applied sciences

MDPI

Review

Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review

Anna Markella Antoniadi [1,2], Yuhan Du [1], Yasmine Guendouz [3,4], Lan Wei [1], Claudia Mazo [1,5], Brett A. Becker [1] and Catherine Mooney [1,2,*]

[1] UCD School of Computer Science, University College Dublin, Dublin 4, Ireland; anna.antoniadi@ucdconnect.ie (A.M.A.); yuhan.du@ucdconnect.ie (Y.D.); lan.wei@ucdconnect.ie (L.W.); claudia.mazovargas@ucd.ie (C.M.); brett.becker@ucd.ie (B.A.B.)

[2] FutureNeuro SFI Research Centre, Royal College of Surgeons in Ireland, Dublin 2, Ireland

[3] Trinity Centre for Biomedical Engineering, Trinity College Dublin, Dublin 2, Ireland; GUENDOUY@tcd.ie

[4] Department of Mechanical, Manufacturing & Biomedical Engineering, School of Engineering, Trinity College Dublin, Dublin 2, Ireland

[5] CeADAR: Ireland's Centre for Applied AI, Dublin 4, Ireland

* Correspondence: catherine.mooney@ucd.ie

Abstract: Machine Learning and Artificial Intelligence (AI) more broadly have great immediate and future potential for transforming almost all aspects of medicine. However, in many applications, even outside medicine, a lack of transparency in AI applications has become increasingly problematic. This is particularly pronounced where users need to interpret the output of AI systems. Explainable AI (XAI) provides a rationale that allows users to understand why a system has produced a given output. The output can then be interpreted within a given context. One area that is in great need of XAI is that of Clinical Decision Support Systems (CDSSs). These systems support medical practitioners in their clinic decision-making and in the absence of explainability may lead to issues of under or over-reliance. Providing explanations for how recommendations are arrived at will allow practitioners to make more nuanced, and in some cases, life-saving decisions. The need for XAI in CDSS, and the medical field in general, is amplified by the need for ethical and fair decision-making and the fact that AI trained with historical data can be a reinforcement agent of historical actions and biases that should be uncovered. We performed a systematic literature review of work to-date in the application of XAI in CDSS. Tabular data processing XAI-enabled systems are the most common, while XAI-enabled CDSS for text analysis are the least common in literature. There is more interest in developers for the provision of local explanations, while there was almost a balance between post-hoc and ante-hoc explanations, as well as between model-specific and model-agnostic techniques. Studies reported benefits of the use of XAI such as the fact that it could enhance decision confidence for clinicians, or generate the hypothesis about causality, which ultimately leads to increased trustworthiness and acceptability of the system and potential for its incorporation in the clinical workflow. However, we found an overall distinct lack of application of XAI in the context of CDSS and, in particular, a lack of user studies exploring the needs of clinicians. We propose some guidelines for the implementation of XAI in CDSS and explore some opportunities, challenges, and future research needs.

Keywords: artificial intelligence; explainable AI; XAI; clinical decision support systems; CDSS; medicine; machine learning; deep learning; explainability; transparency; interpretability

check for updates

Citation: Antoniadi, A.M.; Du, Y.; Guendouz, Y.; Wei, L.; Mazo, C.; Becker, B.A.; Mooney, C. Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Appl. Sci.* **2021**, *11*, 5088. https://doi.org/10.3390/app11115088

Academic Editors: Giorgio De Nunzio and Luana Conte

Received: 28 April 2021
Accepted: 27 May 2021
Published: 31 May 2021

1. Introduction

Artificial Intelligence (AI), generally, and Machine Learning (ML), specifically, have demonstrated remarkable potential in varied application domains, from self-driving cars [1] to beating humans at increasingly complex games such as Go [2]. Almost all processes driven by software can benefit from techniques that can automatically learn from previous data, gaining knowledge from experience, and gradually improving the ability to make predictions based on new data. Recent rapid progress in ML has been driven by many factors including the development of new statistical learning algorithms, the availability of large datasets, and low-cost yet powerful hardware including storage and networking advances which make tools such as cloud storage not only feasible but the norm [3].

Aside from hardware and other advances, the recent growth in ML systems is partly due to the widespread use of increasingly complex models, for example, Deep Neural Networks [4]. However, this complexity comes at a cost. Such systems are often, in effect, black boxes [3] with users or those otherwise affected having little to no understanding of how they make predictions. This lack of understanding presents numerous problems with serious consequences, including, potentially catastrophic errors when flawed models (or decisions based on them) are deployed in real-world contexts [5]. Additionally, even when successful, this opacity can prevent these tools from being accepted by regulated industries, legislation, and society at large [6]. Humans seem to be programmed to seek cause behind action, likely for good reasons. Because of this, humans are reticent to adopt techniques that are not directly interpretable, tractable, and trustworthy [7], especially given the increasing demand for ethical AI [8].

The use of ML systems is expanding not just in software engineering [9] but also into socially delicate application domains such as education [10], law enforcement and forensics [11], and healthcare [12], which further complicates their use especially when their inner workings are simply beyond the understanding of many of those affected by the predictions that these systems make.

The medical domain is home to many critical challenges that stand to be overcome with the use of AI. Many examples have gained traction recently with large volumes of work on automated diagnosis, prognosis, drug design, and testing [13–17]. This is fueled by the importance of medical care and the generation of data in massive quantities from sources such as medical imaging, biosensors, molecular data, and electronic medical records [12]. The aims of AI in medicine include the personalization of medical decisions, health practices, and therapies to individual patients [3]. However, the current state of AI in medicine has been summed up as "high on promise and relatively low on data and proof" [12]. A number of AI-based systems have been validated in real-world settings for diabetic retinopathy, detection of wrist fractures, histologic breast cancer metastases, very small colonic polyps, and congenital cataracts; however, many of the systems that have been shown to be equivalent or superior to experts in experimental settings have demonstrated high false positive rates in real-world clinical environments [12].

Other concerns surrounding the use of ML in medicine include bias, privacy, security, and lack of transparency [12], as well as causality, transferability, informativeness, fairness, and confidence [18]. As decisions made or influenced by such systems ultimately affect human health there is an urgent need for understanding of how such decisions are made [18]. This is amplified in some areas in particular. One example is life-changing outcomes and decisions as a result of disease diagnosis [19]. Another is precision medicine where experts require far more information from the model than a simple binary prediction for supporting their diagnosis [18].

Explanations for how and why a model outputs what it does are crucial in overcoming these challenges [19]. For this reason, explainability and the related concepts of interpretability and transparency have become central issues of concern for ML in medicine over the last few years [20]. It has been argued that despite the existence of much evidence supporting their usefulness, ML-based systems are unlikely to be adopted in routine medical practice beyond a limited number of niche applications unless these challenges are addressed, most likely by having systems provide satisfactory explanations for their decisions [20,21]. Unfortunately, secondary factors also complicate solutions, for instance, different applications usually have different interpretability and explainability needs [20] working against generalizable solutions that span a number of situations.

Clinical Decision Support Systems (CDSS) are computer systems designed to assist in the delivery of healthcare, and ML is being exploited for their development. The explainability of such systems is a relatively new area of study and this work aims to present its application, benefits, gaps, and future opportunities by conducting a systematic literature review. Our hypothesis is that despite the plethora of ML-based CDSS, there is only a limited number of systems that have been specifically developed with explainability as one of their features, and that there are still challenges that need to be addressed. Future systems should be created according to current reported benefits and gaps. As a result, this study aims to first identify the state-of-the-art in explainable ML-based CDSS, in terms of the area of use and current prevalent methodologies, and then discover what benefits have been reported as a result of this combination and what the areas for improvement are.

The remainder of the paper is structured as follows. Section 2 presents a background of CDSS, XAI, and the considerations of applying XAI to CDSS including the need for explainability, its application in medicine, types of explanations, the matter of interpretability vs. performance, and the needs of clinicians in terms of explainability. Section 3 describes our materials, methodology, and research questions. Section 4 presents findings and answers to the research questions. Section 5 discusses these findings, along with guidelines for the future implementation of explainable ML-based CDSS. Section 6 presents our conclusions.

2. Fundamental Concepts and Background

2.1. Clinical Decision Support Systems

Clinical Decision Support Systems are computer systems that "provide clinicians, staff, patients, or other individuals with knowledge and person-specific information, intelligently filtered, or presented at appropriate times, to enhance health and health care" [22]. CDSSs are designed for a variety of purposes such as diagnosis, treatment response prediction, treatment recommendation (personalization), prognosis, and the prioritization of patient care according to their level of risk. They can be helpful in clinical practice as a "second set of eyes" for clinicians, combining their human knowledge with the "knowledge" that is embedded in the system. CDSS can help to improve patients' safety, quality of care, and healthcare efficiency [23–25], as well as reducing the costs of healthcare [26]. They can improve patient safety not only by reducing medical errors but also through reminders for medications or other medical events for patients or clinicians [25]. Additionally, CDSS can be useful in low-resource settings where the number of medical institutions, equipment, and qualified clinicians is limited.

CDSS can be classified as knowledge-based and non-knowledge-based [27]. CDSS that are knowledge-based depend on medical guidelines and knowledge while non-knowledge-based CDSS typically use ML. ML-based CDSS find patterns in historical clinical data and develop predictive models that are able to predict clinical outcomes based on new inputs. These outcomes can then be used as recommendations for clinicians to help them in their practice. ML-based CDSS have great potential in clinical practice. They can help to enhance the accuracy of clinical decisions and minimize medical errors because they are objective, depending only on the input data, and the inner decision-making logic. However, they rely on the quality and quantity of data provided [28]. When the data used to train an ML model are biased, this bias is captured by the model and consequently can make biased or incorrect predictions. This can ultimately lead to a biased or incorrect human decision.

Companies such as IBM, Elsevier, Intermedica, and Microsoft have developed or are currently developing such systems. IBM's "Watson Health" [29] aims to help in treatment-related decisions for patients. However, significant challenges still remain as Watson Health does not perform as well in the clinical world as it did in the game show Jeopardy! [30]. Elsevier's "Via Pathways", rebranded "ClinicalPath" [31] provides evidence-based care maps for the treatment of patients with cancer, and ClinicalKey [32] is a search engine that provides clinical decision support using research-based recommendations. Infermedica has developed a mobile application called Symptomate [33] which is a popular symptom checker, recently updated to perform a COVID-19 checkup. Finally, Microsoft is developing the "Hanover Project" [34] which aims to identify the most relevant pieces of information that experts will need to make the best possible decisions regarding treatment plans for patients with cancer.

CDSS have been developed and applied across a wide range of pathologies including rare diseases [35], oncology [36] and specifically, breast cancer [37], chronic obstructive pulmonary disease [38], the prevention of venous thromboembolism [39], the prediction of chronic kidney disease [40] and Alzheimer's disease [41], for diabetes care [42], and risk-level prediction of heart disease [43].

