pipeline description	Code shared?			Data shared?		
			No	Upon request	Yes	No	Upon request	Yes
2008-2019	79	1,87	72,2%	5,1%	22,8%	58,2%	15,2%	26,6%
2020	108	2,15	68,5%	3,7%	27,8%	54,6%	18,5%	26,9%
2021	200	2,16	80,5%	4,0%	15,5%	52,0%	22,0%	26,0%
2022	63	2,30	76,2%	3,2%	20,6%	57,1%	12,7%	30,2%
Overall	450	2,15	75,6%	4,0%	20,4%	54,4%	18,7%	26,9%

to patients if it met all the following criteria:

1. The output of the explainability method does not require a deeper medical understanding.
2. If variables are the output, they must be explained or self-explanatory.
3. If codes are the output (1 = Female, 2 = Male), they must be explained.
4. If color scales are the output, they must be explained with a legend and the meaning of the marginal values.

Only 16.4 % of the papers met all the above criteria to be considered potentially understandable by patients. A detailed overview of which of the explanatory methods presented in the papers could be understood by patients is given in Fig. 6. Fig. 7 again shows that explanation methods with image data as input were generally rated as better understood by patients, mainly due to the intuitive nature of heatmap displays of pixel relevance.

How well is the ML pipeline described? The mean ML pipeline score was 2.15 (0.60 std). From the aggregation of years in Fig. 3, it appears that the description of the ML pipeline improves over time. This suggests that more weight was given to the ML pipeline in later work. Fig. 8 shows the granularity of the ML pipeline description by input type.

Is the source code provided and is the data used publicly available? Overall, 75.6 % of all included papers do not make their source code available to reproduce the results. Surprisingly, the willingness to share source code was higher in 2020 (27.8 %) than in 2021 (15.5 %) or 2022 (20.6 %). We did not find any association between increasing years and the code sharing ratio. A chi-square test of independence showed that there was no significant association between publication year and code sharing ratio, $\chi^2(6, N = 405) = 11.0$, $p = .09$. The willingness to share data is generally higher than the willingness to share code. In 2022, data availability was highest at 30.2 %; in previous years, it was 27 %. The increase in 2022 may also be since many papers have been published on the coronavirus radiograph use case and this data is publicly available [78]. For an overview, see Fig. 9 with further details in Table 3. Moreover, when a proprietary method was developed, willingness to share data was 4.4 % points higher.

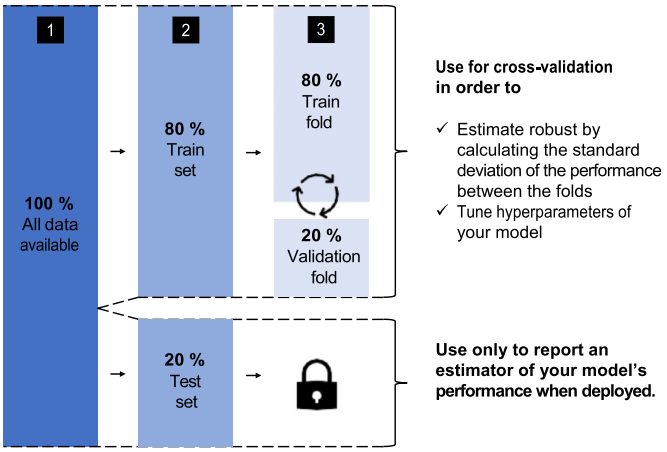


Fig. 10. Our best practice recommendation for splitting all available data into training and validation folds and a test set. The 80–20 split is only a rule of thumb and can be adjusted depending on the amount of data available. The arrow circle refers to the shifting validation folds of the cross validation. Note that we use the terms *validation* and *development* interchangeably. We denote *set* to data that is split into train and test sets. Within the train set, and when applying cross-validation, use the term *folds* to emphasize that this is not test, but validation data within the train set to estimate the model's generalization error. Av. = Available.

6. Discussion

This paper has explored two main tasks. First, we have provided an overview of the taxonomy of machine learning explainability and described representative methods. The machine learning taxonomy has converged in recent years, although homonyms and synonyms still exist, and some technical terms are not used consistently. Second, we examined the last 20 years of medical machine learning publications in the PubMed database for their use cases, input types, AI comprehensibility, code and data sharing, and XAI methods used. The most popular methods are SHAP, LIME, and intrinsically interpretable methods. Most input data is structured, tabular data (65 %) or images (32 %). Text data is rare at 3 %, and we found no use case for our criteria for audio data. We estimate that 16 % of the explanatory methods reported in publications can be understood by patients. The description of machine learning pipelines has become more detailed over time, while data and code sharing has stagnated.

