Review

A Scoping Review on the Progress, Applicability, and Future of Explainable Artificial Intelligence in Medicine

Raquel González-Alday *, Esteban García-Cuesta *, Victor Maojo , Casimir Kulikowski

Posted Date: 8 September 2023

doi: 10.20944/preprints202309.0581.v1

Keywords: artificial intelligence; medicine; explainable AI; interpretable AI

Article

A Scoping Review on the Progress, Applicability, and Future of Explainable Artificial Intelligence in Medicine

Raquel González-Alday [1,*,†], Esteban García-Cuesta [2,*], Victor Maojo [3] and Casimir A. Kulikowski [4]

[1] Department of Artificial Intelligence, Universidad Politécnica de Madrid, Madrid, Spain.; raquel.gonzalez.alday@alumnos.upm.es
[2] Department of Artificial Intelligence, Universidad Politécnica de Madrid, Madrid, Spain.; esteban.garcia@upm.es
[3] Department of Artificial Intelligence, Universidad Politécnica de Madrid, Madrid, Spain.; victor.maojo@upm.es
[4] Department of Computer Science, Rutgers University; kulikows@cs.rutgers.edu
* Corresponding authors: (R.G) rgalday@iib.uam.es; (E.G.) esteban.garcia@upm.es
† Current address: Instituto de Investigaciones Biomédicas "Alberto Sols" CSIC-UAM, Madrid, Spain.; rgalday@iib.uam.es

Abstract: Due to the success of artificial intelligence (AI) applications in the medical field over the past decade, concerns about the explainability of these systems have increased. The reliability requirements of black-box led algorithms for making decisions affecting patients pose a challenge even beyond their accuracy. Recent advances in AI increasingly underscore the need to incorporate explainability into these systems. While most traditional AI methods and expert systems are inherently interpretable, recent literature has focused primarily on explainability techniques for more complex models such as deep learning. This scoping review analyzes the existing literature on explainability and interpretability of AI methods in the medical and clinical field, providing an overview of past and current research trends, and limitations that might impede the development of Explainable Artificial Intelligence (XAI) in medicine, challenges, and possible research directions. In addition, this review discusses possible alternatives for leveraging medical knowledge to improve interpretability in clinical settings, while taking into account the needs of users.

Keywords: artificial intelligence; medicine; explainable AI; interpretable AI

1. Introduction

1.1. AI in medicine: opportunities and challenges

Today's Artificial Intelligence (AI) with its capability to automate and ease almost any kind of task, frequently appearing to surpass human performance, has made it a popular and widespread technology for many applications, especially over the last decade, and thanks to advances in deep learning (DL), with clinical healthcare being no exception.

Medicine has been one of the most challenging, but also most attention-getting application field for AI for the past five decades, with diagnostic decision support, the interpretation of medical images and clinical lab tests, drug development, patient management, and others, all demonstrating the broad and diverse scope of AI techniques applied to medical issues.

AI methods have promised a range of potential advantages for medical informatics systems. Automating burdensome tasks can be of great help, alleviating clinicians from unnecessary efforts and allowing them to focus on more important issues surrounding patient care. AI systems can perform these tasks with high precision and efficiency, and also, they can assist the extraction of relevant information from the large quantities of data being produced by modern medicine [1]. AI systems

 doi:10.20944/preprints202309.0581.v1

might be particularly beneficial in settings such as developing countries, where advanced medical expertise could not be accessed otherwise.

Therefore, as it has the potential to transform and improve healthcare, there is already plenty of research on AI applied to clinical decision support systems or medical imaging analysis for diagnosis and prognosis [2], as well as to other areas such as drug —including vaccines— development or healthcare organisation [3].

However, despite the promising opportunities that AI holds, the number of clinical AI applications currently deployed and actually used in real-world healthcare workflow is still relatively limited [4], although the US FDA has recently increased the number of approvals for AI-based medical applications. The reasons behind this "adoption gap" are complex, and depend not only on pure scientific and technological issues, but also on the special components of medical reasoning, ethical issues, organisational aspects and medical professionals' training and acceptance of novel procedures and tools, among others. One of the main problems, which we are going to focus on in this review, is the lack of transparency of the AI systems, which hinder their explainability and adoption.

When a novel AI algorithm is introduced as a tool to support medical decisions, that can directly influence people's lives and wellbeing, a number of questions —including ethical issues— arise concerning the trustworthiness of these systems and the possibilities of making theme more predictable [5].

These concerns become especially acute when using the classical, so-called "black box" AI algorithms. A recent, singular "mainstream" example is that of deep learning models that cannot provide direct, human understandable insights on the knowledge being used to achieve their final conclusions. Clinicians continue to be reluctant to trust the results of those kinds of systems that cannot easily provide detailed explanations [6]. Even clinicians with more advanced technological knowledge are also being concerned with related issues such as algorithmic biases or more technical aspects —such as overfitting or the quality of the original data— which are much harder to uncover without considerable understanding of the underlying mathematical and computational models, as well as the statistical assumptions behind in the actual implementations [7].

Finally, we can't finish this introductory section without talking about ChatGPT that has been openly released recently [1]. It is based on transformer deep neural network (DNN) architecture (GPT-3) that helps to encode and decode the input text, and its main purpose is to extend the input size to allow larger texts analysis and extract a latent space that contains the "meaning" of that text. Based on that latent space the system is able to search for sentence queries or generate related responses. It has a conversational type human-machine interface and language capabilities that allows an easy interaction in natural language extending the current query capabilities of web search tools. It is still unclear what are the main applications of this technology in medicine (it can be used to access similar previous patients' reports, take clinical decisions, documentation, summarize, etc.) but from the explainability perspective it can be considered as a technology based on deep neural networks. To show its current potential at [8] the authors demonstrate that Large Languge Models (LLMs) are able to approach or exceed the required threshold for passing the United States Medical Licensing Examination and its potential to generate novel insights that can assist human learners in a medical education setting. Yet there are no guarantees or explanations as to why it gives one answer or another [9].

1.2. The emergence of the field of XAI

Burgeoning expectations about AI coupled with the need to resolve or skirt around the above problems have led to the field known as eXplainable Artificial Intelligence (XAI). The latter deals with the different methods and approaches that enhance the transparency of AI models, whether by building and designing more efficient or alternative intrinsically interpretable models, or by providing

[1] ChatGPT was released November 20th, 2022 by OpenAI

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 September 2023 doi:10.20944/preprints202309.0581.v1

explanations for the "black box" models relying on statistical and optimization methods, by using auxiliary techniques [10]. Concerning medical applications, there is great hope that these XAI-based approaches might be one of the keys to improve understanding, trust and the final adoption of AI models in clinical practice [11,12].

Explainability and transparency are not new research topics for AI in medical applications. For consultation and decision support, in particular, the first expert systems developed in the 70s, such as MYCIN [13], tried to lay out the reasoning behind every decision by showing simple logical rules to the user. Other early systems such as CASNET [14] —codesigned by one of the coauthors of this paper— used complex causal models of human pathophysiology to support and justify its recommendations. Years after these pioneering expert systems, knowledge engineers have not yet been able to facilitate comprehensive and useful explanation facilities for knowledge engineering systems, which are still limited.

Lots of research was conducted on knowledge acquisition and knowledge representation over this first generation of AI systems —also medical—, and several impediments to creating knowledge bases from human expertise were identified [15]. These issues are strongly linked to the topic of explainability. In the pursuit of an understandable AI model that provides explanations to its decisions, we have to consider what kind of knowledge and how it should be represented in the explanation given in order to be sufficient and useful for the user (in the context of medicine, a physician).

1.3. Relevance of explainability in medical AI

Despite the difficulties, it is clear that the potential benefits of explainability and transparent models are quite a few, motivating the renewed interest in the field (Figure 1). Besides helping increase acceptance of AI models, enabling clinical professionals to take advantage of their conveniences and addressing ethical concerns, transparency could also allow better validation of these systems and facilitate troubleshooting during development [16]. Going even further, it could also be argued that transparent models could assist knowledge discovery in some cases, if able to understand the intrinsic relations hidden in the data found by AI models [17].

Figure 1. Different purposes and benefits of XAI approaches: technical, scientific and ethical reasons.

In addition, explainability might become a future legal requisite in fields such as medicine, since patients need to be properly informed about how their personal data is being used. Currently, some legal regulations, such as the GDPR (General Data Protection Regulation) in the European Union [18], are starting to include the requirement of transparent AI when dealing with sensitive data, pointing out that in the future, the use of explainability in AI-based systems might be furtherly enforced [19].

For those reasons, much recent literature is increasingly available appraising the benefits of XAI and presenting many different new methods which claim to be appropriate solutions. This is

 doi:10.20944/preprints202309.0581.v1

specially required in the field of healthcare, where responsibility and understanding become peculiarly important issues. However, there is persistent controversy in the field, since some researchers remain skeptical about current approaches, stating that capabilities of XAI may be overestimated, while several important limitations may be disregarded [20]. Such a controversy advocates for rigorous internal and external validation of AI models [21].

Such concerns need actually encourage further critical research in this area, as interpretability of AI methods remains as an open challenge but also a crucial need for the deployment of these systems, specially in areas such as clinical practice. This scoping review aims to be a starting point for those who want to get a wider and comprehensive view of the field of XAI applied to medicine.

There are already a few review articles on this topic, as shown in our preliminary research, either more general [22,23] or focused in more specific subfields [24–29]. Here we aim to present a broader perspective on the state of the art of XAI methods used in clinical applications, particularly taking into account the lessons learned from pioneering research in medical AI —since the 1970s— spanning until the recent boom of AI around Deep Learning, with the actual participation trhoughout these years of some of the authors. With such a direct and long perspective in time, we highlight not only the benefits but also the open challenges posed by interpretability based on medical AI models.

1.4. The influence of data quality

Before moving on to discuss the different methods used in explainable medical AI, we want to remark on one crucial aspect that can seriously affect our results, regardless of the used technique: the quality of the used data. When discussing explainability and transparency of AI models, the quality of the data analyzed and utilized in the machine learning process is critical to ensure that the generated output is trustworthy, transparent and does not lead to misinterpretation.