In constructing CDSS, developers are confronted with "unknown, incomplete, imbalanced, heterogeneous, noisy, dirty, erroneous, inaccurate, and missing datasets in arbitrarily high-dimensional spaces" [3]. Additionally, systems such as CDSS do not work in isolation, but within other systems, institutions, and with human actors whose efforts must be coordinated for AI in medicine to have the most beneficial impact. There is an overall perception that humans are more tolerant towards human error than machine error, and Prahl and Van Swol [44] found that decision-makers considered human advisers to be more expert and useful, while they showed more negative emotions when a human advisor was replaced by a machine. Error tolerance needs to be identified based on current standards and needs, and discussed with the CDSS vendor. Deviations from this rate will lead to mistrust against the system. In addition to tolerance rate, it is also important to analyze the concordance rate between machine learning models and what the physician recommends as best treatment [30]. Watson for Oncology was found to have 83%, 73%, and 49% concordance in three studies mentioned in [30]. As a result, incorporating XAI principles into CDSS is essential if the potential beneficial impact is to be fulfilled.

2.2. Explainable AI (XAI)

The earliest use of the term XAI that we encountered was in Van Lent et al. in 2004 [45]. XAI can simply be described as aiming to make AI systems more understandable to humans; however, there is no accepted technical definition of XAI at this time, and more clarity and consistency is required in terms of the terminology in use [18,19,46]. One of the issues is that the terms transparency, interpretability, and explainability are often used interchangeably. However, there are differences between these concepts.

Interpretability is related to how much a model can be understood [21] although it is also used instead of the term "explainability" [46]. *Transparency* either refers to a holistic characteristic of "providing stakeholders … with relevant information about how the model works: this includes documentation of the training procedure, analysis of training data distribution, code releases, and feature-level explanations" [47], or an algorithm-specific clarity on how the model works, as opposed to opacity [18,46,48]. *Explainability* gives insight into the reasons for the decision-making of the system, but is sometimes connected to understandability which was defined by a consensus as "loosely referring to tools that empower a stakeholder to understand and, when necessary, contest the reasoning of model outcomes"[49]. In this work, we focus on explainability.

2.2.1. The Need for XAI: Fair and Ethical Decision-Making

"Black box" AI systems that give prediction without any explanation are problematic for numerous reasons, not only because of their lack of transparency but also because they hide potential biases within the system [50]. There are many examples where bias in AI-based predictive systems has been uncovered. These systems have been shown to reinforce social and historical human prejudices and people who are traditionally marginalized in our society are disproportional negatively impacted [8].

Predictive software used in courtrooms to assess the likelihood of recidivism have proven to be extremely unreliable due to their bias towards race, revealing higher scores for Black people [51]. For example, the AI-based system COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) has been widely criticized for being unreliable and racially biased [52]. Several case studies have demonstrated that "dirty data" used in policy-making systems have led to skewed predictions [53]. Another example of bias in AI concerns search engines that tend to favor certain sites over others revealing a political bias [54], and hiring algorithms tainted by "societal noise" tend to perpetuate discriminatory behaviors impacting certain individuals or groups [55]. This algorithmic discrimination is also observed in systems such as targeted advertisements where gender discrimination occurred in the display of STEM career ads [56,57]. Moreover, some vision detection systems have demonstrated a bias toward subject skin tone. This predictive inequity has been characterized by higher performance for lower Fitzpatrick skin tones [58]. These cases, to mention but a few, illustrate how such systems deployed in a real-world context can become "Weapons of Math Destruction" reinforcing inequalities [59].

Similarly issues arise when tools built on biased data are used in precision medicines [60] and there are many examples of medical datasets where the lack of inclusion of minorities has led to the development of biased models. For example, European populations were found to be significantly over-represented, while the other races were underrepresented in genomic studies in US [61]. Another example concerns the Framingham Heart Risk functions used to assess the risk of coronary heart diseases which suffered from an overestimation of risk for the German population [62]. This bias was due to the Caucasian sample the initial study was based on. This reveals the particular attention needed in order to implement AI-based models developed on medical datasets.

Slack et al. [63] determined that existing XAI techniques cannot provide explanations that adequately identify discriminatory behavior in some sensitive applications. Although, giving explanations can increase understanding of and trust in a system [19,46,64], simple explanations can hide undesirable attributes of the system and may mislead users into coming to dangerous or unfounded conclusions that could ultimately be unethical [21]. Additionally, an awareness of the dangers of blindly embracing explanations that may disguise racial or gender discrimination [46] or provide fair-washing, i.e., "promoting the false perception that a ML model respects some ethical values" [65] is needed in all medical areas where such systems may be utilized.

In this paper, we focus on XAI as a technician solution that can help to expose systemic bias in CDSS; however, this does not address underlying deep-rooted discriminatory assumptions [8]. XAI can be used to help us evaluate if predictions are biased and defend algorithmic decisions as being fair and ethical [18,19,46,66]. Additionally, XAI can help to shed some light onto causality [18,46,67] although, there is a recognized need to go beyond causality/correlation to true "causability" [3].

Furthermore, there are now regulations in the EU which give subjects the right to obtain an explanation of the decision made using their data which need to be considered in the development of CDSS. To ensure that the processing of data is conducted in a manner that respects the rights of the data subjects and leads to the development of fair systems, different regions have adopted regulations governing the use of data. The General Data Protection Regulation (GDPR) [6] protects the personal data of all EU residents, irrespective of the processing location. GDPR gives EU residents the right to access rectification, erasure, and restriction of processing of their personal data. Specifically, data subjects who will be affected by a decision have a "right to nondiscrimination" and should be able to be informed of the reasons for the automated decision. According to the GDPR, data subjects have "the right not to be subject solely on automated processing" (Article 22). More specifically, according to Article 22(3) and Recital 71 "such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision".

2.3. XAI in Medicine

The use of XAI in medicine is rooted in demand for the added value that comes from medical professionals being able to understand how and why a machine-based decision has been made. Thus, there is growing demand for AI approaches that not only perform well, but are trustworthy, transparent, interpretable, and explainable for a human expert. This also has important implications for the public, policy, and governance as the explainability of AI tools will enhance the trust of medical professionals [3]. In many ways, medical professionals act as translators for patients—translating knowledge that is too complex for patients themselves to understand and act on. Having CDSS which assist medical professionals in this task makes sense, provided the CDSS aid, not hinder, that translation.

Translating ML models effectively to clinical practice requires establishing clinicians' trust in the system. However, there have been a number of high-profile cases that have undermined trust in the use of AI in medicine. For example, many of the recommendations for treatment by "Watson for Oncology" (IBM) have been shown to be incorrect and potentially harmful [30]. In another famous example by Caruana et al. [68], it was found that a ML-based system that was trained to predict which patients with pneumonia should be admitted to hospital identified patients with asthma as being at lower risk of dying from pneumonia. This reflected a true pattern in the training data—patients with asthma were less likely to die from pneumonia—but this was because they tended to be admitted directly to the Intensive Care Units (ICU) and received more aggressive treatments. However, if this model was deployed in a real-world clinical environment without understanding why this prediction was being made, and without human/expert intervention, it is possible that patients with asthma would not be admitted to hospital, and would not receive the aggressive treatment required to prevent death. The use of explainable models could help to prevent such mistakes being made.

However, there is a lack of consensus upon which usable explanations can be used in different settings [69]. Monteath and Sheh [70] proposed a novel XAI approach to incremental decision support for medical diagnosis using decision trees; their approach allows AI systems to work alongside human experts, each informing the other and coming to a decision together. Their system is able to guide physicians in determining which test results are most useful given existing data. The system is also able to explain how a particular decision was made, tracing right back to the underlying training data. This provides the transparency that is crucial for patient confidence, regulation compliance, detecting and correcting errors, and improving patient outcomes. Another example of humans and AI systems working alongside is the work by Wu et al. [71] who proposed an expert-in-the-loop interpretation method to label the behavior of internal units in Convolutional Neural Networks. They demonstrate that several Convolutional Neural Networks models can produce explanatory descriptions to support the final classification decisions. Their findings are an important first step towards XAI in classification of diseased tissue.

Developers of ML-based models in medicine are increasingly focusing on explainability, and their results are promising. Zheng et al. [72] proposed a novel and explainable method to classify cardiac pathology by extracting image-derived features to characterize the shape and motion of the heart. Their proposed model achieves 95% classification accuracy, a performance comparable to that of the state-of-the-art that enables explanations and transparency to become more trustworthy. Tosun et al. [73] described an initial XAI enabled software application, HistoMapr-Breast, for breast core biopsies. HistoMapr-Breast automatically previews breast core whole slide images and recognizes the regions of interest to rapidly present the key diagnostic areas in an interactive and explainable manner. HistoMapr-Breast can work for pathologists in a trustworthy fashion using its explanation interface. They believe that the concept of XAI system must be integrated in pathology workflows promoting safety, reliability, and accountability in addressing issues with bias, transparency, safety, and causality. They also highlight that an XAI system augments pathologists and works with them but does not replace them. Hicks et al. [74] introduced Mimir, an automated multimedia reporting software dissecting the neural network to learn the intermediate analysis steps, which directly adds explainability to Deep Neural Network models in medical problems by producing structured and semantically correct reports, composed of text and images. Mimir enables investigation, explainability, and understanding of the deep learning algorithms decision processes. Ultimately, better explanations will result in patients that understand and trust the reasoning chain, leading to improved confidence, allowing doctors to provide better diagnoses [75].

2.4. Types of Explanations

Techniques can be grouped by scope into those providing global explanations of the entire system and those providing local explanations of single predictions. Global explanations facilitate the understanding of the entire model behavior and reasoning leading to expected outcomes. For local explanations, the reasons for a single prediction are provided to justify why the model made a specific decision for that instance [19]. Techniques can be grouped by whether they are model agnostic (i.e., they can be applied to any ML algorithm), or model specific (i.e., they can be only applied to a specific ML algorithm) [19].

2.4.1. Ante-Hoc Methods

Additionally, techniques can be split into ante-hoc and post hoc explainability methods. Ante-hoc methods are explainable by design or inherently explainable methods, and are also referred to as transparent or white box/glass box approaches. These methods, which are model specific by definition, include linear and logistic regression, decision trees, k-nearest neighbors, fuzzy inference systems, rule-based learners, general additive models, and Bayesian models [3,18]. However, even for these methods, they can only be considered explainable, transparent, or interpretable up to a point, for example, in high-dimensional scenarios with complex interaction terms or deep decision trees, these methods can become difficult to interpret [67].