In this discussion section, we address the limitations of our study, provide recommendations on how to further improve the reproducibility and explainability of AI systems, address the interdisciplinary nature of medicine and machine learning, and mention challenges that may arise in the future and with the use of systems.

6.1. Limitations of this review

We carefully discussed the search terms and the cases to be excluded before searching the databases to obtain as many precise hits as possible. However, because of the variable taxonomy in the ML community and the wide range of terms used in this inherently interdisciplinary field, we may not have identified all the papers that would have been relevant. We used the PubMed database because of our focus on medical use cases. PubMed is the largest medical database available. Google Scholar, i. e., has a more technical focus. In future work, we would like to extend our search on other databases. However, we do not expect our main findings to change. The final selection of 450 studies does not include papers on medical time-series data (e.g., ECG, EEG) or papers using support vector machines as a modeling framework because none of these papers considered explainability. Despite these limitations, we believe that our selection of relevant papers is large enough to derive representative conclusions. The results of our review are not free of subjectivity, especially when it comes to the evaluation of ML pipeline quality. Our assumptions regarding the comprehensibility of XAI methods to patients are also debatable. By setting criteria for when it is presumably understandable, a uniform assessment of the reviewed papers is ensured. However, the best estimator of understandability is admittedly obtained from a representative survey with examples from the papers, which is out of scope as mentioned earlier.

6.2. Recommendations to improve medical ML explainability and reproducibility

In reviewing the literature, we found that not all papers that include an explainability method explain it well. We have therefore developed several specific recommendations to improve the comprehensibility of medical XAI research.

First, the ML pipeline often requires more information about the split between training, validation, and testing. Some papers do not mention the data split at all, while others do not distinguish between validation and testing. We recommend at least mentioning the percentage split between training, validation, and test, and confirming that the final model's performance was calculated on the test set only. If cross-validation was used, we recommend indicating the robustness of the model by reporting the standard deviation of the performance metric across validation folds. In some deep learning literature [79], a validation set is mostly used to avoid overfitting during training which is like maximizing robustness: Overfitting is indicated by a performance drop between training and validation fold, low robustness is indicated by a high variance between training and validation fold. We also recommend using the term "fold" when cross-validation was used, and "split" or "set" otherwise. A graphical concept for this can be seen in Fig. 10. To give an example: there are 1000 samples in a survey with 1000 patients. We divide the samples into 800 for a training set and 200 for a test set. The test set is not used until the final evaluation. The training set is divided into 5 parts for 5-fold-cross validation. 4 folds (=640 samples) are used for the first training while the 5th fold (160 samples) is tested. The average performance of the 5th folds is an estimator for the performance in the test set. The standard deviation in the test set, which results from the deviations between the test folds, is an estimator for the robustness of the model at deployment.

Second, "explaining the explanations" is often critical for XAI to fully deliver its intended benefits. For tabular data, it is important to explain what each variable means. Rather than using an abbreviation such as "BP," it is helpful to write out the full variable description: "blood pressure." Sometimes variables were not even abbreviated, but presented as numeric coding without a legend, e.g., "1", "2". When numeric coding is used, the meaning of each variable should be indicated, e.g., "1 = blood pressure," "2 = respiratory rate." For images, it is useful to indicate important anatomical structures or abnormalities with arrows so that non-radiologists can identify whether the model explanation overlaps with relevant parts of the medical image. **It** is also often helpful to recognize when an explanatory method is beyond a size that a human can easily understand. Decision Trees (DT) continue to be a popular tool in medical XAI. We have observed that Decision Trees are used both as an intrinsic method, i. e., the tree is the model and its own explanation, and in other work as a post-hoc method to approximate a more complicated model, such as a gradient boosting machine or a neural network. A problem arises when the decision tree becomes too large. We recommend avoiding Decision Trees with several dozen levels, as these are understandable to humans in theory but not in practice: The sum of all decision rules along a path might be too long to be comprehended in a clinical daily routine. However, in future, better explanation methods must be also investigated. It is also important to clarify the splitting direction on the leaves. It is also relevant to consider the overall system of a use case with the three elements of the use case, the AI system, and the explanation of the AI system, rather than the XAI

method separately. A suggestion from the community is to extend the descriptions of the XAI methods with standardized metadata that are uniform for all XAI methods and thus simplify the implementation and the technical access, analogous to FAIR [80] (Findable, Accessible, Interoperable, Reusable) principle [81].

Finally, to facilitate replication and faster progress of medical ML research, we encourage increased de-identification (anonymization of personal data) and sharing of datasets, as it is often difficult to build directly on medical ML research when a new dataset needs to be created from scratch.