The scarcity and poor quality of available data is one of the biggest technical and methodological challenges in medical AI, and it is directly related to the issues of ethical trust, algorithmic vs. human expert bias and interpretability. Data based models have the advantage of not needing to explicitly capture and encode expert knowledge, which is quite difficult in most cases, and particularly in healthcare, as will be discussed in following sections of this review. However, poor quality or artifacts in measurement, gathering, or other processing of the data can make these models fragile and fail easily [30].

Predictive as well as explanatory model performance is strongly related to variations and quality of the dataset used for training in a machine learning context. An inappropriate dataset can make a model unable to generalise to unseen scenarios in deployment due to overfitting, caused by over or underrepresented classes in the training data, by wrong annotations or by other reasons [31]. Moreover, besides these issues of overfitting and bad performance, the issue of bias from various sources related to the data and its acquisition and subsequent selection or filtering through preprocessing are also central to explainability and its relation to the robustness of a model to changes in assumptions and future applicability scenarios.

One of the goals of XAI methods is to build transparent models in order to avoid algorithmic biases. However, biases are inevitably introduced in the application of models by the training data they rely on. Several reasons can be behind data bias, such as poorly represented populations or different conditions affecting data collection and model deployment [32]. In addition, data can incorporate human biases, and specially in the field of medicine, where the individuality and particularities of each patient and clinician will exert great influence [33].

Overfitting of models in training, and biases can have devastating consequences on the field of healthcare, but one of the potential benefits of XAI techniques is that they might be designed to help detect how and when these problems arise before deployment of faulty models [34]. That more emphasis should be put on data quality is obvious, such as checking the dataset's consistency, integrity and adequacy for the specific application as a crucial step for the development of a trustworthy and robust AI model [35]. Exhaustive data curation is needed for building large, well labelled and

 doi:10.20944/preprints202309.0581.v1

good quality datasets [36] and should go hand in hand with XAI approaches in order to avoid the aforementioned problems, as well as with exhaustive testing of AI systems before clinical deployment [37].

2. Objectives

The main goal of this review is to analyse the existing literature dealing with explainability and interpretability of AI methods applied to the medical field. We aim to provide an overview of current research trends and identify the challenges and limitations that this research area presents, as well as potential research directions and alternative approaches that could help improve trust and transparency for clinical AI applications.

There is much recent literature on explainability methods, with a special focus on deep learning models in the last few years. However, as mentioned before, the issue of interpretability has been addressed since the beginnings of AI and its application to medicine with models such as the first medical expert systems.

For that reason, besides giving an overview of current approaches, this review aims to cover some studies on explainability from classical AI methods as well for applications in medicine and the clinic.

In order to make concrete the objectives of this review, the following research questions are posed:

- **RQ1.** What are the XAI approaches being used in medical AI applications?
- **RQ2.** Does any technical limitation to explainability exist that might particularly hinder XAI development in medicine?
- **RQ3.** What are current research gaps and directions in the field of XAI that could be of special interest for medical applications?
- **RQ4.** Is it possible to include some kind of medical validation or knowledge to enhance the explainability of clinical AI systems?

3. Methodology

3.1. Search strategy

This scoping review has been conducted according to the framework and guidelines proposed by the Joanna Briggs Institute Manual for Evidence Synthesis [38]. An electronic search over three scientific research databases was conducted, namely Web of Science, PubMed and Scopus, looking for relevant literature on the topic of explainability and transparency of AI methods in the field of medicine. The search was conducted until February 2022, so articles up until that date were included in the systematic search.

To narrow down the search, several synonyms and related words were included so that each article title contained at least one term associated with medicine, another with artificial intelligence and another with explainability (Table 1).

 doi:10.20944/preprints202309.0581.v1

Table 1. Keywords used for literature search.

Explainable	XAI, explanation, explainability, interpretable, interpretability, trust, trustworthy, ethical, causality, understandable
Artificial Intelligence	AI, machine intelligence, computer intelligence, machine learning, deep learning, neural network, convolutional, computer vision, bayesian, black box, classifier, expert system, prediction model, algorithm, big data, data mining, knowledge discovery, pattern recognition, natural language processing, supervised learning, reinforcement learning
Medicine	Medical, clinical, health, healthcare, doctor, physician, diagnosis, prognosis, drug, illness, disease, cancer, mri, ct, treatment, therapy, patient, radiology, surgery, dermatology, hematology, oncology, cardiology, neurology, urology, psychiatry, immunology, immune, virus, bacteria

In addition to the systematic search, a few articles identified through the references of selected studies, recommended by the authors of this review or searched by hand for topics not thoroughly covered in the available literature were also included. Some of them were also selected to include recent advances not covered in the systematic search time range.

3.2. Inclusion and exclusion criteria

In order to meet the objectives of this review, the following criteria were defined for source inclusion:

- Directly related to the context of medicine and healthcare.
- Addressing clinical problems with AI algorithms, such as machine learning (ML), DL, computer vision or other kinds of either data or knowledge based methods.
- Including a clear focus on any kind of ethical or technical information about explainability, transparency or related issues.
- Published in English and with available full text.

A bottom time limit for the publications included was not established, as one of the objectives of this review was to cover some studies from the first medical AI systems (1970s -) to take into account the lessons learned from that period that might be of interest for current and future AI-focused systems. No constraints about the kind of publication were defined either: articles, reviews and studies of any kind were both as long as they met the inclusion criteria and were of interest for this review.

Articles that did not meet this criteria were excluded, such as those that used other types of computer approaches that do not fall under the scope of AI, those that did not talk about clinical applications or those that only included matters of explainability as a side topic without going into much detail.

3.3. Data extraction and analysis

The process of source of evidence screening conducted in this review are shown in Figure 2. After performing the systematic search and excluding duplicates between the results of the different databases, a first stage of selection was performed by analysing the title and abstract of the resulting papers. At this stage, all papers that were clearly not related to the topic of this review, either because they were not related to medicine, because they did not actually focus on AI methods, or because they did not discuss interpretability at all, as well as those that were not published in English.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 September 2023 doi:10.20944/preprints202309.0581.v1

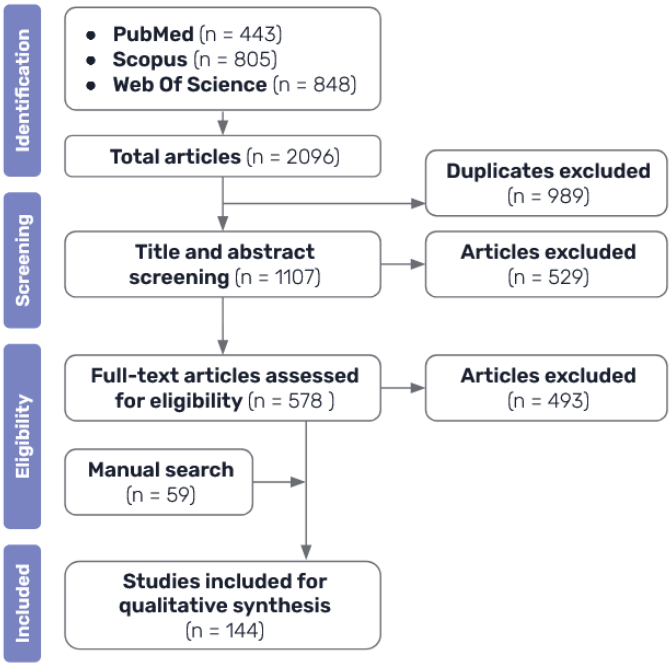

Figure 2. Source of evidence analysis and selection workflow.

Once all the potential candidate articles had been selected, a second stage of selection was performed to reach the final selection of literature included in this review. A more thorough assessment was made by looking into the full text papers, in order to exclude the ones that were of no interest for this review: those that were too short or general, articles that did not focus sufficiently on the topic of explainability, studies that were too similar to others already included or that employed identical techniques and did not add much value to the synthesis of the review, in addition to ones that had no available full text. This final selection of articles, plus the ones identified by manual search, make up the final bibliography included on this review.

3.4. Literature search results

The literature analysis carried out in this review corroborates the rising popularity over the past few years of the topics of explainability and interpretability on medical AI applications. Even though our search shows that some related studies were published in the 1980s, it is not until about 2017 that the number of publications starts to substantially increase, similar to what other related studies show [39]. Moreover, as expected, there is a special focus on explainability approaches to DL models, as they are currently the most popular ML methods.

Studies included in this review mainly consist of medical AI applications that employ explainability methods or that are built with interpretable models, but there are also some review papers and ethical commentaries. In what follows, we review the approaches and methods employed in the reviewed literature to make medical AI models explainable will be presented, as well other important issues covered in the various articles.

4. XAI methods in medicine

4.1. Classification of explainability approaches

Several taxonomies have been proposed in the literature to classify XAI methods and approaches, depending on different criteria [40–42] (Figure 3).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 September 2023 doi:10.20944/preprints202309.0581.v1

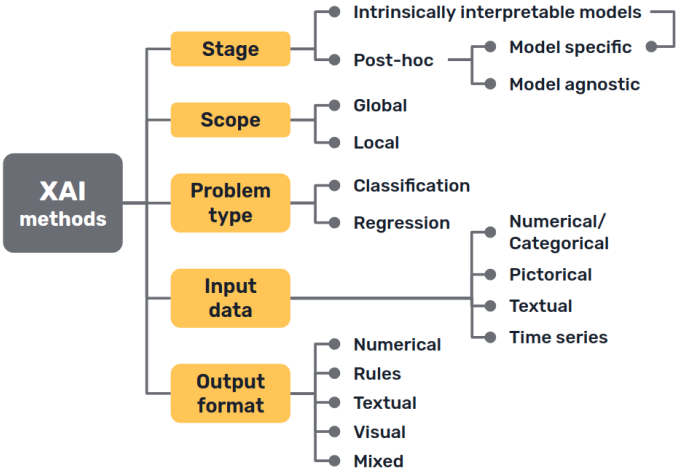

Figure 3. Several diverse approaches to the classification of XAI methods. Adapted from [42].