ML algorithms such as random forests, support vector machines, neural networks (including Deep Neural Networks) are, within practical limits, inherently non-explainable and are typically referred to as "black-box" models [3,19,21]. Post-hoc methods, which are typically model agnostic, might not explain how black-box models work but they may provide local explanations for a specific decision [3,18,46]. One way to do this is to build simpler transparent models that provide interpretable approximations of the black-box [63].

2.4.2. Post-Hoc Methods

Post-hoc methods can be divided into global explanations, for example, the model-agnostic method BETA [76] and neural networks specific method GAM [77] or local explanations including model-agnostic approaches such as LIME [78], SHAP [79] and Anchors [80]. These methods provide feature level explanations by learning an interpretable model that attempts to approximate the behavior of the original model. Global explanations can also be provided by these methods by summarizing local explanations, such as with SHAP summary plots [79] or SP-LIME [78]. CLEAR [81] and CERTIFAI [82] are both model-agnostic methods that generate local explanations supported by the provision of counterfactual explanations that show examples of inputs that are generated to be close to the original input but for which the model provides a different outcome. Other common methods used for the explanations of DL models include gradient-based attribution methods [83], such as integrated gradients [84], or DeepLIFT [85]. Deconvolution [86,87], Class Activation Maps, or CAM [88], and Grad-CAM [89] are techniques to visualize Convolutional Neural Networks. Variations on all these techniques are being developed and apply to different scenarios. Visual explanation techniques are also a means to providing model-agnostic explanations, and a summary of them is presented in the work by [18].

There are two main ways of evaluating post-hoc methods: mathematically quantifiable metrics and human-centered evaluations [67]. However, there is currently no consensus on how to evaluate how interpretable a model is, how correct an explanation is, or how to benchmark methods against each other [18,66,67]. There is some concern around the reliability of post-hoc explanations [63,90]. There are also concerns that post-hoc methods could expose the original models to adversarial attacks [18] or could lead to the generation of classifiers whose post-hoc explanations could be arbitrarily controlled [63]. Adversarial attacks can "trick" the ML algorithm and significantly affect its output with slight changes in the input data. As XAI provides insight into the functionality of the CDSS, it can allow for more effective attacks. Solutions for DL models, SVM models, or even unsupervised ML models have been proposed [18]. Others have deployed explainable techniques such as SHAP to discriminate between normal and adversarial inputs in Deep Neural Networks [91]. Techniques such as a "goodness checklist, explanation satisfaction scales, elicitation methods for mental models, computational measures for explainer fidelity, explanation trustworthiness and model reliability" have been suggested as appropriate methods of evaluation [18].

2.5. Trade-off between Interpretability and Performance

There is often a perceived trade-off between the performance (predictive accuracy) of a model and explainability [3,18,19]. The algorithms that currently often perform the best (e.g., deep learning) are the least explainable, creating a demand for explainable models which can achieve high performance [18]. Simple models are often preferred for their ease of interpretation despite a general trade-off between model performance and explainability that is often assumed [68,69]. However, linear models, for example, are not strictly more interpretable than, for example, neural networks, especially when high-dimensional or heavily engineered features are used. In these cases, the interpretability or the explainability of the model can be lost [46]. Likewise, more complex models may not be more accurate.

One could argue that it would not be ethical to apply in clinical practice a model that does not have the best possible performance, as the ultimate goal is to provide the best possible assistance to patients [92]. Amann et al. [93] provided an example comparing advanced laboratory testing and AI-based CDSS, which are similar in terms of the fact that they support clinical decisions and that accuracy is important. In the case of the first, there is some general understanding on behalf of clinicians but not for each result. Some level of understanding for AI-based CDSS is also possible in terms of "the agent view of AI, i.e., what it takes as input, what it does with the environment, and what it produces as output, and (2) explaining the training of the mapping which produces the output by letting it learn from examples, which encompasses unsupervised, supervised, and reinforcement learning" [93] and might suffice for certain scenarios. The authors also consider the fact that the first requirement of AI systems in medicine is clinical validation, while explainability is a second aspect. Medical certification comes after the system is compliant with regulatory standards and prediction accuracy is usually the main measurement of clinical validation. However, as perfect performance is not possible, while from a patient perspective there is more trust towards clinicians and less tolerance for "machine" error, explainability is required, making this a difficult dilemma for developers. With the availability of larger datasets there are increasing benefits of using more complex models which allow for more complex functions to be approximated [18,46,79] and future developments in XAI may allow for an optimal balance between the explainability and performance of more complex models.

2.6. What Do Clinicians Want?

The needs of clinicians are critically important for the success of XAI in medicine and extend far beyond better, more accurate, cheaper, or faster decisions. Clinicians are the primary users (if not beneficiaries) of XAI-enabled CDSS and their requirements must be met. Different clinicians will have different views, but all clinicians share a common ground—that of explainability through the eyes of patients [69]. Bussone et al. [75] found that clinicians wanted better explanations from the CDSS to help them interpret the system's confidence, to verify that the clinical disorder fit the CDSS suggestion, to better understand the reasoning chain of the system, and to make different diagnoses in order to help them make an assessment of the reliability of the system's decisions. Tonekaboni et al. [69] found that the model's overall accuracy was not sufficient on its own to allow clinicians to make an informed decision, clinicians wanted to know the subset of features driving a prediction to allow them to compare the model decision to their clinical judgment. They explored what makes a model explainable for clinicians through exploratory interviews and found the following:

- Clinicians view explainability as a means of justifying their clinical decision-making (e.g., to patients and colleagues) in the context of a model's decision.
- The implemented system/model needs to provide information about the context within which the model operates and promote awareness of situations where the model may fall short (e.g., model did not use specific history or did not have information around certain aspects of a patient's context). Models that fall short in accuracy were deemed acceptable provided there is clarity around why the model under-performs.
- Familiar metrics such as reliability, specificity, and sensitivity were important for the initial uptake of an AI tool. However, a critical factor for continuing use was whether the tool was repeatedly successful in prognosticating their patient's condition in their personal experience. Real-world application was crucial to developing "a sense of when it's working and when it's limited" which meant "alignment with expectations and clinical presentation".
- Clinical thought processes for acting on predictions of any assistive tool appear to consist of two primary steps following presentation of the model's prediction: (i) understanding and (ii) rationalizing the predictions. Thus, classes of explanations for clinical ML models should be designed with the purpose of facilitating the understanding and rationalization process. Clinicians believe that carefully designed visualization and presentation can facilitate further understanding of the model.
- A well designed explanation should augment or supplement clinical ML systems to (a) recalibrate clinician (stakeholder) trust of model predictions, (b) provide a level of transparency that allows users to validate model outputs with domain knowledge, (c) reliably disseminate model prediction using task specific representations (e.g., confidence scores), and (d) provide parsimonious and actionable steps clinicians can undertake.

3. Materials and Methods

This review aims to explore the literature surrounding the use of XAI in CDSS by identifying publications that are of interest to the ML/AI and medical communities, the contributions of these publications, and the evidence for findings reported. We conducted a systematic literature review by adapting the guidelines proposed by Kitchenham [94]. In this review, we followed a structured process that involved the following:

1. Specifying research questions
2. Conducting searches of specified databases
3. Selecting studies by criterion
4. Filtering studies by evaluating their pertinence
5. Extracting data
6. Synthesizing results.

3.1. Research Questions

The research questions that we aim to address are as follows:

- RQ1: What AI-based CDSS have been developed that incorporate XAI?
- RQ2: What aspects/methods of the use of XAI in CDSS have been the focus of the literature?
- RQ3: What benefits have been reported when addressing different aspects of the use of XAI in CDSS?
- RQ4: What open problems, challenges, and needs of explainable CDSS are expressed in literature?

3.2. Conducting Searches

Selecting search terms for a broad and inclusive review of XAI in CDSS proved challenging. Terms that were too general resulted in an unwieldy set of many irrelevant papers, while terms that were too specific were likely to miss relevant studies. After some trial and error with a range of terms, we performed the following six searches (S1–S6). The search terms were used to search Google Scholar (currently the most comprehensive academic search engine according to recent studies [95,96]) on 24 July 2020 and the number of papers returned by each search are shown in brackets after the search terms.

- S1 "clinical decision support system" XAI (35)
- S2 "clinical decision support system" explainable AI (165)
- S3 "clinical decision support system" explainable ML (181)
- S4 CDSS XAI (41)
- S5 CDSS explainable AI (122)
- S6 CDSS explainable ML (124)

The combined output of the six individual searches returned 261 unique publications. The six searches had a minimum of 35 and a maximum of 181 results each, with a total of 668. A ratio of $n_{unique}/n_{total} = 0.39$ is indicative of a cohesive set of searches that together have a desired degree of internal consistency.

3.3. Paper Selection and Filtering

The next stage was selecting papers that formed the basis of the review. We eliminated papers that were not peer-reviewed conference or journal papers (e.g., theses, dissertations, books, book chapters, pre-prints, or other archived articles and posters) and 10 papers that were not written in English, leaving 132 papers. The search results were then examined by title, abstract, and full-text if deemed necessary to remove papers that were clearly out of scope. For instance, we removed non-medical, legal, or human-factor studies e.g., "Experimental Strategies for Regulating Fintech" and "Human–Agent Interaction for Human Space Exploration". This reduced our set to 121 papers (39 conference and 82 journal).

We then performed a quality check of the remaining papers. We only retained conference papers published by ACM or IEEE, or those listed in the CORE 2020 rankings (http://portal.core.edu.au/conf-ranks, accessed on 24 July 2020), and journal publications that were listed in the JCR 2018 Impact Factors (https://clarivate.com/webofsciencegroup/tag/jcr-2018/, accessed on 24 July 2020). Additionally, npj Digital Medicine and two ACM journals that were not listed in JCR Impact Factors 2018 were also included. After this filtering step, 76 conference and journal publications remained. Three of these were not available on any platform at our disposal, and one additional paper was removed when we discovered that it was a pre-print using an ACM TOIIS template but was not published in the ACM Digital Library.

The remaining 72 papers were divided randomly into three groups and shared between three pairs of authors. Each pair took one group of papers and classified them as either include or exclude based on inclusion criteria. The inclusion criteria was XAI or explainability discussed in relation to CDSS. If CDSS and/or XAI was only mentioned in the introduction or related work section of the paper, the paper was excluded and marked as "related work only". After both authors had independently classified each paper, 16 papers were removed by agreement. At this point, each pair met to reconcile differences. The author pairs disagreed on 23 papers giving a 68% agreement rate on paper inclusion/exclusion. Reconciliation resulted in seven disputed papers being excluded and 16 retained, leaving 33 papers. These papers were published between 2008 and 24 July 2020: one paper was published in 2008, three papers were published in 2018, 13 papers were published in 2019, and 16 papers were published in 2020 (until the 24 July). We see an upward trend in the number of relevant studies published over time which indicated the increased interest in XAI-enabled CDSS.