6.3. Understandability of an XAI method in medical ML is related to medical knowledge

In medical ML applications, medical knowledge is often useful to understand the results of an XAI method. Using image-based heatmap XAI methods as an example, we can consider different degrees of understanding. A general reader can reach a basic level of understanding, which we define as awareness that the highlighted pixels are relevant to the prediction. A physician who is not a radiologist would be able to reach an intermediate level of understanding, meaning that he or she is able to recognize organs and major abnormalities in the underlying medical image and consider how these relate to the relevance of the pixels. A radiologist would eventually be able to recognize even subtle anomalies in the medical image and assess their relationship to XAI pixel relevance. This means on the one hand that the degree of comprehensibility is essentially a subjective assessment, and on the other hand that the potential of comprehensibility depends on the background knowledge of the viewer. The more specific the AI application, the higher the dependency on domain knowledge for the comprehensibility of the XAI system.

6.4. Challenges

Some enthusiasts believe that the use of black-box ML systems is unproblematic [82], while the most conservative work argues that not even existing explainable ML methods are sufficiently understandable to justify the use of ML in a clinical setting, since explainable ML cannot confirm the correctness of a decision [83].

We take an intermediate perspective in which we believe that explainable ML has the potential to improve clinical care in certain circumstances. In our opinion, any deployment of a medical ML model should involve close collaboration between medical professionals, ML engineers, software developers, and computer security experts. Medical professionals have the deepest understanding of the model's explanations and can confirm whether a model's behavior appears medically appropriate. Only explanatory methods that are demonstrably faithful to the model should be used. Bias and fairness metrics should be calculated to ensure that the models used do not exhibit discriminatory behavior. Further, we think that the model must be protected from unauthorized access, and ML experts must be available to update the model in the event of a concept or data mismatch. In many countries, regulatory approvals are required for newly trained models.

Deploying a model is no guarantee that it will be used clinically. We think that the likelihood that a model will impact clinical care is greatest when the program's user interface for using the model has been carefully developed with significant input from medical professionals and when the model's outputs can be seamlessly integrated into existing software tools and workflows. Explainable ML methods with demonstrable guaranteed fidelity to the underlying model have the potential to improve the quality of medical ML models and prevent the use of possibly critical, biased, or ineffective models. The more medical ML research incorporates explainable techniques, the more clinical relevance it could achieve.

Supplementary information

Supplementary materials, such as detailed descriptions of the XAI methods, as well as the Python code to replicate numbers, figures and tables are available on github.com/joa24jm/literature_review.

CRediT authorship contribution statement

Johannes Allgaier (J. A.), Rachel Draelos (R. D.), and Lena Mulansky (L. M.) together wrote and revised the paper. J. A. and L. M. together did the literature search, screened abstracts, and did the full text reviews. J. A. created the supplementary materials, the tables, and most of the figs. L. M. created some of the figures. Rüdiger Pryss (R. P.) revised the paper and supervised the project.

Declaration of competing interest

The authors declare no competing interests.