First, there is a clear distinction between auxiliary techniques that aim to provide explanations for either the model's prediction or its inner workings, which are commonly called post-hoc explainability methods, and AI models that are intrinsically interpretable, either because of their simplicity, the features they use or because they have a straightforward structure that is readily understandable by humans.

Secondly, we can distinguish interpretability techniques by their scope, where explanations provided by an XAI approach can be local, meaning that they refer to particular predictions of the model, or global, if they try to describe the behaviour of the model as a whole.

Other differentiations can be made between interpretability techniques that are model specific, because they have requirements regarding the kind of data or algorithm used, and model agnostic methods, that are general and can be applied in any case. Intrinsically interpretable models are model specific by definition, but post-hoc explainability methods can be generally seen as model agnostic, though some of them can have some requisites regarding the data or structure of the model.

More classifications can be made regarding how characteristics of the output explanations are displayed (textual, visual, rules...), the type of input data required, the type of problem they can be applied to [42] or how they are produced [23].

However, since these classifications overlap [43], in the following sections we have chosen to go over the different techniques included in the reviewed literature structured under the most popular taxonomy: interpretable models and post-hoc explainability methods.

4.2. Intrinsically interpretable models

Intrinsically interpretable models are those built using logical relations, statistical or probabilistic frameworks and similar strategies that represent human-interpretable systems, since they use rules, relationships or probabilities assigned to known variables.

This approach to transparency in AI, despite receiving less attention in recent years while the focus has been on DL, is historically the original one, and the perspective taken by knowledge-based systems.

4.2.1. Classical medical knowledge-based systems

Some knowledge-based systems, commonly known as "expert systems", are some of the classical AI models that were first developed at the end of the 1960s. Explanations were sometimes introduced as a feature of these first rule-based expert systems by design, as they were needed not only by users, but also by developers to troubleshoot their code during the design of these models. Thus, the importance of AI explainability has been discussed since the 1970's [44,45].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 September 2023 doi:10.20944/preprints202309.0581.v1

In medicine, many of these systems were developed aiming to be an aid for clinicians during diagnosis of patients and treatment assignment [46]. We must remind that explanations for patient cases are neither easily made, in many occasions, by human medical professionals. The most widely known of these classical models was MYCIN [13], but many more based on causal, taxonomic, and other networks of semantic relations such as CASNET, INTERNIST, the Present Illiness Program and others were designed to support rules by models of underlying knowledge that explained the rules and drove the inferences in clinical decision-making [14,47,48]. Subsequently, modeling of explanations was pursued explicitly for the MYCIN type of rule-based models [49]. The role of explanation was frequently recognized as a major aspect of expert systems [50–52].

As shown in Figure 4, expert systems consist in a knowledge base, containing the expertise captured from human experts in the field, usually in the form of rules, including both declarative or terminological knowledge and procedural knowledge of the domain [45]. This knowledge base is consulted by an inference algorithm when the user interacts with the system, and an explanation facility interacts with both the inference system and the knowledge base to construct the corresponding explaining statements [53].

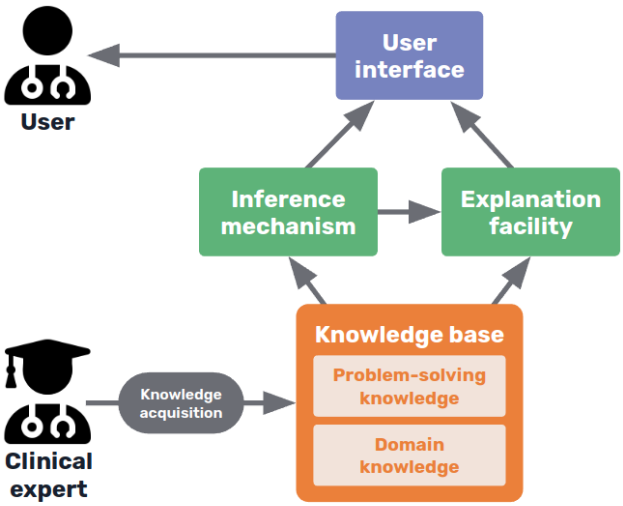

Figure 4. Basic diagram of a medical knowledge-based expert system. Adapted from [53].

The explanations provided were pre-modelled [54] and usually consisted of tracing the rules used by the inference mechanism to arrive at the final decision and presenting them in an intelligible way to the user. Such an approach aimed to explain how and why the system produced a diagnosis [52], and in some more sophisticated cases, even the causal evolution of the patient's clinical status [55].

However, these explanations were limited by the knowledge base constructed by the system's designers: all the justifications for knowledge had to be explicitly captured to produce specific explanations [52]. Knowledge acquisition (KA) and updating is a challenging task (see Section 6.2) and was not efficiently resolved, leading KA to become the bottleneck that resulted in the decline of interest on knowledge-based AI systems (as a main example of symbolic AI) in favour of data-based AI and ML systems (which we can call as subsymbolic AI).

4.2.2. Interpretable machine learning models

As an alternative to knowledge-based systems, from the early days of medical decision making, statistical Bayesian, Hypothesis-Testing, and linear discriminant models were ML models that can be considered interpretable. They are based on the statistical relationships extracted from clinical databases which allow formal probabilistic inferential methods to be applied. Ledley and Lusted proposed the Bayesian approach in their pioneering article in the journal Science in 1959 [56], with

 doi:10.20944/preprints202309.0581.v1

many of the alternatives first discussed in The Diagnostic Process conference [57]. Logistic regression is an effective statistical approach that can be used for classification and prediction (sendi2021). Generalized Linear Models (GLMs) [58] are also used in various problems among the literature [59,60], while Generalized Additive Models (GAMs), an extension of these, allow modeling of non-linear relationships and are used for prediction in medical problems as well [61,62].

Decision trees are considered transparent models because of their hierarchical structure which allows to easily visualise the logical processing of data in decision making processes. Moreover, a set of rules can be extracted to formalise that interpretation. They can be used for classification in the medical context [63,64], however, sometimes show poor generalisation capabilities, so it is most common to use tree ensembles (like the random forest algorithm [65]) that show better performance, in combination with post-hoc explainability methods, as they lose some interpretability [66–68,68,69].

Generalization of formal models through Bayesian networks [70], have become popular for modelling medical prediction problems [71–75], representing conditional dependencies between variables in the form of a graph, so that evidence can be propagated through the network to update the diagnostic or prognostic states of a patient [76]. This reasoning process can be easily visualised in an straightforward manner.

Interpretable models can be used by themselves, but another interesting strategy is using them in ensemble models. Ensemble models consist in combining several different ML methods to achieve a better performance and better interpretability than with a black-box model alone [77]. These approaches can also include these interpretable models in conjunction with DL models such as neural networks [78–81], as well as other post-hoc explainability techniques [82,83] such as the ones that will be presented later in this paper (Section 4.3). However, they pose increasingly and as yet unresolved complex interpretation issues as recently emphasized by Pearl [84].

4.2.3. Interpretation of neural network architectures

Despite the fact that neural networks cannot be fully included in the category of intrinsically interpretable models we can characterize them (also DL are included), such as architectures designed so that they resemble some of the simple neural modelling of brain function and used heuristically to recognise images or perform different tasks, and some neural network architectures have been specially designed to provide interpretability. The first type tries to mimic human decision-making where the decision is based on previously seen examples. In [85] a prototype learning design is presented to provide interpretable samples associated with the different types of respiratory sounds (normal, crackle, and wheeze). This technique learns a set of prototypes in a latent space that are used to make a prediction. Moreover it also allows for a new sample to be compared with the set of prototypes identifying the most similar and decode it to its original input representation. The architecture is based on the work in [86] and it intrinsically provides an automatic process to extract the input characteristics that are related to the associated prototype given that input.

Other methods' main motivation is to behave in a way more similar to how clinicians diagnose, and provide explanations in the form of relevant features. Among this type, attention maps are widely used. It basically extracts the influence of a feature on the output for a given sample. It is based on the gradients of the learned model and in [87] has been used to provide visual MRI explanations of liver lesions. For small datasets it is possible even to include some kind of medical knowledge as structural constraint rules over the attention maps during the process design [88]. Moreover the attention maps can be also applied at different scales concatenating feature maps as proposed at [89] being able to identify small structures on retina images.

These approaches are specific to DL but still surrogate models or post-hoc methods are applicable to add explainability.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 September 2023 doi:10.20944/preprints202309.0581.v1

4.3. Post-hoc explainability methods

Extending the above approach to transparency came with the development of more complex data-based ML methods, such as support vector machines (SVMs), tree ensembles and of course, DL techniques. The latter have become popular due to their impressive performance on a huge variety of tasks, sometimes even surpassing human accuracy for concrete applications, but also unfortunately entailing deeper opacity -for instance, than the detailed explanations that classic statistics can provide.

For this reason, different explainability methods have been proposed in order to shed light on the inner workings or algorithmic implementations used in these blackbox-like AI models. Because they are implemented as added facilities to these models, executed either over the results or the finished models, they are known as post-hoc methods, which produce post-hoc explanations, as opposed to the approach of intrinsically interpretable models.

Many of the approaches included in this category, which are also currently the most widely used as reported in the literature, are model agnostic. Post-hoc model agnostic methods are so popular due to their convenience: they are quick and easy to set up, flexible and well-established. Within this category, there are also some model specific post-hoc techniques designed to work only for a particular type of model.

These are less flexible, but tend to be faster and sometimes more accurate due to their specificity, as they can access the model internals and can produce different types of explanations that might be more suitable for some cases [23].

Regardless of their range of application, post-hoc methods can be also grouped by the basis of their functionalities. Combining the taxonomies proposed in [10] and [66], we can broadly differentiate between explanations by simplification (surrogate models), feature relevance methods, visualization techniques and example-based explanations. In the following sections these ideas will be presented as well as some of the most popular and representative methods belonging to each group.

4.3.1. Explanation by simplification

One way to explain a black-box model is to use a simpler, intrinsically interpretable model for the task of explaining its behaviour.