Finally, we separated the 33 papers into eight literature reviews (24.2%) [97–104], nine papers discussing aspects of the use of XAI in CDSS (27.3%) [48,105–112], four papers that discussed the implementation of a CDSS without XAI (12.1%) [113–116], and 12 papers describing the implementation of a CDSS with (36.4%) XAI [117–128] (see Table 1 for a high-level overview of these papers, as well as the type of data the described CDSS can process to provide suggestions).

Table 1. Twelve publications reporting on CDSS that have implemented XAI.

Paper Subject Area	Main Contribution	Data Processed
Sadeghi et al. [117] *Sleep quality prediction*	Use of time domain features for transparency and explainability	Tabular
Wang et al. [118] *Intensive care phenotyping*	Used SHAP [79] for attribution, LORE [129] for counterfactual rules and multiobjective evolutionary algorithm based on decomposition (MOEA/D) for sensitivity analysis [130]	Tabular
Lee et al. [119] *Alzheimer's Disease (AD)*	Regional abnormalities in the brain space are visualized to create a "regional abnormality map" which is used to interpret regional statuses based on the probability that a region represents later stages of AD progression for a target task, and to draw potential relationships between symptomatic observations	Image
Hu et al. [120] *Critically ill influenza*	SHAP [79] is used to illustrate the individual feature-level impacts on the 30-day mortality	Tabular
Militello et al. [121] *Epicardial fat volume*	A user-centred Graphical User Interface (GUI) design is used to allow for safe interaction of the physician as well as for an effective integration into the existing clinical workflow	Image
Blanco et al. [122] *Cause of death*	A bidirectional Gated Recurrent Units (GRU) with attention mechanism allows for exploration of how much each fragment of the text contributed in the prediction	Text
Lamy et al. [123] *Antibiotic treatment*	The CDSS uses rainbow boxes [131] a visualization technique that displays all the antibiotics present in the ontology in columns and their properties in colored boxes, using labels and icons	Tabular
Tan et al. [124] *Breast cancer*	Implemented a novel method: Complementary Learning Fuzzy Neural Network (CLFNN)	Tabular
El-Sappagh et al. [125] *Diabetes*	Implemented a novel Fuzzy Rule-Based Systems (FRBS) for diagnosis	Tabular
Lamy et al. [126] *Breast cancer*	Implemented a visual case-based reasoning approach for breast cancer management	Tabular
Cai et al. [127] *Prostate cancer*	Algorithmic predictions (benign, grade 3, 4, and 5) were displayed as visual overlays on the image	Image
Kunapuli et al. [128] *Renal mass classification*	XAI based on the Relational Functional Gradient Boosting (RFGB), a statistical relational learning method which provides explanations in terms of tumor shape, size, and texture metrics as well as clinical, demographic, and other factors when they are available	Image

4. Results

4.1. RQ1: What AI-Based CDSS Have Been Developed that Incorporates XAI?

Although AI has achieved notable momentum in medicine since the early 1970s, the use of XAI has only risen notably over the last few years. In AI-based CDSS particularly, XAI did not appear until nearly a decade into the 2000s [124]. However, given the undeniable need for transparency and explainability in medical practice and the growing use of CDSS leveraging AI, XAI has started to be incorporated in recent AI-based CDSS.

Previous works have evaluated XAI in AI-based CDSS, but only as a secondary aspect [99,101,103]. However, there is a number of CDSS in literature that have incorporated XAI (Table 1). Image-based CDSS using XAI are common [119,121,126–128]. Lamy et al. [126] present a CDSS for diagnosing breast cancer using visualization methods for XAI. Additionally, a graphical user interface was presented to medical experts for usability and acceptability validation. Kunapuli et al. [128] proposed a CDSS for renal mass classification. Their XAI is based on tumor shape, size, and texture metrics as well as clinical, demographic, and other factors when they are available. Militello et al. [121]

proposed a CDSS for epicardial fat volume quantification. This CDSS used visualization representations to provide explanations. They developed a user-centered graphical user interface design, allowing them to optimize the interface for safe interaction with the physician (user experience) as well as for effective integration into the existing clinical workflow. Lee et al. [119] proposed a CDSS for magnetic resonance imaging based Alzheimer's disease or mild cognitive impairment diagnosis which presents a "regional abnormality map" to visualize regional abnormalities in the brain space. Cai et al. [127] developed a Deep Neural Network based CDSS for prostate cancer that presents its predictions on the image as visual overlays.

Linguistic reasoners [118,122,124] and ontology-based CDSS [123,125] are the second most common class of CDSS using XAI. Blanco et al. [122] present a CDSS to rank the cause of death from verbal autopsy. This CDSS provides interpretable outputs by evaluating the most important words. Tan et al. [124] proposed a CDSS based on the Wisconsion diagnostic breast cancer dataset. The authors developed a method to improve the CDSS tractability using human-like reasoning, step-by-step inference, clinical differential diagnosis methodology procedure, explanation capacity, and user-familiar terms to gain user acceptance. Wang et al. [118] presented a framework for human-centered, decision-theory-driven XAI building. Visualization methods, data structures, and atomic elements were used to represent explanations in this CDSS. El-Sappagh et al. [125] present a CDSS to diagnose diabetes which mimics the medical expert in both knowledge representation and reasoning process. Lamy et al. [123] developed a CDSS for antibiotic treatment. They used a graphical user interface (GUI) to identify the recommended antibiotic, and also to explain why it is recommended and preferred over alternatives. This CDSS used a set visualization technique called rainbow boxes for XAI.

We also found a CDSS using physiological signals [117] and a feature-based CDSS that incorporated XAI [120]. Sadeghi et al. [117] describe the implementation of a CDSS to predict sleep quality based on physiological signal trends in deep sleep state. Time-domain features were used to make their system transparent and explainable. Hu et al. [120] developed a CDSS for predicting mortality in critically ill influenza patients using feature importance to quantify the importance of each variable based on SHAP.

4.2. RQ2: What Aspects/Methods of the Use of XAI in CDSS Have Been the Focus of the Literature?

We classified the CDSS according to three main categories of XAI: algorithmic transparency, explainer generalizability, and explanation granularity. More specifically, we examined whether they implemented a *post-hoc* or *ante-hoc* explainability method, a model-specific or model-agnostic technique, and whether they provided global or local explanations. The classifications of the CDSS in these categories are presented in Table 2.

Almost all studies aimed for the provision of local explanations, for a specific prediction. The work by Hu et al. [120] was the only one that focused on global explanations, using a model-agnostic *post-hoc* technique. Most studies implemented a model-specific *ante-hoc* technique that provided local explanations [122,124–126,128]. One of the studies used a variety of model-agnostic *post-hoc* methods to provide local explanations in the form of feature attribution, counterfactual rules, and sensitivity analysis [118]. The remaining CDSS implemented model-agnostic *post-hoc* methods for local explanations [119,123].

In terms of *post-hoc* explainability, perturbation-based models that use model-agnostic explanations are common in the literature [98,99,118,120,127] and have been used in two of the proposed CDSS [118,120]. An additional system that incorporated *post-hoc* explainability was designed by Lamy et al. [123] and provided local explanations of the preference model using rainbow boxes [56]. Deep Neural Networks, generally a black-box technique, were explained visually with regional abnormality maps in the system proposed by Lee et al. [119].

The remaining CDSS were created in an *ante-hoc* explainable manner. Rule-based systems were developed in two studies to provide local explanations, as they were considered closer to human reasoning, and thus more preferable by clinicians [124,125]. Case-Based Reasoning is an intrinsically explainable method that was used by Lamy et al. [126] for their CDSS for breast cancer, which was also supported by visual explanations in the form of rainbow boxes and a polar multidimensional scaling scatter plot. In the study by Blanco et al. [122], the CDSS was developed using a bidirectional gated recurrent unit (BiGRU) with attention mechanism, which allowed for the exploration of how much each fragment of the text contributed to a prediction, thus providing local explanations. Kunapuli et al. [128] built a CDSS using Relational Functional Gradient Boosting (RFGB), a statistical relational learning method which attributes its explainability to the usage of tree models and the provision of explanations in terms of features of interest.

Most studies used visualization as a key aspect to enhance explainability, either with SHAP plots [118,120], regional abnormality maps [119], rainbow boxes [123,126], or the attention mechanism that highlighted the important words that lead to a prediction [122].

Militello et al. [121] focused on the use of a user-centered GUI that functions with a semi-automatic strategy, requiring input from the clinicians, and allows for safe interaction. Considering that the focus of this work was on the interface, we did not include this study in Table 2. Sadeghi et al. [117] proposed a CDSS that used a Random Forest to predict the outcome. The authors stated that the use of time-domain features leads to a transparent and explainable CDSS, but there is not sufficient information towards this claim. For this reason, this study is not included in Table 2.

Table 2. CDSS classified by XAI method.

Paper	XAI Method	Model-Agnostic/Specific	*Ante-Hoc*/*Post-Hoc*	Local/Global
Wang et al. [118]	SHAP [79] for attribution, LORE [129] for counterfactual rules, MOEA/D [130] for sensitivity analysis	agnostic	*post-hoc*	local
Lee et al. [119]	Pre-processing to obtain regions, application of randomised Deep Neural Networks on each region and extraction of regional abnormality representations in the form of a map	specific	*post-hoc*	local
Hu et al. [120]	SHAP [79] for summary plot and partial dependence plot	agnostic	*post-hoc*	global
Blanco et al. [122]	BiGRU with attention mechanism to show the contribution of each fragment of text to the prediction	specific	*ante-hoc*	local
Lamy et al. [123]	Visualised the created preference model using rainbow boxes [131]	agnostic	*post-hoc*	local
Tan et al. [124]	CLFNN, which autonomously generates fuzzy rules to provide human-like reasoning	specific	*ante-hoc*	local
El-Sappagh et al. [125]	Semantically interpretable FRBS with the integration of semantic ontology-based reasoning	specific	*ante-hoc*	local
Lamy et al. [126]	Visual (using rainbow-boxes [131] and a polar multidimensional scaling scatter plot) case-based reasoning approach	specific	*ante-hoc*	local
Kunapuli et al. [128]	RFGB, a statistical relational learning method which uses tree models and provides explanations in terms of features of interest	specific	*ante-hoc*	local

4.3. RQ3: What Benefits Have Been Reported When Addressing Different Aspects of the Use of XAI in CDSS?

Several benefits of XAI used in CDSS have been reported. Some researchers presented their XAI-based approaches to doctors or clinicians and collected feedback for usability and acceptability validation. Vorm [106] created vignettes of intelligent systems including a hypothetical CDSS and asked participants (graduate human–computer interaction students) to write down any questions that they would want to ask the system to help them determine whether or not to accept or reject the system recommendation. They reported that XAI could provide different information types to make intelligent systems explainable and more acceptable and trustworthy to users. Liao et al. [111] developed an XAI question bank to bridge the spaces of user needs for AI explainability and technical capabilities provided by XAI work. They interviewed 20 participants to identify gaps between the current XAI algorithmic work and practices to create explainable AI products. The results showed that XAI could gain further insights or evidence, and thus enhance decision confidence or generate the hypothesis about causality. In some cases, users also believed that the interpretation of AI decisions might alleviate their own decision-making biases. XAI can also adapt usage or interaction behaviors to utilize the AI better [111].