References

[1] Benda NC, et al. “How did you get to this number?” stakeholder needs for implementing predictive analytics: a pre-implementation qualitative study. J Am Med Inform Assoc 2020;27:709–16.
[2] Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based fda-approved medical devices and algorithms: an online database. NPJ Digit Med 2020;3:1–8.
[3] Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJ. Artificial intelligence in radiology. Nat Rev Cancer 2018;18:500–10.
[4] Lee C, Nagy P, Weaver S. Cognitive and system factors contributing to diagnostic errors in radiology. Am J Roentgenol 2013;201:611–7.
[5] Iyawe EP, Idowu BM, Omoleye OJ. Radiology subspecialisation in africa: a review of the current status. S Afr J Radiol 2021;25:1–7.
[6] Dov D, et al. Thyroid cancer malignancy prediction from whole slide cytopathology images. In: Machine Learning for Healthcare Conference. PMLR; 2019. p. 553–70.
[7] Esteva A, et al. Dermatologist-level classification of skin cancer with deep neural networks. nature 2017;542:115–8.
[8] Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. Nat Rev Cardiol 2021;18:465–78.
[9] Draelos RL, et al. Genesis: gene-specific machine learning models for variants of uncertain significance found in catecholaminergic polymorphic ventricular tachycardia and long qt syndrome-associated genes. Circ Arrhythm Electrophysiol 2022;15:e010326.
[10] González-Nóvoa JA, et al. Using explainable machine learning to improve intensive care unit alarm systems. Sensors 2021;21:7125.
[11] Echle A, et al. Deep learning in cancer pathology: a new generation of clinical biomarkers. Br J Cancer 2021;124:686–96.
[12] Taghiakbari M, Mori Y, von Renteln D. Artificial intelligence-assisted colonoscopy: a review of current state of practice and research. World J Gastroenterol 2021;27:8103.
[13] Ć osić, K., et al. Ai-based prediction and prevention of psychological and behavioral changes in ex-covid-19 patients. Front Psychol 2021;12.
[14] Zech JR, et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med 2018;15:e1002683.
[15] McCoy LG, Brenna CT, Chen SS, Vold K, Das S. Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based. J Clin Epidemiol 2022;142:252–7.
[16] Allgaier J, Schlee W, Probst T, Pryss R. Prediction of tinnitus perception based on daily life mhealth data using country origin and season. J Clin Med 2022;11:4270.
[17] Tjoa E, Guan C. A survey on explainable artificial intelligence (xai): toward medical xai. IEEE Transactions on Neural Networks Learn. Syst; 2020.
[18] Chakrobartty S, El-Gayar O. Explainable artificial intelligence in the medical domain: a systematic review. 2021.
[19] Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable ai: a review of machine learning interpretability methods. Entropy 2021;23:18.
[20] Fuhrman JD, et al. A review of explainable and interpretable ai with applications in covid-19 imaging. Med Phys 2021;49:1–14.
[21] Singh A, Sengupta S, Lakshminarayanan V. Explainable deep learning models in medical image analysis. J Imaging 2020;6.
[22] Hauser K, et al. Explainable artificial intelligence in skin cancer recognition: a systematic review. Eur J Cancer 2022;167:54–69.
[23] Zhang Y, Weng Y, Lund J. Applications of explainable artificial intelligence in diagnosis and surgery. Diagnostics 2022;12.
[24] van der Velden BH, Kuijf HJ, Gilhuijs KG, Viergever MA. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. Med Image Anal 2022;79:102470.