One method that uses this idea, which is undoubtedly one of the most employed ones throughout all the literature, is LIME (Local Interpretable Model-agnostic Explanations) [90]. This method builds a simple linear surrogate model to explain each of the predictions of the learned black-box model. The prediction's input to be explained is locally perturbed creating a new dataset that is used to build the explainable surrogate model. Explanation of instances can help to enforce trust in assisted AI clinical diagnosis within a patient diagnosis workflow [91].

Knowledge distillation is another technique included in this category. It was developed to compress neural networks for efficiency purposes, but it can also be used to construct a global surrogate interpretable model [92]. It consists of using the more complex black box model as a "teacher" to a simpler model that learns to mimic its output scores. If the "student" model demonstrates sufficient empirical performance, a domain expert may even prefer to use it in place of the teacher model and LIME. The main rationale behind this type of modelling is the assumption that some potential noise and error in the training data may affect the training efficacy of simple models. [93] used knowledge distillation to create an interpretable model achieving strong prediction performance for ICU outcome prediction.

Under this category we could also include techniques that attempt to simplify the models by extracting knowledge in a more comprehensive way. For example, rule extraction methods try to approximate the decision-making process of the black-box model, such as a neural network, with a set of rules or decision trees. Some of the methods try decomposing the units of the model to extract these rules [94], while others keep treating the original model as a black box and use the outcomes to perform a rule search [95]. There are also combinations of both approaches [96].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 September 2023 doi:10.20944/preprints202309.0581.v1

4.3.2. Explanation by Feature relevance methods

In the category of feature relevance methods we can find many popular examples of explainability techniques. These approaches try to find which are the most relevant variables or features to the model's predictions, those that influence the most the outcome in each case or in general.

The ancestry to these techniques can be found in both statistical and heuristic approaches dating back to the 1930's with Principal Component Analysis (PCA), which explains the weightings of features, or contributions to relevance in in terms of their contribution to inter and intra population patterns of multinomial variance and covariance [97]. These techniques were also shown to be central to both dimensionality reduction and its explanation in terms of information content for pattern recognition [98] and clinical diagnostic classification and prediction using subspace methods from atomic logic [99]. Later, related techniques for feature extraction by projection pursuit were developed and applied to clinical decision-making.

More recently, with LIME (that could also be included in this group), SHAP (SHapley Additive exPlanations) is one of the most widely used XAI model agnostic techniques, and it is the main example of the category of feature relevance methods. It is based on concepts from game theory that allow to compute which are the features that contribute the most to the outcomes of the black box model, by trying different feature set permutations [100]. SHAP explanations increase trust by helping to test prior knowledge and also can help to get insights into new ones [101].

Other well known similar example that measures the importance of different parts of the input by trying different changes is SA (Sensitivity Analysis) [102], and LRP (Layer-Wise Relevance Propagation) [103], Deep Taylor Decomposition (an evolution of LRP) [104] and DeepLIFT [105] are other model-specific alternatives for neural networks, that propagate the activation of neurons with respect to the inputs to compute feature importance.

4.3.3. Explanation by visualization techniques

Some of the aforementioned methods can produce visual explanations in some cases. Still, in this section we can mention some other methods that visualize directly the inner workings of the models, like Grad-CAM [106], that helps showing the activation of the layers of a convolutional neural network. In addition, there are other techniques that visualize the inputs and outputs of a model and the relationship between them, such as PDP (Partial Dependence Plots) [82] and ICE (Individual Conditional Expectation) plots [107]. It is worth mentioning that visualization can help to build explicable interfaces to interact with users, but it is complex to use them as an automatic step of the general explainability process.

4.3.4. Explanations by examples

Finally, another approach to produce explanations is to provide examples of other similar cases that help understanding why one instance has been classified as one object or structure or another by the model, or instead, dissimilar instances (counterfactuals) that might provide insights on why not.

For instance, MMD-critic [108] is an unsupervised algorithm that finds prototypes (the most representative instances of a class) as well as criticisms, instances that belong to a class but are not well represented by the prototypes. Another example are counterfactual explanations [109], that describe the minimum conditions that would lead to a different prediction by the model.

5. Evaluation of explainability

Despite the growing body of literature on different XAI methods and the rising interest on the topics of interpretability, explainability, and transparency, there is still limited research on the field of formal evaluations and measurements for these issues [110]. Most studies just employ XAI techniques without providing any kind of quantitative evaluation or appraisal of whether the produced explanations are appropriate.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 September 2023 doi:10.20944/preprints202309.0581.v1

Developing formal metrics and a more systematic evaluation of different methods can be difficult because of the variety of the available techniques and the lack of consensus on the definition of interpretability [111]. Moreover, contrary to usual performance metrics, there is no ground-truth when evaluating explanations of a black-box model [20,111]. However, this is foundational work of great importance, as such evaluation metrics would help towards not only assessing the quality of explanations and somehow measuring if the goal of interpretability is met, but also to compare between techniques and help standardise the different approaches, making it easier to select the most appropriate method for each case [112].

In short, there is a need for more robust metrics, standards and methodologies that help data scientists and engineers to integrate interpretability of in medical AI applications in a more detailed, verified, consistent and comparable way, along the whole methodology, design and algorithmic development process [113]. Nevertheless, in the few studies available on this topic, there are some common aspects that establish a starting point for further development, and there are some metrics such as robustness, consistency, comprehensibility, and importance of explanations.

A good and useful explanation for an AI model is one that is in accordance with human intuition and easy to understand [114]. To evaluate this, some qualitative and quantitative intuitions have already been proposed.

- On the one hand, **qualitative** intuitions include notions about the cognitive form, complexity and structure of the explanation. For example, what are the basic units that compose the explanation and how many are there (more units mean more complexity), how are they related (rules or hierarchies might be more interpretable for humans), if any uncertainty measure is provided or not, and so on [110].
- On the other hand, **quantitative** intuitions are easier to formally measure, and include, for example, notions like identity (for identical instances, explanations should be the same), stability (instances from the same class should have comparable explanations) or separability (distinct instances should have distinct explanations) [114,115]. Metrics based on these intuitions mathematically measure the similarity between explanations and instances as well as the agreement between the explainer and the black-box model.

Other options to evaluate XAI techniques include factors such as the time needed to output an explanation or the ability to detect bias in the data [114].

Another interesting strategy is to quantify the overlap between human intuitions (such as expert annotations) and the explanations obtained [116,117], or using human ratings by experts on the topic [112,118–120].

There are also different options regarding the context in which these various metrics can be used. The evaluation of an XAI system can be made either in the context of the final target task with the help of domain experts, in simpler tasks, or using formal definitions [110]. Depending on the specific characteristics of the problem and available resources, different approaches and metrics can be chosen.

6. Challenges of XAI approaches

As discussed before, there is an increasing recognition of XAI in recent years, particularly in fields like medicine, rediscovering previous concerns already raised for 50-60 years about statistical, pattern recognition, and earlier AI models, where the ethical requirements for transparency based on the Hippocratic Oath and derived professional ethics of practitioners is even stronger [121]. But, despite the benefits of using explainable systems being so clear and with some of the advances already made in the field and its range of available methods, the use of XAI techniques is still not widespread [10].

But why might explainability methods still not be enough? The problem of AI interpretability is a most challenging issue, as there are many unresolved technical limitations, related ethical questions and controversies surrounding current approaches.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 September 2023 doi:10.20944/preprints202309.0581.v1

6.1. Controversy around current XAI methods

Currently, one of the frequently discussed issues among researchers in the field of XAI is the choice between different approaches, mainly between post-hoc explainability and other transparency criteria for systems. There is no consensus about whether it is better to use intrinsically interpretable models or to develop techniques that try to explain the outcomes of black-box models, given the contrasts and complementarities between them that make comparisons incommensurate.

Despite being more understandable (and therefore meeting some of the transparency needs that make other types of systems more problematic), interpretable models are commonly rated as less accurate or efficient than more complex data-based approaches such as DL (see Figure 5), implying that there is a trade-off between explainability and classification or prediction accuracy [10,23,122].

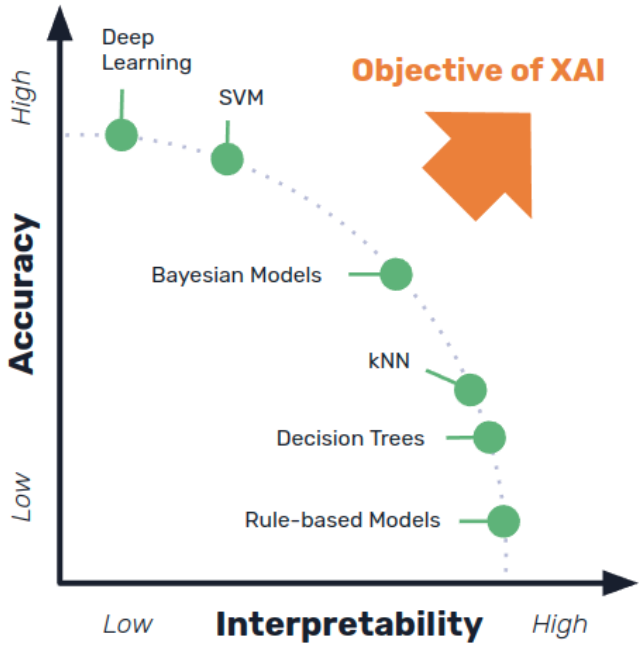

Figure 5. Trade-off between explainability and performance of different AI methods. Adapted from [66].

If such a trade-off exists, it could be alleged that it would not be ethical to use a model that does not perform at the best possible level, therefore being more adequate to use top-notch accurate systems, regardless of their black-box nature [7]. Some researchers maintain that, in fact, explainability might not be so necessary if good accuracy is proven by conducting empirical validations for example [123,124], while others disagree [125,126]. The practical ethical problem is the heterogeneity and lack of "gold standards" of comparable training and testing datasets for any particular problem with statistical and heuristic approaches, which do not take into account the wide range of qualitative differences between different types of data and knowledge, and the implementation of responsible evaluatory judgments by experts which cannot be reduced to simple measures. Complementary approaches for this might be official certifications or criteria for controlling data biases, beside explainability, in order to build trust on these systems [127,128].