Xie et al. [116] developed CheXplain that enables physicians to explore and understand AI-enabled chest X-ray analysis. They asked 39 referring physicians and 38 radiologists to summarize how CheXplain changed their understanding of the underlying AI and how such systems can be integrated into their existing workflow. They showed that XAI provides implications for how physicians can explore and understand data-driven, AI-enabled medical imaging analysis to assist physicians in the medical decision-making process. Cai et al. [127] introduced the critical type of information needs of medical experts to an AI Assistant. They interviewed 21 pathologists to learn about the type of information they desired from the AI assistant. Their findings revealed that users seeking a second opinion compare their information needs to the collaborative mental models they have developed and their compatibility with their diagnostic models. This suggests that AI transparency in collaborative decision making could allow experts to integrate AI assistants into daily practice and gain a richer understanding of the key issues they find.

Lamy et al. [126] presented a visual and interpretable case-based reasoning system. The system displays the dimension names and their associated values to explain why similar cases are similar to the query case and on which dimensions and values the similarities are contained. Such a visual interface can explain the reasoning process to the user, the user can consider their own personal knowledge to enrich the reasoning process, and automatic algorithms can better formalize the visual reasoning process. Lamy et al. [126] reporting that a visual approach could explain why cases are similar via the visualization of shared patient characteristics. This was useful to medical experts, as the physician needs to be aware of the recommendations and confident in their application and use. They presented their interface to 11 medical experts for usability and acceptability validation, demonstrating that XAI could provide the user with a good indication of the confidence level of their choice [126].

Even though some other XAI-based approaches have not yet been tested on users, the benefits of the XAI presented in these works still seem likely to be useful in practice. Kunapuli et al. [128] indicated that XAI could support specific rational reasoning processes, enabling CDSS to support their decisions with understandable interpretations to users with/without ML expertise. Wang et al. [118] identified that XAI could support different explanation types by articulating how people understand events or observations through explanations and can be leveraged to mitigate decision biases and cognitive biases [98,103,112]. Moreover, XAI facilities do support specific rational reasoning processes and can be designed to target decision errors. They could help organize explanations, identify gaps to develop new explanations given an unmet reasoning need, and identify appropriate mitigation strategies to select specific XAI facilities [118]. As discussed in Section 2.2.1, in 2016 the European Union passed the GDPR which has been interpreted as a requirement for any decision made based on an algorithm to be explainable to the user [106]. XAI could help the user understand when to trust a model and why an error may occur [97,116]. Therefore, XAI can support compliance with the GDPR [98,103,106].

Moreover, Hu et al. [120] supported that XAI could provide a description of the cumulative importance of domain-specific features, and a visual explanation of their importance would enable the physicians to understand the critical features in the model intuitively. Therefore, explainability of the support system can improve the acceptability of CDSS by clinicians [48] increase the chances of the complex AI systems' adoption and clinical feasibility of a novel CDSS [105,121]. XAI therefore could greatly enhance the effectiveness of decision-support and clinician confidence [128] especially when high-stakes decisions are being made [127], which is the key factor for the success of the model in the practical use stage [122,124].

4.4. RQ4: What Open Problems, Challenges, and Needs of Explainable CDSS Are Expressed in Literature?

The development of XAI-based CDSS still faces a series of challenges. There is no universal definition of explainability [48,98] nor a well-recognized equivalence or distinction between "interpretable" and "explainable" ML [98]. Furthermore, the concept of interpretability is often highly subjective [115]. Richard et al. [48] proposed a definition stating that a transparent classification system should be understandable, use an interpretable type of classifier and learning system, produce traceable results, and use a revisable classifier. Moreover, Wang et al. [118] believe that what constitutes a good explanation should draw from social science instead of depending on researchers' intuition, and justification is required for choosing different explanation types or representations. Furthermore, Luz et al. [97] argue that a thorough reasoning is required for choosing between transparent and black-box ML algorithms, because post-hoc interpretation methods that develop a mirror model of the original one to add explainability could provide an inaccurate representation of the original model.

In addition, there are some challenges associated with the clinical implementation of XAI-based CDSS. Cai et al. [127] interviewed pathologists and found that beyond local interpretations, clinicians also require insights of models' overall properties, for instance, their capacities, limitations, functionality, medical perspectives, characteristics, and design objectives. This information enriches the explainability of CDSS and is desired prior to the adoption of these systems in routine practice. Liao et al. [111] interviewed user experience and design practitioners of AI products, finding that users recognized the importance of a comprehensive transparency of the training data, in particular: their limitations; explanations of how to best utilize the output; global interpretation with an appropriate level of detail; and local interpretations; understanding of the changes and adaptation of AI and social explanations. However, it was uncovered that users give low rankings to the explainability needs of the performance and counterfactual explanations. The authors agree with the human–computer interaction community that interdisciplinary cooperation and user-centered approaches to explainability are required to close the gaps between XAI and practices. They also discovered that identifying the motivation of explainability helps to select XAI techniques, foresee their limitations, and fill in the gaps occurring while designing user experiences. XAI needs to be interactive and human-like with customized explanations for different users. User experience of XAI design is challenged by the current availability of XAI techniques and other goals. Additionally, guidance for explainability needs specification and creating explainability solutions are desired. Tan et al. [124] argue that a CDSS should have high tractability, which requires human like reasoning, step by step inference, explanation capacity, and user-familiar terms, to gain user acceptance. Moreover, Jin et al. [99] reported that there is a lack of evaluation of XAI techniques on glioma imaging due to lack of focus on practical challenges relating to clinical implementation of XAI. As to testing the systems, Lamy et al. [126] raised the need to confirm their results on a larger user study, in addition to the lack of user studies for some XAI-based approaches as discussed in Section 4.3 (RQ3).

5. Discussion

The massive amounts of data generated and increasing availability of computational resources in healthcare systems make many clinical problems ripe for the development of AI applications. These systems will make diagnosis, treatment, prognostic efforts, follow-up, and decision-making more straightforward, precise, and efficient. This is aided by the fact that physicians worldwide are becoming more receptive towards, and accepting of, AI solutions [12]. However, medical experts struggle with the gap between what is output by an ML-based solution and human explanations. To close the gap requires interdisciplinary work that studies how humans explain, formalizes the patterns in algorithmic forms, and explains outputs in a transparent, easy to interpret manner [67].

A rapid advancement of XAI is evident, and the recent study by Linardatos et al. [132] has identified four main areas of focus: methods for explaining complex black-box models, methods for creating white-box models, methods that promote fairness and restrict the existence of discrimination, and methods for analyzing sensitivity of model predictions. The authors noticed a significant amount of work on explaining complex black-box models, especially on neural networks [132], probably due to the fact that there is great potential in terms of complex analyses and performance. On the other hand, white-box models are described as more challenging to create and, as a result, they seem to have lost their popularity among developers, while despite the progress in methods to promote fairness, the studies that have addressed this issue are also limited.

However, there is an open debate on whether or not XAI in these contexts is necessary and/or worth the substantial overall cost [19]. Nevertheless, XAI may lead to greater uptake and use of CDSS, and may become a requirement in the future due to societal, regulatory, and ethical pressures [18], which could make the difference between success and failure of the system [19]. In some scenarios, explainability of AI output will be a requirement for the output to be used at all, in particular in high-stakes or high-pressure scenarios [3]. Currently, most of the decisions made by AI-based CDSS cannot be interpreted in a transparent way potentially limiting the uptake, trust, and usability of these systems in practice [20,133]. On the other hand, there are some studies supporting that XAI is not always necessary [46,134,135]. London [134] defends the ability to produce results and empirically verify their accuracy as more important than the ability to explain how such results are produced. Baldi [135] explains this argument using examples of the lack of explainability of many processes in our daily lives, for example, how cars, computers, cell phones, or even our brains work. Lipton [46] argues that the short-term goal of building trust with doctors by developing transparent models might clash with the longer-term goal of improving health care. Note that Sullivan [136] states that the opaqueness of models such as deep neural networks is not what is limiting our understanding, but rather the "link uncertainty", meaning the empirical link between the model's features and the phenomenon studied.

Additionally, Bruckert et al. [137] shed light on the difficulties that are presented when rendering ML models explainable for healthcare purposes, highlighting that the implementation of such systems requires overlap between different disciplines and professions. The "right" level of explainability required depends on many factors and is context and resource specific [138]. However, explanations should be at least potentially actionable, parsimonious, and timely [69], warranting further research [19]. Ultimately, it is presumed that explainable CDSS will build trust with clinicians leading to increased adoption of ML-based systems in clinical practice [12,21,69,139].

Guidelines for Implementing Explainable Models in CDSS: Opportunities, Challenges, and Future Research Needs

Developing ML-based CDSS is a multidisciplinary process that should include the needs of all stakeholders. This is especially true when incorporating XAI into these systems. Consideration should be given to the designers of the system, the decision-makers using the systems, and those ultimately impacted by the consequences of those decisions [18,20]. Models should be built in collaboration with input from those with expertise from the fields of social and behavioral science, philosophy, psychology, and cognitive science [19,20,64]. Although XAI can assist with identifying issues with the data, the problem with unstructured medical data remains a challenge for the development of usable AI-based systems. Angehrn et al. [103] discuss the problem and propose solutions that include (a) data exchange between different sources, provided that appropriate safeguards for data privacy are in place; (b) considering the use of data mining techniques to extract crucial clinical information which might have been captured in free text; and (c) a controlled design process that uses AI to develop and collect data during clinical use. Additionally, we might surpass the problem of data availability and heterogeneity by using the least amount of data available and the easiest-to-collect information to initiate the development of the system [140].

Domain-specific needs must be taken into account including a thorough understanding of the purpose of the system, the performance and interpretability of existing systems, and the level and nature of the explanations that are required [18]. Additionally, Arrieta et al. [18] recommend that those black-box models should be selected only when necessary and, when possible, the use of interpretable or transparent by design algorithms should be prioritized over complex algorithms that require the application of post-hoc XAI techniques. Additionally, ethics, fairness, and safety-related implications, as well as the cognitive skills and limitations of the audience must be considered when deciding what type of explanations should be provided [18].