[25] Antoniadi AM, et al. Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: a systematic review. Appl Sci 2021;11:5088.
[26] Quinn TP, Jacobs S, Senadeera M, Le V, Coghlan S. The three ghosts of medical ai: can the black-box present deliver? Artif Intell Med 2022;124:102158.
[27] Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip. Rev.* Data Min Knowl Disc 2019;9:e1312.
[28] Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE Access 2018;6:52138–60.
[29] Longo L, Goebel R, Lecue F, Kieseberg P, Holzinger A. Explainable artificial intelligence: concepts, applications, research challenges and visions. In: International cross-domain conference for machine learning and knowledge extraction. Springer; 2020. p. 1–16.
[30] Lou Y, Caruana R, Gehrke J. Intelligible models for classification and regression. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining; 2012. p. 150–8.
[31] Caruana R, et al. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining; 2015. p. 1721–30.
[32] Lipton ZC. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. Queue 2018;16:31–57.
[33] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1:206–15.
[34] Draelos RL, Carin L. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks. In: *arXiv e-prints* arXiv–2011; 2020.
[35] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. In: *arXiv e-prints* arXiv–1702; 2017.
[36] Miller T. Explanation in artificial intelligence: insights from the social sciences. Artif Intell 2019;267:1–38.
[37] Kim B, Khanna R, Koyejo OO. Examples are not enough, learn to criticize! Criticism for interpretability. Adv Neural Inf Proces Syst 2016;29.
[38] Gilpin LH, et al. Explaining explanations: An overview of interpretability of machine learning. In: *2018 IEEE 5th international conference on data science and advanced analytics (DSAA)*, 80–89. IEEE; 2018.
[39] Arrieta AB, et al. Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible ai. Inf Fusion 2020;58:82–115.
[40] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;1189–1232.
[41] Altmann A, Tolos IL, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics 2010;26:1340–7.
[42] Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L. Distribution-free predictive inference for regression. J Am Stat Assoc 2018;113:1094–111.
[43] Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Stat 2015;24:44–65.
[44] Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016. p. 1135–44.
[45] Zhang Y, Song K, Sun Y, Tan S, Udell M. "Why should you trust my explanation?" understanding uncertainty in lime explanations. In: *arXiv e-prints* arXiv–1904; 2019.
[46] Zafar MR, Khan N. Deterministic local interpretable model-agnostic explanations for stable explainability. Mach Learn Knowl Extr 2021;3:525–41.
[47] Zhou Z, Hooker G, Wang F. S-lime: stabilized-lime for model explanation. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining; 2021. p. 2429–38.
[48] Ribeiro MT, Singh S, Guestrin C. Anchors: high-precision model-agnostic explanations. In: Proceedings of the AAAI conference on artificial intelligence. 32; 2018.
[49] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Guyon I, et al., editors. Advances in Neural Information Processing Systems 30. Curran Associates, Inc.; 2017. p. 4765–74.
[50] Shapley L. Notes on the n-person game—ii: the value of an n-person game, the rand corporation, the rand corporation. Res Memo 1951;670.
[51] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: International conference on machine learning. PMLR; 2017. p. 3145–53.
[52] Koh PW, Liang P. Understanding black-box predictions via influence functions. In: International Conference on Machine Learning. PMLR; 2017. p. 1885–94.
[53] Louppe G, Wehenkel L, Sutera A, Geurts P. Understanding variable importances in forests of randomized trees. Adv Neural Inf Proces Syst 2013;26:431–9.
[54] Kim B, et al. Interpretability beyond feature attribution: quantitative testing with concept activation vectors (tcav). In: International conference on machine learning. PMLR; 2018. p. 2668–77.
[55] Frosst N, Hinton G. Distilling a neural network into a soft decision tree. In: *arXiv e-prints* arXiv–1711; 2017.
[56] Irsoy O, Yildiz OT, Alpaydin E. Soft decision trees. In: International Conference on Pattern Recognition; 2012.
[57] Arık SO, Pfister T. Tabnet: attentive interpretable tabular learning. In: arXiv; 2020.
[58] Ghorbani A, Wexler J, Zou JY, Kim B. Towards automatic concept-based explanations. Adv Neural Inf Proces Syst 2019;32.
[59] You S, Ding D, Canini K, Pfeifer J, Gupta MR. Deep lattice networks and partial monotonic functions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017. p. 2985–93.
[60] Confalonieri R, Weyde T, Besold TR, del Prado Martín FM. Using ontologies to enhance human understandability of global post-hoc explanations of black-box models. Artif Intell 2021;296:103471.
[61] Ancona M, Ceolini E, Öztireli C, Gross M. Gradient-based attribution methods. In: Explainable AI: interpreting, explaining and visualizing deep learning. Springer; 2019. p. 169–91.
[62] Balduzzi D, et al. The shattered gradients problem: if resnets are the answer, then what is the question?. In: arXiv preprint arXiv:1702.08591; 2017.
[63] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. ICLR; 2014.
[64] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: European conference on computer vision. Springer; 2014. p. 818–33.
[65] Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for simplicity: the all convolutional net. In: *arXiv e-prints* arXiv–1412; 2014.
[66] Nie W, Zhang Y, Patel A. A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. In: International Conference on Machine Learning. PMLR; 2018. p. 3809–18.
[67] Adebayo J, et al. Sanity checks for saliency maps. Adv Neural Inf Proces Syst 2018; 31.
[68] Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. Smoothgrad: removing noise by adding noise. In: *arXiv e-prints* arXiv–1706; 2017.
[69] Bach S, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One 2015;10:e0130140.
[70] Lipton P. Contrastive explanation. R Inst Philos Suppl 1990;27:247–66. https://doi.org/10.1017/S1358246100005130.
[71] Sayres R, et al. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. Ophthalmology 2019;126: 552–64.
[72] Recio-García JA, Parejas-Llanovarced H, Orozco-del Castillo MG, Brito-Borges EE. A case-based approach for the selection of explanation algorithms in image classification. In: International conference on case-based reasoning. Springer; 2021. p. 186–200.
[73] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2921–2929; 2016.
[74] Selvaraju RR, et al. Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. 618–626; 2017.
[75] Draelos RL, Carin L. Explainable multiple abnormality classification of chest ct volumes. Artif Intell Med 2022;132:102372.
[76] Yeh C-K, Kim J, Yen IE-H, Ravikumar PK. Representer point selection for explaining deep neural networks. Adv Neural Inf Proces Syst 2018;31.
[77] Page MJ, et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev 2021;10:1–11.
[78] Chowdhury ME, et al. Can ai help in screening viral and covid-19 pneumonia?. In: *arXiv e-prints* arXiv–2003; 2020.
[79] Bishop CM, Nasrabadi NM. Pattern recognition and machine learningvol. 4. Springer; 2006.
[80] Wilkinson MD, et al. The fair guiding principles for scientific data management and stewardship. Sci Data 2016;3:1–9.
[81] Adhikari A, et al. Towards fair explainable ai: a standardized ontology for mapping xai solutions to use cases, explanations, and ai systems. In: Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments; 2022. p. 562–8.
[82] Azarpanah A, et al. On the ethics of artificial intelligence. In: CSDH-SCHN 2020; 2020.
[83] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 2021;3: e745–50.