Moreover, some scholars argue that, while it is important that research continues trying to unravel black-box models, we should also focus on pursuing more efficient and stable models, and putting so much emphasis on interpretability of these models before clinical application might restrain innovation in that direction and might not be that crucial [129].

However, from an opposite point of view, other researchers advocate for intrinsically interpretable models arguing that there is no such trade-off between explainability and accuracy in reality and that

 doi:10.20944/preprints202309.0581.v1

these models are more appropriate and safe for high-stakes fields like healthcare, where explanations of black-box models can be misleading and inaccurate, inducing effects opposite to the intended [125]. At the same time, some interpretable models have been recognized to require further explanations for non-technical end users such as clinicians or patients [24].

While there is no clear agreement on what is the best approach to the problem of transparency and interpretability or if it is even always necessary, some studies show that physicians, not surprisingly, do prefer AI systems that include some form of explainability, while seeming not to have a clear preference between methods [6].

In the following section, we will discuss further limitations of explanations related to this controversy, and that should be taken into account when designing these kinds of models and choosing among explainability approaches.

6.2. Technical limitations of explainability

While there is a common trend in favour of employing XAI among researchers, the limitations of these approaches are often underestimated, which can be especially high risk in the field of medicine [130]. These obstacles might not fully invalidate the use of XAI methods, but researchers and developers should be aware of them, as they can negate the application of these techniques in certain cases, and urgently require study [21].

Importantly, it should be noted that explanations and transparency might not translate into understanding underlying biomedical causal relationships [131], so we must be very cautious when using explainability as a tool to attempt knowledge discovery. For instance, using a surrogate model to explain a black-box system is only an approximation to the original model, and the produced explanations are unlikely to be faithful to the true inner workings of it [132]. Beyond this, it can not be assumed that behind the explanations or decisions of an AI system there is always an equivalent or at least a truly comparable reason that human experts could infer [133]. For that reason, we should be careful not to build systems that are persuasive rather than transparent [134,135], by properly analysing, validating and interpreting the results obtained.

Another point to be taken into account is that usually models and explanations that are more simple or compact are more interpretable, but it might not always be that way. Explanations depend on the specific context of the task and expectations of the user [44], and if they are not meaningful, detailed enough or in the needed form, they might not be useful and only complicate explanations of decision pathways [135].

It should also be considered that there can also be serious cognitive limits to explainability, as it was acknowledged about knowledge acquisition in the development of expert systems in the first decades of AI development. When human experts, like clinicians, become proficient in an area, they perform their tasks in a kind of automated or at least compiled way, effortlessly and efficiently. At this point, the knowledge needed for performing these tasks has become tacit, meaning that is compiled in their mind and not available for conscious access and sharing [15]. For that reason, sometimes they cannot completely explain what are the exact reasons behind their decisions.

This issue of after-the-fact justificatory explanations by experts makes it really hard to accurately or veridically model expert knowledge when the problem to solve is complex or vaguely defined, which is typical for the medical domain [30]. Unlike scientific disciplines such as biology, in clinical practice where reproducible experimentation is approximately feasible at least, there is a considerable uniquely patient-and-expert encounter specific practical expertise, and lack of generalizable knowledge for many practical problem-solving situations [136]. In consequence, it is really difficult to elicit reproducible and useful models of physician reasoning in the form of concrete rules [30].

This is one of the reasons that have contributed to making it impossible to fully capture the knowledge needed for building expert systems [15], and, ironically it presents parallels to the problems with black-box AI heuristic or statistical models [136], so maybe it is unrealistic to expect comprehensive explanations from these systems as we might desire, in the same way that we do not entirely understand

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 September 2023 doi:10.20944/preprints202309.0581.v1

much of human neural processes within their complex biochemical and genetic and developmental living contexts. Moreover, one must take into account that post-hoc explanations frequently suffer from the same problems of interpretability as human-expert explanations [20].

7. Research opportunities for medical XAI

The research limitations and shortcomings surrounding XAI discussed in this review, include the need for more robust evaluations (Section 5) and for more studies about cognitive limitations of explainability (Section 6), as well as efforts for improving data quality (Section 1.4). Here we discuss explainability as it is especially related to the field of medicine.

Before getting specific,it ought to be emphasized that much more interdisciplinary work would benefit the building of XAI systems. Not only do users' needs have to be taken into account more explicitly, but detecting errors in models that are related to a particular field of application has to be learned and imbued by technical designers, in order to improve current explainability techniques and develop and implement novel and more effective ones.

Lately, most of the studies in the field of interpretability tend to focus and develop some particular techniques, the ones that are more popular at the moment (for example, feature extractions or deep learning models visualisations), give less attention to other categories (such as interpretable models, for instance). Popular purely technological techniques are hardly most appropriate solutions for every case and, moreover, combining ideas from different fields could lead to really interesting advances to achieve more effective explanations [135].

7.1. Alternative ways to incorporate medical knowledge

We have already commented on the difficulties of capturing and modelling knowledge from experts, specially in healthcare, due to the uncertainty and incompleteness of knowledge in clinical practice [136]. However, it is undeniable that knowledge would help building more robust AI systems, with enhanced clinical interpretability. For that reason, research on alternative ways of embedding clinicians' expertise in AI models will be most useful.

One way of incorporating medical knowledge can involve human-in-the-loop evaluations of these systems [2]. Interaction with medical experts during development, or discussing the results of XAI models can help detect errors, validate these systems and identify possible underlying causes of the model's behaviour that would go unnoticed by technical developers. The use of the novel technique ChatGPT could also be helpful to automatize this process or part of it.

Another interesting research path is information fusion. Combining different types of multimodal data, such as medical images, family histories, genomic data or electronic health records can help specify, define and incorporate clinical context into a model, improving not only its diagnostic accuracy, but also its interpretability [28,137,138]. Moreover, using data collected from different centres can also help with domain variability and shift and enhance AI models [28].

There are already some studies on how to achieve this fusion of different kinds of data [137] as well as studies taking this kind of approach [139], however, more research and comprehensive data collection and annotation should be made to facilitate the building of these enhanced AI systems [28].

7.2. Taking into account user needs

Choosing the right XAI approach for each case is not a straightforward task, as the decision depends on many factors, such as characteristics of the concrete problem, of the application environment or of the available data, as different applications have different interpretability needs [111]. However, the most important aspect involves the requirement of the actual end-users of a system: clinicians and, in some cases, patients.

The questions from the users will vary depending on their expertise and knowledge of the domain, and their views on an application problem. A clinician using an AI system to confirm a diagnosis who wants to know if the model is working properly but has no technical knowledge will need different

 doi:10.20944/preprints202309.0581.v1

explanations than the system builder who wants to check the models' performance, or a patient who is using an AI system by themselves. For these reasons, the kind and extent of explanations need to be adjusted to the specific type of user needs in order to build trust [140], without falling into over-trust [1].

Depending on the prospective end users of the explanations, whether patients, clinicians or technical designers, it might be preferable to design different types of explainability: about the exhaustive workings of the model or about the relevant features being used, for example [17]. It should be identified in each case why users want explanations from the model, what information they need that said explanations contain and how they prefer them to be presented [141,142].

If the factors above are not considered, accurate explanations will not match the needs of the users and be informative and understandable to them [143]. Achieving this user understanding might as well need interaction between the system and its human user, in order to obtain further answers to different questions [10].

To summarize, in order to enable the collaboration between humans and AI, XAI outcomes have to be appropriately tailored to different end users, so more attention has to focus on these aspects of research: human-machine interaction and users' mental models. Most likely, general solutions will not be feasible, so the context of the problem has to be taken into account, preferably with interdisciplinary collaboration, and combining different types of explainability to fulfill users needs [144].

8. Conclusions

AI has the potential to transform and improve healthcare, nevertheless, without explainable and trustworthy systems, its application will continue to be limited.

In this paper, we have reviewed the precedents and background with the state of the art of XAI as applied in medicine. Several popular approaches and techniques have been discussed providing an overview of available options to include explainability as an aspect of clinical AI systems (**RQ1**).

The benefits of explainable systems could be considerable: promoting trust, enabling better interpretation of the data and predictions or enhancing the detection of errors, for example. However, considerable challenges identified in this review need to be overcome in the field of medical XAI. Besides the high dimensionality and black box nature of many AI models, in medicine the problem of data quality is especially serious, if we are to be able to develop accurate XAI techniques without the risk of being influenced by unrepresentative or poorly selected or curated and filtered data. Moreover, cognitive limitations to knowledge acquisition about clinicians' reasoning are also related to the extent of how far XAI methods might be able to prove useful (**RQ2**).

In order to develop and consolidate further robust explainability methods and interpretable models, making them a useful tool for clinicians to trust medical AI applications and therefore support their implementation in the real world, future research on this topic should be focused on overcoming these challenges, as well as better analysing user needs, enhancing human-system interaction and studying alternatives such as data fusion or clinicians' feedback to include medical validation and knowledge in different ways without the need of explicitly modelling it. Moreover, involving medical experts in the process of design and development of these systems would also help building more robust models and improving user understanding (**RQ3** and **RQ4**).

All in all, the prospects for XAI methods in clinical applications is that they are essential in many ways, but that further research is needed to overcome the current limitations enhancing these techniques in order to build secure, trustworthy and efficient systems that benefit both patients and clinicians.