Metrics to evaluate the performance of XAI techniques require further study [18]. According to Arrieta et al. [18], the majority of studies are focused on subjective measurements, for example, user satisfaction, the goodness of an explanation, acceptance, and trust in the system [18]. Subjective measurements can provide valuable insight into the user's experience, however, there is an overall lack of validated and reliable evaluation metrics. A summary of many quantitative metrics for the evaluation of explainability properties (i.e., clarity, broadness, parsimony, completeness, and soundness) for different explanation types, is presented in the work of Zhou et al. [141]. They found that some properties (clarity, broadness, and completeness) are still in shortage of appropriate metrics, and so is the class of explanations that are example-based. The authors have also discussed human-grounded experiments for the evaluation of ML-explanations. They conclude their survey by stating that "the evaluation of ML explanations is a multidisciplinary research topic. It is also not possible to define an implementation of evaluation metrics, which can be applied to all explanation methods." Holzinger et al. [142] introduced the "System Causability Scale" as means to measure explanation quality. This metric is based on "how useful an explanation is".

Finally, there is a need for more robust user studies [66,143]. Bussone et al. [75] found that giving clinicians a fuller explanation of the facts that led to the system's proposed diagnosis had a positive effect on trust but caused over-reliance issues. On the other hand, less detailed explanations had the opposite effect, as this made participants question the system's reliability and caused self-reliance issues [75]. Through a case study, Jacobs et al. [144] found that incorrect ML recommendations may affect clinicians and lower the accuracy of decisions, while explanations were found insufficient for addressing over-reliance on a model that suggests erroneous decisions. They found that explanation strategies ought to be selected according to the clinicians' prior experience with ML and that those with prior experience perceived a higher utility from the ML recommendations. However, there are very few studies like this that give insight into what clinicians want or need. Similarly, there is little discussion on the impact of XAI on patients from the patients perspective. These are areas that will benefit from future research.

6. Conclusions

In order for Clinical Decision Support Systems (CDSS) to be used effectively in practice, they need to be trustworthy, easy to understand, and, most of all, positively augment the human decision-making process. Explainability is a critical component in achieving these goals. Explainability allows developers to identify shortcomings in a system and allows clinicians to be confident in the decisions they make with CDSS assistance. While there are many studies on XAI in medicine, there is a limited number that focus on the context of CDSS. In this review of XAI in CDSS, we focused on the "where" and "how" of XAI use in CDSS, and were able to gauge some realized benefits as well as identify future needs in this area. However, despite some user studies reporting positive views on CDSS, especially in light of explainability, there is still skepticism around their use in practice. A lack of research in general is likely both a symptom and cause of this. A main challenge remains the selection of methods used to present explanations in an informative and efficient—and therefore clinically useful—manner. Significant work lies ahead in order to integrate useful explainablity into CDSS. Studies focusing on all stages of CDSS development are required to establish more firmly how explainability can be put into useful practice in this important context.

Author Contributions: Conceptualization, C.M. (Catherine Mooney) and B.A.B.; methodology, C.M. (Catherine Mooney) and B.A.B.; investigation, A.M.A., Y.D., L.W., C.M. (Claudia Mazo), C.M. (Catherine Mooney), and B.A.B.; data curation, A.M.A., Y.D., L.W., C.M. (Claudia Mazo), C.M. (Catherine Mooney), and B.A.B.; writing—original draft preparation, A.M.A., Y.D., Y.G., L.W., C.M. (Claudia Mazo), C.M. (Catherine Mooney) and B.A.B.; writing—review and editing, A.M.A., Y.D., Y.G., L.W., C.M. (Claudia Mazo), C.M. (Catherine Mooney) and B.A.B.; supervision, C.M. (Catherine Mooney). All authors have read and agreed to the published version of the manuscript.

Funding: This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number 16/RC/3948 and co-funded under the European Regional Development Fund and by FutureNeuro industry partners. This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (D-REAL) under Grant No. 18/CRT/6224. This project has received funding from Enterprise Ireland (EI) and from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 713654.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ML	Machine Learning
CDSS	Clinical Decision Support Systems
XAI	Explainable AI
GDPR	General Data Protection Regulation

References

1. Falcone, P.; Borrelli, F.; Asgari, J.; Tseng, H.E.; Hrovat, D. Predictive active steering control for autonomous vehicle systems. *IEEE Trans. Control Syst. Technol.* **2007**, *15*, 566–580. [CrossRef]
2. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484–489. [CrossRef]
3. Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K.; Müller, H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2019**, *9*, e1312. [CrossRef]
4. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
5. Marcus, G. Deep learning: A critical appraisal. *arXiv* **2018**, arXiv:1801.00631.
6. Goodman, B.; Flaxman, S. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Mag.* **2017**, *38*, 50–57. [CrossRef]
7. Holzinger, A.; Biemann, C.; Pattichis, C.S.; Kell, D.B. What do we need to build explainable AI systems for the medical domain? *arXiv* **2017**, arXiv:1712.09923.
8. Birhane, A. Algorithmic injustice: A relational ethics approach. *Patterns* **2021**, *2*, 100205. [CrossRef]
9. Li, T.; Wang, S.; Lillis, D.; Yang, Z. Combining Machine Learning and Logical Reasoning to Improve Requirements Traceability Recovery. *Appl. Sci.* **2020**, *10*, 7253. [CrossRef]
10. Becker, B.A. Artificial Intelligence in Education: What is it, Where is it Now, Where is it Going? In *Ireland's Yearbook of Education 2017–2018*; Mooney, B., Ed.; Education Matters: Dublin, Ireland, 2017; Volume 1, pp. 42–48. ISBN 978-0-9956987-1-0.
11. Du, X.; Hargreaves, C.; Sheppard, J.; Anda, F.; Sayakkara, A.; Le-Khac, N.A.; Scanlon, M. SoK: Exploring the State of the Art and the Future Potential of Artificial Intelligence in Digital Forensic Investigation. In Proceedings of the 13th International Workshop on Digital Forensics (WSDF) and 15th International Conference on Availability, Reliability and Security (ARES'20), Virtually, 25–28 August 2020; ACM: New York, NY, USA, 2020.
12. Topol, E.J. High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* **2019**, *25*, 44–56. [CrossRef] [PubMed]

13. Hwang, E.J.; Park, S.; Jin, K.N.; Im Kim, J.; Choi, S.Y.; Lee, J.H.; Goo, J.M.; Aum, J.; Yim, J.J.; Cohen, J.G.; et al. Development and validation of a deep learning–based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA Netw. Open* **2019**, *2*, e191095. [CrossRef]
14. Geras, K.J.; Wolfson, S.; Shen, Y.; Wu, N.; Kim, S.; Kim, E.; Heacock, L.; Parikh, U.; Moy, L.; Cho, K. High-resolution breast cancer screening with multi-view deep convolutional neural networks. *arXiv* **2017**, arXiv:1703.07047.
15. Chilamkurthy, S.; Ghosh, R.; Tanamala, S.; Biviji, M.; Campeau, N.G.; Venugopal, V.K.; Mahajan, V.; Rao, P.; Warier, P. Deep learning algorithms for detection of critical findings in head CT scans: A retrospective study. *Lancet* **2018**, *392*, 2388–2396. [CrossRef]
16. Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine learning: Support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5–14. [CrossRef]
17. Kourou, K.; Exarchos, T.P.; Exarchos, K.P.; Karamouzis, M.V.; Fotiadis, D.I. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **2015**, *13*, 8–17. [CrossRef]
18. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
19. Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
20. Vellido, A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Comput. Appl.* **2019**, *32*, 18069–18083. [CrossRef]
21. Gilpin, L.H.; Bau, D.; Yuan, B.Z.; Bajwa, A.; Specter, M.; Kagal, L. Explaining explanations: An overview of interpretability of machine learning. In Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), Turin, Italy, 1–3 October 2018; pp. 80–89.
22. Osheroff, J.A.; Teich, J.M.; Middleton, B.; Steen, E.B.; Wright, A.; Detmer, D.E. A roadmap for national action on clinical decision support. *J. Am. Med. Inform. Assoc.* **2007**, *14*, 141–145. [CrossRef]
23. Coiera, E. Clinical decision support systems. *Guide Health Inform.* **2003**, *2*, 331–345.
24. Shahsavarani, A.M.; Azad Marz Abadi, E.; Hakimi Kalkhoran, M.; Jafari, S.; Qaranli, S. Clinical decision support systems (CDSSs): State of the art review of literature. *Int. J. Med. Rev.* **2015**, *2*, 299–308.
25. Sutton, R.T.; Pincock, D.; Baumgart, D.C.; Sadowski, D.C.; Fedorak, R.N.; Kroeker, K.I. An overview of clinical decision support systems: Benefits, risks, and strategies for success. *NPJ Digit. Med.* **2020**, *3*, 17. [CrossRef]
26. Belard, A.; Buchman, T.; Forsberg, J.; Potter, B.K.; Dente, C.J.; Kirk, A.; Elster, E. Precision diagnosis: A view of the clinical decision support systems (CDSS) landscape through the lens of critical care. *J. Clin. Monit. Comput.* **2017**, *31*, 261–271. [CrossRef]
27. Abbasi, M.; Kashiyarndi, S. *Clinical Decision Support Systems: A Discussion on Different Methodologies Used in Health Care*; Marlaedalen University Sweden: Västerås, Sweden, 2006.
28. Obermeyer, Z.; Emanuel, E.J. Predicting the future—Big data, machine learning, and clinical medicine. *N. Engl. J. Med.* **2016**, *375*, 1216–1219. [CrossRef] [PubMed]
29. IBM Watson Health. Available online: https://www.ibm.com/watson-health (accessed on 25 April 2021).
30. Strickland, E. IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectr.* **2019**, *56*, 24–31. [CrossRef]
31. ClinicalPath. Available online: https://www.elsevier.com/solutions/clinicalpath (accessed on 25 April 2021).
32. ClinicalKey. Available online: https://www.clinicalkey.com (accessed on 25 April 2021).
33. Symptomate. Available online: https://symptomate.com/ (accessed on 25 April 2021).
34. Hanover Project. Available online: https://www.microsoft.com/en-us/research/project/project-hanover/ (accessed on 25 April 2021).
35. Schaaf, J.; Sedlmayr, M.; Schaefer, J.; Storf, H. Diagnosis of Rare Diseases: A scoping review of clinical decision support systems. *Orphanet J. Rare Dis.* **2020**, *15*, 1–14. [CrossRef]
36. Walsh, S.; de Jong, E.E.; van Timmeren, J.E.; Ibrahim, A.; Compter, I.; Peerlings, J.; Sanduleanu, S.; Refaee, T.; Keek, S.; Larue, R.T.; et al. Decision Support Systems in Oncology. *JCO Clin. Cancer Inform.* **2019**, *3*, 1–9. [CrossRef] [PubMed]
37. Mazo, C.; Kearns, C.; Mooney, C.; Gallagher, W.M. Clinical decision support systems in breast cancer: A systematic review. *Cancers* **2020**, *12*, 369. [CrossRef] [PubMed]
38. Velickovski, F.; Ceccaroni, L.; Roca, J.; Burgos, F.; Galdiz, J.B.; Marina, N.; Lluch-Ariet, M. Clinical Decision Support Systems (CDSS) for preventive management of COPD patients. *J. Transl. Med.* **2014**, *12*. [CrossRef] [PubMed]
39. Durieux, P.; Nizard, R.; Ravaud, P.; Mounier, N.; Lepage, E. A Clinical Decision Support System for Prevention of Venous Thromboembolism Effect on Physician Behavior. *JAMA* **2000**, *283*, 2816–2821. [CrossRef]
40. Lakshmanaprabu, S.; Mohanty, S.N.; Sheeba, R.S.; Krishnamoorthy, S.; Uthayakumar, J.; Shankar, K. Online clinical decision support system using optimal deep neural networks. *Appl. Soft Comput.* **2019**, *81*, 105487. [CrossRef]
41. Mattila, J.; Koikkalainen, J.; Virkki, A.; van Gils, M.; Lötjönen, J. Design and Application of a Generic Clinical Decision Support System for Multiscale Data. *IEEE Trans. Biomed. Eng.* **2012**, *59*, 234–240. [CrossRef]
42. Sim, L.L.W.; Ban, K.H.K.; Tan, T.W.; Sethi, S.K.; Loh, T.P. Development of a clinical decision support system for diabetes care: A pilot study. *PLoS ONE* **2017**, *12*, e0173021. [CrossRef]