Author Contributions: Conceptualization, García-Cuesta. E., Maojo. V. and Kulikowski. C.A.; methodology, González-Alday. R., García-Cuesta. E., González-Alday. R. and Maojo. V.; formal analysis, González-Alday. R. García-Cuesta. E.; investigation, Maojo. V., Kulikowski. C.A., García-Cuesta. E. and González-Alday. R.; writing—original draft preparation, González-Alday. R.; writing—review and editing, Kulikowski. C.A., Maojo. V., García-Cuesta. E.; supervision, Maojo. V. and García-Cuesta. E.; funding acquisition, Maojo. V. and García-Cuesta. E. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the IAX 'Inteligencia Artificial eXplicable' grant from the Comunidad de Madrid through the call Universidad Politénica de Madrid-Jóvenes Investigadores 2022/2024, and the Proyecto colaborativo de integración de datos genómicos (CICLOGEN) (No. PI17/01561) funded by the Carlos III Health Institute from the Spanish National Plan for Scientific and Technical Research and Innovation 2017-2020 and the European Regional Development Fund (FEDER).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Asan, O.; Bayrak, A.E.; Choudhury, A.; others. Artificial intelligence and human trust in healthcare: focus on clinicians. *Journal of Medical Internet Research* **2020**, *22*, e15154.
2. Holzinger, A. Explainable AI and Multi-Modal Causability in Medicine. *i-com* **2020**, *19*. doi:10.1515/icom-2020-0024.
3. Adadi, A.; Berrada, M. Explainable AI for healthcare: from black box to interpretable models. In *Embedded Systems and Artificial Intelligence*; Springer, 2020; pp. 327–337.
4. He, J.; Baxter, S.L.; Xu, J.; Xu, J.; Zhou, X.; Zhang, K. The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine* **2019**, *25*, 30–36.
5. Han, C.; Rundo, L.; Murao, K.; Nemoto, T.; Nakayama, H. Bridging the gap between AI and healthcare sides: towards developing clinically relevant AI-powered diagnosis systems. IFIP International Conference on Artificial Intelligence Applications and Innovations. Springer, 2020, pp. 320–333.
6. Diprose, W.; Buist, N.; Hua, N.; Thurier, Q.; Shand, G.; Robinson, R. Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *Journal of the American Medical Informatics Association* **2020**, *27*. doi:10.1093/jamia/ocz229.
7. Kerasidou, A. Ethics of artificial intelligence in global health: Explainability, algorithmic bias and trust. *Journal of Oral Biology and Craniofacial Research* **2021**, *11*. doi:10.1016/j.jobcr.2021.09.004.
8. TH, K.; M, C.; A, M.; C, S.; L, D.L.; C, E.; M, M.; R, A.; G, D.C.; J, M.; V., T. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health* **2023**, *2*, 164–174. doi:10.1371/journal.pdig.0000198.
9. Lee, P.; Goldberg, C.; Kohane, I. *The AI Revolution in Medicine: GPT-4 and Beyond*; Pearson, 2023.
10. Adadi, A.; Berrada, M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160.
11. Sethi, T.; Kalia, A.; Sharma, A.; Nagori, A. Interpretable artificial intelligence: Closing the adoption gap in healthcare. *Artificial Intelligence in Precision Health: From Concept to Applications* **2020**. doi:10.1016/B978-0-12-817133-2.00001-X.
12. Yoon, C.H.; Torrance, R.; Scheinerman, N. Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned? *Journal of Medical Ethics* **2021**. doi:10.1136/medethics-2020-107102.
13. Shortliffe, E.H.; Davis, R.; Axline, S.G.; Buchanan, B.G.; Green, C.C.; Cohen, S.N. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. *Computers and Biomedical Research* **1975**, *8*, 303–320.
14. Weiss, S.M.; Kulikowski, C.A.; Amarel, S.; Safir, A. A model-based method for computer-aided medical decision-making. *Artificial intelligence* **1978**, *11*, 145–172.
15. Musen, M.A. An overview of knowledge acquisition. *Second Generation Expert Systems* **1993**, pp. 405–427.
16. Cruz, H.F.d.; Pfahringer, B.; Schneider, F.; Meyer, A.; Schapranow, M.P. External validation of a "black-box" clinical predictive model in nephrology: can interpretability methods help illuminate performance differences? Conference on Artificial Intelligence in Medicine in Europe. Springer, 2019, pp. 191–201.
17. Vilone, G.; Longo, L. Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093* **2020**.
18. Goodman, B.; Flaxman, S. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine* **2017**, *38*, 50–57.
19. Amann, J.; Blasimme, A.; Vayena, E.; Frey, D.; Madai, V.; Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making* **2020**, *20*. doi:10.1186/s12911-020-01332-6.
20. Lipton, Z.C. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* **2018**, *16*, 31–57.

21. Ghassemi, M.; Oakden-Rayner, L.; Beam, A.L. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet. Digital health* **2021**, *3*. doi:10.1016/S2589-7500(21)00208-9.
22. Antoniadi, A.; Du, Y.; Guendouz, Y.; Wei, L.; Mazo, C.; Becker, B.; Mooney, C. Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Applied Sciences-Basel* **2021**, *11*. doi:10.3390/app11115088.
23. Abdullah, T.; Zahid, M.; Ali, W. A Review of Interpretable ML in Healthcare: Taxonomy, Applications, Challenges, and Future Directions. *Symmetry-Basel* **2021**, *13*. doi:10.3390/sym13122439.
24. Payrovnaziri, S.; Chen, Z.; Rengifo-Moreno, P.; Miller, T.; Bian, J.; Chen, J.; Liu, X.; He, Z. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *Journal of the American Medical Informatics Association* **2020**, *27*. doi:10.1093/jamia/ocaa053.
25. Gulum, M.; Trombley, C.; Kantardzic, M. A Review of Explainable Deep Learning Cancer Detection Models in Medical Imaging. *Applied Sciences-Basel* **2021**, *11*. doi:10.3390/app11104573.
26. Nazar, M.; Alam, M.; Yafi, E.; Su'ud, M. A Systematic Review of Human-Computer Interaction and Explainable Artificial Intelligence in Healthcare With Artificial Intelligence Techniques. *IEEE Access* **2021**, *9*. doi:10.1109/ACCESS.2021.3127881.
27. Salahuddin, Z.; Woodruff, H.; Chatterjee, A.; Lambin, P. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in Biology and Medicine* **2022**, *140*. doi:10.1016/j.compbiomed.2021.105111.
28. Yang, G.; Ye, Q.; Xia, J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information Fusion* **2022**, *77*. doi:10.1016/j.inffus.2021.07.016.
29. Zhang, Y.; Weng, Y.; Lund, J. Applications of Explainable Artificial Intelligence in Diagnosis and Surgery. *Diagnostics* **2022**, *12*, 237.
30. Quinn, T.P.; Senadeera, M.; Jacobs, S.; Coghlan, S.; Le, V. Trust and medical AI: the challenges we face and the expertise needed to overcome them. *Journal of the American Medical Informatics Association* **2021**, *28*, 890–894.
31. Subbaswamy, A.; Saria, S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics* **2020**, *21*, 345–352.
32. Brady, A.; Neri, E. Artificial Intelligence in Radiology-Ethical Considerations. *Diagnostics* **2020**, *10*. doi:10.3390/diagnostics10040231.
33. Maojo, V. Domain-specific particularities of data mining: Lessons learned. International Symposium on Biological and Medical Data Analysis. Springer, 2004, pp. 235–242.
34. Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 2015, pp. 1721–1730.
35. Gudivada, V.; Apon, A.; Ding, J. Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software* **2017**, *10*, 1–20.
36. Papadimitroulas, P.; Brocki, L.; Chung, N.C.; Marchadour, W.; Vermet, F.; Gaubert, L.; Eleftheriadis, V.; Plachouris, D.; Visvikis, D.; Kagadis, G.C.; others. Artificial intelligence: Deep learning in oncological radiomics and challenges of interpretability and data harmonization. *Physica Medica* **2021**, *83*, 108–121.
37. Chandler, C.; Foltz, P.W.; Elvevåg, B. Using machine learning in psychiatry: the need to establish a framework that nurtures trustworthiness. *Schizophrenia bulletin* **2020**, *46*, 11–14.
38. Peters, M.D.; Godfrey, C.; McInerney, P.; Munn, Z.; Tricco, A.C.; Khalil, H. Chapter 11: scoping reviews (2020 version). *JBI Manual for Evidence Synthesis, JBI* **2020**, *2020*.
39. Deshpande, N.; Gite, S. A Brief Bibliometric Survey of Explainable AI in Medical Field. *Library Philosophy and Practice* **2021**, *2021*.
40. Stiglic, G.; Kocbek, P.; Fijacko, N.; Zitnik, M.; Verbert, K.; Cilar, L. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery* **2020**, *10*. doi:10.1002/widm.1379.
41. Singh, A.; Sengupta, S.; Lakshminarayanan, V. Explainable Deep Learning Models in Medical Image Analysis. *Journal of Imaging* **2020**, *6*. doi:10.3390/jimaging6060052.
42. Vilone, G.; Longo, L. Classification of explainable artificial intelligence methods through their output formats. *Machine Learning and Knowledge Extraction* **2021**, *3*, 615–661.

43. Arya, V.; Bellamy, R.K.; Chen, P.Y.; Dhurandhar, A.; Hind, M.; Hoffman, S.C.; Houde, S.; Liao, Q.V.; Luss, R.; Mojsilović, A.; others. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012* **2019**.
44. Biran, O.; Cotton, C. Explanation and justification in machine learning: A survey. IJCAI-17 workshop on explainable AI (XAI), 2017, Vol. 8, pp. 8–13.
45. Preece, A. Asking 'Why'in AI: Explainability of intelligent systems–perspectives and challenges. *Intelligent Systems in Accounting, Finance and Management* **2018**, *25*, 63–72.
46. Vourgidis, I.; Mafuma, S.J.; Wilson, P.; Carter, J.; Cosma, G. Medical expert systems–a study of trust and acceptance by healthcare stakeholders. UK Workshop on Computational Intelligence. Springer, 2018, pp. 108–119.
47. Miller, R.A.; Pople Jr, H.E.; Myers, J.D. Internist-I, an experimental computer-based diagnostic consultant for general internal medicine. In *Computer-assisted medical decision making*; Springer, 1985; pp. 139–158.
48. Long, W.; Naimi, S.; Criscitiello, M.; Pauker, S.; Szolovits, P. An aid to physiological reasoning in the management of cardiovascular disease. 1984 Computers in Cardiology Conference. IEEE Computer Society Long Beach, CA, 1984, pp. 3–6.
49. Clancey, W.J.; Shortliffe, E.H. *Readings in medical artificial intelligence: the first decade*; Addison-Wesley Longman Publishing Co., Inc., 1984.
50. Ford, K.M.; Coffey, J.W.; Cañas, A.; Andrews, E.J. Diagnosis and explanation by a nuclear cardiology expert system. *International Journal of Expert Systems* **1996**, *9*, 4.
51. Hogan, W.R.; Wagner, M.M. The use of an explanation algorithm in a clinical event monitor. Proceedings of the AMIA Symposium. American Medical Informatics Association, 1999, p. 281.
52. Darlington, K. Using explanation facilities in healthcare expert systems. *HEALTHINF 2008: Proceedings of the First International Conference on Health Informatics, vol 1* **2008**.
53. Darlington, K.W. Designing for explanation in health care applications of expert systems. *Sage Open* **2011**, *1*, 2158244011408618.
54. Rennels, G.D.; Shortliffe, E.H.; Miller, P.L. Choice and explanation in medical management: a multiattribute model of artificial intelligence approaches. *Medical Decision Making* **1987**, *7*, 22–31.
55. Molino, G.; Console, L.; Torasso, P. Causal expert systems supporting medical decision making and medical education: explanations based on simulated situations. Images of the Twenty-First Century. Proceedings of the Annual International Engineering in Medicine and Biology Society,. IEEE, 1989, pp. 1827–1828.
56. Ledley, R.S.; Lusted, L.B. Reasoning foundations of medical diagnosis: symbolic logic, probability, and value theory aid our understanding of how physicians reason. *Science* **1959**, *130*, 9–21.
57. Jacquez, J. The Diagnostic Process: Proceedings of a Conference Held at the University of Michigan. *Malloy Lithographing, Inc., Ann Arbor, Michigan* **1963**.
58. Nelder, J.A.; Wedderburn, R.W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)* **1972**, *135*, 370–384.
59. Meacham, S.; Isaac, G.; Nauck, D.; Virginas, B. Towards explainable AI: Design and development for explanation of machine learning predictions for a patient readmittance medical application. Intelligent Computing-Proceedings of the Computing Conference. Springer, 2019, pp. 939–955.
60. Banegas-Luna, A.J.; Peña-García, J.; Iftene, A.; Guadagni, F.; Ferroni, P.; Scarpato, N.; Zanzotto, F.M.; Bueno-Crespo, A.; Pérez-Sánchez, H. Towards the interpretability of machine learning predictions for medical applications targeting personalised therapies: a cancer case survey. *International Journal of Molecular Sciences* **2021**, *22*, 4394.
61. Karatekin, T.; Sancak, S.; Celik, G.; Topcuoglu, S.; Karatekin, G.; Kirci, P.; Okatan, A. Interpretable machine learning in healthcare through generalized additive model with pairwise interactions (GA2M): predicting severe retinopathy of prematurity. 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML). IEEE, 2019, pp. 61–66.
62. Wang, H.; Huang, Z.; Zhang, D.; Arief, J.; Lyu, T.; Tian, J. Integrating co-clustering and interpretable machine learning for the prediction of intravenous immunoglobulin resistance in kawasaki disease. *IEEE Access* **2020**, *8*, 97064–97071.
63. Itani, S.; Rossignol, M.; Lecron, F.; Fortemps, P. Towards interpretable machine learning models for diagnosis aid: a case study on attention deficit/hyperactivity disorder. *PloS one* **2019**, *14*, 1–20.

64. Brito-Sarracino, T.; dos Santos, M.R.; Antunes, E.F.; de Andrade Santos, I.B.; Kasmanas, J.C.; de Leon Ferreira, A.C.P.; others. Explainable machine learning for breast cancer diagnosis. 2019 8th Brazilian Conference on Intelligent Systems (BRACIS). IEEE, 2019, pp. 681–686.
65. Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5–32.
66. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; others. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* **2020**, *58*, 82–115.
67. Mattogno, P.P.; Caccavella, V.M.; Giordano, M.; D'Alessandris, Q.G.; Chiloiro, S.; Tariciotti, L.; Olivi, A.; Lauretti, L. Interpretable Machine Learning–Based Prediction of Intraoperative Cerebrospinal Fluid Leakage in Endoscopic Transsphenoidal Pituitary Surgery: A Pilot Study. *Journal of Neurological Surgery Part B: Skull Base* **2022**.
68. Alsinglawi, B.; Alshari, O.; Alorjani, M.; Mubin, O.; Alnajjar, F.; Novoa, M.; Darwish, O. An explainable machine learning framework for lung cancer hospital length of stay prediction. *Scientific reports* **2022**, *12*, 1–10.
69. El-Sappagh, S.; Alonso, J.M.; Islam, S.; Sultan, A.M.; Kwak, K.S. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. *Scientific reports* **2021**, *11*, 1–26.
70. Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*; Morgan kaufmann, 1988.
71. Chang, C.C.; Cheng, C.S. A Bayesian decision analysis with fuzzy interpretability for aging chronic disease. *International Journal of Technology Management* **2007**, *40*, 176–191.
72. Casini, L.; McKay Illari, P.; Russo, F.; Williamson, J. Recursive Bayesian nets for prediction, explanation and control in cancer science. *Theoria* **2011**, *26*, 495–4548.
73. Kyrimi, E.; Marsh, W. A progressive explanation of inference in 'hybrid' Bayesian networks for supporting clinical decision making. Conference on Probabilistic Graphical Models. PMLR, 2016, pp. 275–286.
74. Xie, W.; Ji, M.; Zhao, M.; Zhou, T.; Yang, F.; Qian, X.; Chow, C.Y.; Lam, K.Y.; Hao, T. Detecting symptom errors in neural machine translation of patient health information on depressive disorders: developing interpretable bayesian machine learning classifiers. *Frontiers in Psychiatry* **2021**, *12*.
75. Yun, J.; Basak, M.; Han, M.M. Bayesian rule modeling for interpretable mortality classification of covid-19 patients. *Cmc-Computers Materials & Continua* **2021**, pp. 2827–2843.
76. Kyrimi, E.; Mossadegh, S.; Tai, N.; Marsh, W. An incremental explanation of inference in Bayesian networks for increasing model trustworthiness and supporting clinical decision making. *Artificial intelligence in medicine* **2020**, *103*, 101812.
77. Kanda, E.; Epureanu, B.I.; Adachi, T.; Tsuruta, Y.; Kikuchi, K.; Kashihara, N.; Abe, M.; Masakane, I.; Nitta, K. Application of explainable ensemble artificial intelligence model to categorization of hemodialysis-patient and treatment using nationwide-real-world data in Japan. *Plos one* **2020**, *15*, 1–23.
78. Chen, J.; Dai, X.; Yuan, Q.; Lu, C.; Huang, H. Towards interpretable clinical diagnosis with Bayesian network ensembles stacked on entity-aware CNNs. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3143–3153.
79. Ahmed, Z.U.; Sun, K.; Shelly, M.; Mu, L. Explainable artificial intelligence (XAI) for exploring spatial variability of lung and bronchus cancer (LBC) mortality rates in the contiguous USA. *Scientific reports* **2021**, *11*, 1–15.
80. Singh, R.K.; Pandey, R.; Babu, R.N. COVIDScreen: explainable deep learning framework for differential diagnosis of COVID-19 using chest X-rays. *Neural Computing and Applications* **2021**, *33*, 8871–8892.
81. Yu, T.H.; Su, B.H.; Battalora, L.C.; Liu, S.; Tseng, Y.J. Ensemble modeling with machine learning and deep learning to provide interpretable generalized rules for classifying CNS drugs with high prediction power. *Briefings in Bioinformatics* **2022**, *23*, bbab377.
82. Peng, J.; Zou, K.; Zhou, M.; Teng, Y.; Zhu, X.; Zhang, F.; Xu, J. An explainable artificial intelligence framework for the deterioration risk prediction of hepatitis patients. *Journal of Medical Systems* **2021**, *45*, 1–9.
83. Kim, S.H.; Jeon, E.T.; Yu, S.; Oh, K.; Kim, C.K.; Song, T.J.; Kim, Y.J.; Heo, S.H.; Park, K.Y.; Kim, J.M.; others. Interpretable machine learning for early neurological deterioration prediction in atrial fibrillation-related stroke. *Scientific reports* **2021**, *11*, 1–9.
84. Pearl, J.; Mackenzie, D. *The book of why: the new science of cause and effect*; Basic books, 2018.