43. Anooj, P. Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. *J. King Saud Univ. Comput. Inf. Sci.* **2012**, *24*, 27–40. [CrossRef]
44. Prahl, A.; Van Swol, L. Out with the Humans, in with the Machines?: Investigating the Behavioral and Psychological Effects of Replacing Human Advisors with a Machine. *Hum.-Mach. Commun.* **2021**, *2*, 11.
45. Van Lent, M.; Fisher, W.; Mancuso, M. An explainable artificial intelligence system for small-unit tactical behavior. In Proceedings of the National Conference on Artificial Intelligence, San Jose, CA, USA, 25–29 July 2004; pp. 900–907.
46. Lipton, Z.C. The mythos of model interpretability. *Queue* **2018**, *16*, 31–57. [CrossRef]
47. Bhatt, U.; Xiang, A.; Sharma, S.; Weller, A.; Taly, A.; Jia, Y.; Ghosh, J.; Puri, R.; Moura, J.M.; Eckersley, P. Explainable machine learning in deployment. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; pp. 648–657.
48. Richard, A.; Mayag, B.; Talbot, F.; Tsoukias, A.; Meinard, Y. Transparency of Classification Systems for Clinical Decision Support. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems*; Communications in Computer and Information Science (CCIS); Springer: Cham, Switzerland, 2020; Volume 1239, pp. 99–113. [CrossRef]
49. Bhatt, U.; Andrus, M.; Weller, A.; Xiang, A. Machine learning explainability for external stakeholders. *arXiv* **2020**, arXiv:2007.05408.
50. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A survey of methods for explaining black box models. *ACM Comput. Surv. (CSUR)* **2018**, *51*, 1–42. [CrossRef]
51. Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. Machine bias. *ProPublica May* **2016**, *23*, 139–159.
52. Dressel, J.; Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.* **2018**, *4*, eaao5580. [CrossRef]
53. Richardson, R.; Schultz, J.M.; Crawford, K. Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *NYUL Rev. Online* **2019**, *94*, 15.
54. Introna, L.D.; Nissenbaum, H. Shaping the Web: Why the politics of search engines matters. *Inf. Soc.* **2000**, *16*, 169–185.
55. Ifeoma, A. The Auditing Imperative for Automated Hiring (15 March 2019). 34 Harv. J.L. & Tech. (forthcoming 2021). Available online: https://ssrn.com/abstract=3437631 (accessed on 24 July 2020).
56. Lambrecht, A.; Tucker, C. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Manag. Sci.* **2019**, *65*, 2966–2981. [CrossRef]
57. Imana, B.; Korolova, A.; Heidemann, J. Auditing for Discrimination in Algorithms Delivering Job Ads. *arXiv* **2021**, arXiv:2104.04502.
58. Wilson, B.; Hoffman, J.; Morgenstern, J. Predictive inequity in object detection. *arXiv* **2019**, arXiv:1902.11097.
59. O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*; Penguin Random House: New York, NY, USA, 2016.
60. Ferryman, K.; Pitcan, M. Fairness in precision medicine. *Data Soc.* **2018**, *1*. Available online: https://datasociety.net/library/fairness-in-precision-medicine/ (accessed on 24 July 2020).
61. Landry, L.G.; Ali, N.; Williams, D.R.; Rehm, H.L.; Bonham, V.L. Lack of diversity in genomic databases is a barrier to translating precision medicine research into practice. *Health Aff.* **2018**, *37*, 780–785.
62. Hense, H.W.; Schulte, H.; Löwel, H.; Assmann, G.; Keil, U. Framingham risk function overestimates risk of coronary heart disease in men and women from Germany—Results from the MONICA Augsburg and the PROCAM cohorts. *Eur. Heart J.* **2003**, *24*, 937–945. [CrossRef]
63. Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; Lakkaraju, H. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, 7–8 February 2020; pp. 180–186.
64. Miller, T.; Howe, P.; Sonenberg, L. Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *arXiv* **2017**, arXiv:1712.00547.
65. Aïvodji, U.; Arai, H.; Fortineau, O.; Gambs, S.; Hara, S.; Tapp, A. Fairwashing: The risk of rationalization. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 161–170.
66. Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv* **2017**, arXiv:1702.08608.
67. Molnar, C.; Casalicchio, G.; Bischl, B. Interpretable Machine Learning—A Brief History, State-of-the-Art and Challenges. *arXiv* **2020**, arXiv:2010.09337.
68. Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015; pp. 1721–1730.
69. Tonekaboni, S.; Joshi, S.; McCradden, M.D.; Goldenberg, A. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. In Proceedings of the Machine Learning for Healthcare Conference, Boston, MA , USA, 13–14 June 2019; pp. 359–380.
70. Monteath, I.; Sheh, R. Assisted and incremental medical diagnosis using explainable artificial intelligence. In Proceedings of the 2nd Workshop on Explainable Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 104–108.
71. Wu, J.; Peck, D.; Hsieh, S.; Dialani, V.; Lehman, C.D.; Zhou, B.; Syrgkanis, V.; Mackey, L.; Patterson, G. Expert identification of visual primitives used by CNNs during mammogram classification. In *Medical Imaging 2018: Computer-Aided Diagnosis*; International Society for Optics and Photonics: Bellingham, WA, USA, 2018; Volume 10575, p. 105752T.

72. Zheng, Q.; Delingette, H.; Ayache, N. Explainable cardiac pathology classification on cine MRI with motion characterization by semi-supervised learning of apparent flow. *Med. Image Anal.* **2019**, *56*, 80–95. [CrossRef]
73. Tosun, A.B.; Pullara, F.; Becich, M.J.; Taylor, D.; Fine, J.L.; Chennubhotla, S.C. Explainable AI (xAI) for Anatomic Pathology. *Adv. Anat. Pathol.* **2020**, *27*, 241–250. [CrossRef]
74. Hicks, S.A.; Eskeland, S.; Lux, M.; de Lange, T.; Randel, K.R.; Jeppsson, M.; Pogorelov, K.; Halvorsen, P.; Riegler, M. Mimir: An automatic reporting and reasoning system for deep learning based analysis in the medical domain. In Proceedings of the 9th ACM Multimedia Systems Conference, Amsterdam, The Netherlands, 12–15 June 2018; pp. 369–374.
75. Bussone, A.; Stumpf, S.; O'Sullivan, D. The role of explanations on trust and reliance in clinical decision support systems. In Proceedings of the 2015 International Conference on Healthcare Informatics, Dallas, TX, USA, 21–23 October 2015; pp. 160–169.
76. Lakkaraju, H.; Kamar, E.; Caruana, R.; Leskovec, J. Interpretable & explorable approximations of black box models. *arXiv* **2017**, arXiv:1707.01154.
77. Ibrahim, M.; Louie, M.; Modarres, C.; Paisley, J. Global explanations of neural networks: Mapping the landscape of predictions. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA, 27–28 January 2019; pp. 279–287.
78. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
79. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774.
80. Ribeiro, M.T.; Singh, S.; Guestrin, C. Anchors: High-Precision Model-Agnostic Explanations. In Proceedings of the AAAI, New Orleans, LA, USA, 2–7 February 2018; Volume 18, pp. 1527–1535.
81. White, A.; Garcez, A.D. Measurable counterfactual local explanations for any classifier. *arXiv* **2019**, arXiv:1908.03020.
82. Sharma, S.; Henderson, J.; Ghosh, J. CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-box Models. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, 7–8 February 2020; pp. 166–172.
83. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv* **2013**, arXiv:1312.6034.
84. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic attribution for deep networks. In Proceedings of the International Conference on Machine Learning, PMLR, Sydney, Australia, 6–11 August 2017; pp. 3319–3328.
85. Shrikumar, A.; Greenside, P.; Kundaje, A. Learning important features through propagating activation differences. In Proceedings of the International Conference on Machine Learning, PMLR, Sydney, Australia, 6–11 August 2017; pp. 3145–3153.
86. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 818–833.
87. Zeiler, M.D.; Taylor, G.W.; Fergus, R. Adaptive deconvolutional networks for mid and high level feature learning. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2018–2025.
88. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
89. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
90. Garreau, D.; von Luxburg, U. Explaining the explainer: A first theoretical analysis of LIME. *arXiv* **2020**, arXiv:2001.03447.
91. Fidel, G.; Bitton, R.; Shabtai, A. When explainability meets adversarial learning: Detecting adversarial examples using SHAP signatures. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8.
92. Holzinger, A. Explainable AI and Multi-Modal Causability in Medicine. *i-com* **2020**, *19*, 171–179. [CrossRef]
93. Amann, J.; Blasimme, A.; Vayena, E.; Frey, D.; Madai, V.I. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 310. [CrossRef]
94. Kitchenham, B.A.; Charters, S. Guidelines for Performing Systematic Literature Reviews in Software Engineering; Technical Report EBSE 2007-001, Keele University and Durham University Joint Report. 2007. Available online: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=2BE22FED09591B99D6A7ACF8FE2258D5? (accessed on 24 July 2020).
95. Martín-Martín, A.; Orduna-Malea, E.; Thelwall, M.; López-Cózar, E.D. Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *J. Inf.* **2018**, *12*, 1160–1177. [CrossRef]
96. Gusenbauer, M. Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics* **2019**, *118*, 177–214. [CrossRef]
97. Luz, C.; Vollmer, M.; Decruyenaere, J.; Nijsten, M.; Glasner, C.; Sinha, B. Machine learning in infection management using routine electronic health records: Tools, techniques, and reporting of future technologies. *Clin. Microbiol. Infect.* **2020**, *26*, 1291–1299. [CrossRef]
98. Zucco, C.; Liang, H.; Fatta, G.D.; Cannataro, M. Explainable Sentiment Analysis with Applications in Medicine. In Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018, Madrid, Spain, 3–6 December 2019; pp. 1740–1747. [CrossRef]