85. Ren, Z.; Nguyen, T.T.; Nejdl, W. Prototype learning for interpretable respiratory sound analysis. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 9087–9091.
86. Li, O.; Liu, H.; Chen, C.; Rudin, C. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, Vol. 32.
87. Wan, Y.; Zheng, Z.; Liu, R.; Zhu, Z.; Zhou, H.; Zhang, X.; Boumaraf, S. A Multi-Scale and Multi-Level Fusion Approach for Deep Learning-Based Liver Lesion Diagnosis in Magnetic Resonance Images with Visual Explanation. *Life* **2021**, *11*, 582.
88. Xu, Y.; Hu, M.; Liu, H.; Yang, H.; Wang, H.; Lu, S.; Liang, T.; Li, X.; Xu, M.; Li, L.; others. A hierarchical deep learning approach with transparency and interpretability based on small samples for glaucoma diagnosis. *NPJ digital medicine* **2021**, *4*, 1–11.
89. Liao, W.; Zou, B.; Zhao, R.; Chen, Y.; He, Z.; Zhou, M. Clinical interpretable deep learning model for glaucoma diagnosis. *IEEE journal of biomedical and health informatics* **2019**, *24*, 1405–1412.
90. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
91. Magesh, P.R.; Myloth, R.D.; Tom, R.J. An explainable machine learning model for early detection of Parkinson's disease using LIME on DaTSCAN imagery. *Computers in Biology and Medicine* **2020**, *126*, 104041.
92. Tan, S.; Caruana, R.; Hooker, G.; Lou, Y. Distill-and-compare: Auditing black-box models using transparent model distillation. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018, pp. 303–310.
93. Che, Z.; Purushotham, S.; Khemani, R.; Liu, Y. Interpretable deep models for ICU outcome prediction. AMIA annual symposium proceedings. American Medical Informatics Association, 2016, Vol. 2016, p. 371.
94. Krishnan, R.; Sivakumar, G.; Bhattacharya, P. A search technique for rule extraction from trained neural networks. *Pattern Recognition Letters* **1999**, *20*, 273–280.
95. Etchells, T.A.; Lisboa, P.J. Orthogonal search-based rule extraction (OSRE) for trained neural networks: a practical and efficient approach. *IEEE transactions on neural networks* **2006**, *17*, 374–384.
96. Barakat, N.; Diederich, J. Eclectic rule-extraction from support vector machines. *International Journal of Computational Intelligence* **2005**, *2*, 59–62.
97. Fisher, R.A. The logic of inductive inference. *Journal of the royal statistical society* **1935**, *98*, 39–82.
98. Kaminuma, T.; Takekawa, T.; Watanabe, S. Reduction of clustering problem to pattern recognition. *Pattern Recognition* **1969**, *1*, 195–205.
99. Kulikowski, C.A. Pattern recognition approach to medical diagnosis. *IEEE Transactions on Systems Science and Cybernetics* **1970**, *6*, 173–178.
100. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **2017**, *30*.
101. Weis, C.; Cuénod, A.; Rieck, B.; Dubuis, O.; Graf, S.; Lang, C.; Oberle, M.; Brackmann, M.; Søgaard, K.K.; Osthoff, M.; others. Direct antimicrobial resistance prediction from clinical MALDI-TOF mass spectra using machine learning. *Nature Medicine* **2022**, *28*, 164–174.
102. Saltelli, A. Sensitivity analysis for importance assessment. *Risk analysis* **2002**, *22*, 579–590.
103. Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.R.; Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* **2015**, *10*, e0130140.
104. Montavon, G.; Lapuschkin, S.; Binder, A.; Samek, W.; Müller, K.R. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition* **2017**, *65*, 211–222.
105. Shrikumar, A.; Greenside, P.; Kundaje, A. Learning important features through propagating activation differences. International conference on machine learning. PMLR, 2017, pp. 3145–3153.
106. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.
107. Goldstein, A.; Kapelner, A.; Bleich, J.; Pitkin, E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics* **2015**, *24*, 44–65.

108. Kim, B.; Khanna, R.; Koyejo, O.O. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems* **2016**, *29*.
109. Wachter, S.; Mittelstadt, B.; Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* **2017**, *31*, 841.
110. Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* **2017**.
111. Markus, A.; Kors, J.; Rijnbeek, P. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics* **2021**, *113*. doi:10.1016/j.jbi.2020.103655.
112. Kaur, D.; Uslu, S.; Durresi, A.; Badve, S.; Dundar, M. Trustworthy Explainability Acceptance: A New Metric to Measure the Trustworthiness of Interpretable AI Medical Diagnostic Systems. *Complex, Intelligent and Software Intensive Systems, CISIS-2021* **2021**, *278*. doi:10.1007/978-3-030-79725-6_4.
113. Kolyshkina, I.; Simoff, S. Interpretability of Machine Learning Solutions in Public Healthcare: The CRISP-ML Approach. *Frontiers in Big Data* **2021**, *4*. doi:10.3389/fdata.2021.660206.
114. ElShawi, R.; Sherif, Y.; Al-Mallah, M.; Sakr, S. Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Computational Intelligence* **2021**, *37*. doi:10.1111/coin.12410.
115. Honegger, M.R. Shedding light on black box machine learning algorithms. *Development of an Axiomatic Framework to Assess the Quality of Methods that Explain Individual Predictions, MA Karlsruhe* **2018**.
116. Muddamsetty, S.M.; Jahromi, M.N.; Moeslund, T.B. Expert level evaluations for explainable AI (XAI) methods in the medical domain. International Conference on Pattern Recognition. Springer, 2021, pp. 35–46.
117. de Souza, L.; Mendel, R.; Strasser, S.; Ebigbo, A.; Probst, A.; Messmann, H.; Papa, J.; Palm, C. Convolutional Neural Networks for the evaluation of cancer in Barrett's esophagus: Explainable AI to lighten up the black-box. *Computers in Biology and Medicine* **2021**, *135*. doi:10.1016/j.compbiomed.2021.104578.
118. Kumarakulasinghe, N.B.; Blomberg, T.; Liu, J.; Leao, A.S.; Papapetrou, P. Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models. 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS). IEEE, 2020, pp. 7–12.
119. Singh, A.; Balaji, J.; Rasheed, M.; Jayakumar, V.; Raman, R.; Lakshminarayanan, V. Evaluation of Explainable Deep Learning Methods for Ophthalmic Diagnosis. *Clinical Ophthalmology* **2021**, *15*. doi:10.2147/OPTH.S312236.
120. Deperlioglu, O.; Kose, U.; Gupta, D.; Khanna, A.; Giampaolo, F.; Fortino, G. Explainable framework for Glaucoma diagnosis by image processing and convolutional neural network synergy: Analysis with doctor evaluation. *Future Generation Computer Systems* **2022**, *129*. doi:10.1016/j.future.2021.11.018.
121. Kulikowski, C.A. Pandemics: historically slow "learning curve" leading to biomedical informatics and vaccine breakthroughs. *Yearbook of medical informatics* **2021**, *30*, 290–301.
122. Došilović, F.K.; Brčić, M.; Hlupić, N. Explainable artificial intelligence: A survey. 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO). IEEE, 2018, pp. 0210–0215.
123. Durán, J.M.; Jongsma, K.R. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics* **2021**. doi:10.1136/medethics-2020-106820.
124. McCoy, L.G.; Brenna, C.T.; Chen, S.S.; Vold, K.; Das, S. Believing in black boxes: Machine learning for healthcare does not need explainability to be evidence-based. *Journal of Clinical Epidemiology* **2022**, *142*, 252–257.
125. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **2019**, *1*, 206–215.
126. Véliz, C.; Prunkl, C.; Phillips-Brown, M.; Lechterman, T.M. We might be afraid of black-box algorithms. *Journal of Medical Ethics* **2021**, *47*, 339–340.
127. Reimer, U.; Maier, E.; Tödtli, B. Going Beyond Explainability in Medical AI Systems. Modellierung (Companion), 2020, pp. 185–191.
128. Reimer, U.; Tödtli, B.; Maier, E. How to Induce Trust in Medical AI Systems. International Conference on Conceptual Modeling. Springer, 2020, pp. 5–14.
129. Wang, F.; Kaushal, R.; Khullar, D. Should health care demand interpretable artificial intelligence or accept "black box" medicine?, 2020.

130. Babic, B.; Gerke, S.; Evgeniou, T.; Cohen, I.G. Beware explanations from AI in health care. *Science* **2021**, *373*, 284–286.
131. Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K.; Muller, H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery* **2019**, *9*. doi:10.1002/widm.1312.
132. Petch, J.; Di, S.; Nelson, W. Opening the black box: the promise and limitations of explainable machine learning in cardiology. *Canadian Journal of Cardiology* **2021**.
133. Tjoa, E.; Guan, C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems* **2021**, *32*. doi:10.1109/TNNLS.2020.3027314.
134. Herman, B. The promise and peril of human evaluation for model interpretability. *arXiv preprint arXiv:1711.07414* **2017**.
135. Gilpin, L.H.; Bau, D.; Yuan, B.Z.; Bajwa, A.; Specter, M.; Kagal, L. Explaining explanations: An overview of interpretability of machine learning. 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). IEEE, 2018, pp. 80–89.
136. London, A. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report* **2019**, *49*. doi:10.1002/hast.973.
137. Huang, S.C.; Pareek, A.; Seyyedi, S.; Banerjee, I.; Lungren, M.P. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ Digital Medicine* **2020**, *3*, 1–9.
138. Holzinger, A.; Dehmer, M.; Emmert-Streib, F.; Cucchiara, R.; Augenstein, I.; Del Ser, J.; Samek, W.; Jurisica, I.; Diaz-Rodriguez, N. Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. *Information Fusion* **2022**, *79*. doi:10.1016/j.inffus.2021.10.007.
139. Kamal, M.S.; Northcote, A.; Chowdhury, L.; Dey, N.; Crespo, R.G.; Herrera-Viedma, E. Alzheimer's patient analysis using image and gene expression data and explainable-ai to present associated genes. *IEEE Transactions on Instrumentation and Measurement* **2021**, *70*, 1–7.
140. Larasati, R.; De Liddo, A.; Motta, E. AI Healthcare System Interface: Explanation Design for Non-Expert User Trust. ACMIUI-WS 2021: Joint Proceedings of the ACM IUI 2021 Workshops. CEUR Workshop Proceedings, 2021, Vol. 2903.
141. Barda, A.J.; Horvat, C.M.; Hochheiser, H. A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. *BMC medical informatics and decision making* **2020**, *20*, 1–16.
142. Hwang, J.; Lee, T.; Lee, H.; Byun, S.; others. A Clinical Decision Support System for Sleep Staging Tasks With Explanations From Artificial Intelligence: User-Centered Design and Evaluation Study. *Journal of medical Internet research* **2022**, *24*, e28659.
143. Cutillo, C.; Sharma, K.; Foschini, L.; Kundu, S.; Mackintosh, M.; Mandl, K.; Beck, T.; Collier, E.; Colvis, C.; Gersing, K.; Gordon, V.; Jensen, R.; Shabestari, B.; Southall, N.; Hlthcare Workshop Working Grp. Machine intelligence in healthcare-perspectives on trustworthiness, explainability, usability, and transparency. *NPJ Digital Medicine* **2020**, *3*. doi:10.1038/s41746-020-0254-2.
144. Sariyar, M.; Holm, J. Medical Informatics in a Tension Between Black-Box AI and Trust. *Studies in Health Technology and Informatics* **2022**, *289*. doi:10.3233/SHTI210854.