99. Jin, W.; Fatehi, M.; Abhishek, K.; Mallya, M.; Toyota, B.; Hamarneh, G. Artificial intelligence in glioma imaging: Challenges and advances. *J. Neural Eng.* **2020**, *17*, 021002. [CrossRef]
100. Wulff, A.; Montag, S.; Marschollek, M.; Jack, T. Clinical Decision-Support Systems for Detection of Systemic Inflammatory Response Syndrome, Sepsis, and Septic Shock in Critically Ill Patients: A Systematic Review. *Methods Inf. Med.* **2019**, *58*, e43–e57. [CrossRef] [PubMed]
101. Rundo, L.; Pirrone, R.; Vitabile, S.; Sala, E.; Gambino, O. Recent advances of HCI in decision-making tasks for optimized clinical workflows and precision medicine. *J. Biomed. Inform.* **2020**, *108*, 103479. [CrossRef]
102. Fu, L.H.; Schwartz, J.; Moy, A.; Knaplund, C.; Kang, M.J.; Schnock, K.O.; Garcia, J.P.; Jia, H.; Dykes, P.C.; Cato, K.; et al. Development and validation of early warning score system: A systematic literature review. *J. Biomed. Inform.* **2020**, *105*, 103410. [CrossRef] [PubMed]
103. Angehrn, Z.; Haldna, L.; Zandvliet, A.S.; Gil Berglund, E.; Zeeuw, J.; Amzal, B.; Cheung, S.Y.A.; Polasek, T.M.; Pfister, M.; Kerbusch, T.; et al. Artificial Intelligence and Machine Learning Applied at the Point of Care. *Front. Pharmacol.* **2020**, *11*, 759. [CrossRef] [PubMed]
104. Ibrahim, A.; Primakov, S.; Beuque, M.; Woodruff, H.; Halilaj, I.; Wu, G.; Refaee, T.; Granzier, R.; Widaatalla, Y.; Hustinx, R.; et al. Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework. *Methods* **2020**, *188*, 20–29. [CrossRef]
105. Mahadevaiah, G.; RV, P.; Bermejo, I.; Jaffray, D.; Dekker, A.; Wee, L. Artificial intelligence-based clinical decision support in modern medical physics: Selection, acceptance, commissioning, and quality assurance. *Med. Phys.* **2020**, *47*, e228–e235. [CrossRef]
106. Vorm, E.S. Assessing Demand for Transparency in Intelligent Systems Using Machine Learning. In Proceedings of the 2018 Innovations in Intelligent Systems and Applications (INISTA), Thessaloniki, Greece, 3–5 July 2018; pp. 1–7. [CrossRef]
107. Jamieson, T.; Goldfarb, A. Clinical considerations when applying machine learning to decision-support tasks versus automation. *BMJ Qual. Saf.* **2019**, *28*, 778–781. [CrossRef]
108. Choudhury, A.; Asan, O.; Mansouri, M. Role of Artificial Intelligence, Clinicians & Policymakers in Clinical Decision Making: A Systems Viewpoint. In Proceedings of the 2019 International Symposium on Systems Engineering (ISSE), Edinburgh, UK, 1–3 October 2019; pp. 1–8. [CrossRef]
109. Cánovas-Segura, B.; Morales, A.; Martínez-Carrasco, A.L.; Campos, M.; Juarez, J.M.; Rodríguez, L.L.; Palacios, F. Exploring Antimicrobial Resistance Prediction Using Post-hoc Interpretable Methods. In *Artificial Intelligence in Medicine: Knowledge Representation and Transparent and Explainable Systems*; Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer Nature: Cham, Switzerland, 2019; Volume 11979 LNAI, pp. 93–107. [CrossRef]
110. Zihni, E.; Madai, V.I.; Livne, M.; Galinovic, I.; Khalil, A.A.; Fiebach, J.B.; Frey, D. Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome. *PLoS ONE* **2020**, *15*, e0231166. [CrossRef]
111. Liao, Q.V.; Gruen, D.; Miller, S. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; ACM: New York, NY, USA, 2020; pp. 1–15. [CrossRef]
112. Johnson, S.L.J. AI, Machine Learning, and Ethics in Health Care. *J. Leg. Med.* **2019**, *39*, 427–441. [CrossRef]
113. Timotijevic, L.; Hodgkins, C.E.; Banks, A.; Rusconi, P.; Egan, B.; Peacock, M.; Seiss, E.; Touray, M.M.L.; Gage, H.; Pellicano, C.; et al. Designing a mHealth clinical decision support system for Parkinson's disease: A theoretically grounded user needs approach. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 34. [CrossRef] [PubMed]
114. Ben Souissi, S.; Abed, M.; El Hiki, L.; Fortemps, P.; Pirlot, M. PARS, a system combining semantic technologies with multiple criteria decision aiding for supporting antibiotic prescriptions. *J. Biomed. Inform.* **2019**, *99*, 103304. [CrossRef] [PubMed]
115. Gangavarapu, T.; S Krishnan, G.; Kamath S, S.; Jeganathan, J. FarSight: Long-Term Disease Prediction Using Unstructured Clinical Nursing Notes. *IEEE Trans. Emerg. Top. Comput.* **2020**, 1–16. [CrossRef]
116. Xie, Y.; Chen, M.; Kao, D.; Gao, G.; Chen, X.A. CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; ACM: New York, NY, USA, 2020; pp. 1–13. [CrossRef]
117. Sadeghi, R.; Banerjee, T.; Hughes, J.C.; Lawhorne, L.W. Sleep quality prediction in caregivers using physiological signals. *Comput. Biol. Med.* **2019**, *110*, 276–288. [CrossRef]
118. Wang, D.; Yang, Q.; Abdul, A.; Lim, B.Y. Designing theory-driven user-centric explainable AI. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Scotland, UK, 4–9 May 2019; pp. 1–15.
119. Lee, E.; Choi, J.S.; Kim, M.; Suk, H.I. Toward an interpretable Alzheimer's disease diagnostic model with regional abnormality representation via deep learning. *NeuroImage* **2019**, *202*, 116113. [CrossRef] [PubMed]
120. Hu, C.A.; Chen, C.M.; Fang, Y.C.; Liang, S.J.; Wang, H.C.; Fang, W.F.; Sheu, C.C.; Perng, W.C.; Yang, K.Y.; Kao, K.C.; et al. Using a machine learning approach to predict mortality in critically ill influenza patients: A cross-sectional retrospective multicentre study in Taiwan. *BMJ Open* **2020**, *10*, e033898. [CrossRef]
121. Militello, C.; Rundo, L.; Toia, P.; Conti, V.; Russo, G.; Filorizzo, C.; Maffei, E.; Cademartiri, F.; La Grutta, L.; Midiri, M.; et al. A semi-automatic approach for epicardial adipose tissue segmentation and quantification on cardiac CT scans. *Comput. Biol. Med.* **2019**, *114*, 103424. [CrossRef]

122. Blanco, A.; Perez, A.; Casillas, A.; Cobos, D. Extracting Cause of Death from Verbal Autopsy with Deep Learning interpretable methods. *IEEE J. Biomed. Health Inform.* **2020**, *25*, 1315–1325. [CrossRef]
123. Lamy, J.B.; Sedki, K.; Tsopra, R. Explainable decision support through the learning and visualization of preferences from a formal ontology of antibiotic treatments. *J. Biomed. Inform.* **2020**, *104*, 103407. [CrossRef] [PubMed]
124. Tan, T.Z.; Ng, G.S.; Quek, C. Improving tractability of Clinical Decision Support system. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–6 June 2008; pp. 1997–2002. [CrossRef]
125. El-Sappagh, S.; Alonso, J.M.; Ali, F.; Ali, A.; Jang, J.H.; Kwak, K.S. An Ontology-Based Interpretable Fuzzy Decision Support System for Diabetes Diagnosis. *IEEE Access* **2018**, *6*, 37371–37394. [CrossRef]
126. Lamy, J.B.; Sekar, B.; Guezennec, G.; Bouaud, J.; Séroussi, B. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artif. Intell. Med.* **2019**, *94*, 42–53. [CrossRef] [PubMed]
127. Cai, C.J.; Winter, S.; Steiner, D.; Wilcox, L.; Terry, M. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. In *Proceedings of the ACM on Human-Computer Interaction*; ACM: New York, NY, USA, 2019; Volume 3, pp. 1–24. [CrossRef]
128. Kunapuli, G.; Varghese, B.A.; Ganapathy, P.; Desai, B.; Cen, S.; Aron, M.; Gill, I.; Duddalwar, V. A Decision-Support Tool for Renal Mass Classification. *J. Digit. Imaging* **2018**, *31*, 929–939. [CrossRef] [PubMed]
129. Guidotti, R.; Monreale, A.; Ruggieri, S.; Pedreschi, D.; Turini, F.; Giannotti, F. Local rule-based explanations of black box decision systems. *arXiv* **2018**, arXiv:1805.10820.
130. Zhang, Q.; Li, H. MOEA/D: A multiobjective evolutionary algorithm based on decomposition. *IEEE Trans. Evol. Comput.* **2007**, *11*, 712–731. [CrossRef]
131. Lamy, J.B.; Berthelot, H.; Capron, C.; Favre, M. Rainbow boxes: A new technique for overlapping set visualization and two applications in the biomedical domain. *J. Vis. Lang. Comput.* **2017**, *43*, 71–82. [CrossRef]
132. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **2021**, *23*, 18. [CrossRef]
133. Gomolin, A.; Netchiporouk, E.; Gniadecki, R.; Litvinov, I.V. Artificial intelligence applications in dermatology: Where do we stand? *Front. Med.* **2020**, *7*, 100. [CrossRef]
134. London, A.J. Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Cent. Rep.* **2019**, *49*, 15–21. [CrossRef]
135. Baldi, P. Deep learning in biomedical data science. *Annu. Rev. Biomed. Data Sci.* **2018**, *1*, 181–205. [CrossRef]
136. Sullivan, E. Understanding from machine learning models. *Br. J. Philos. Sci.* **2020**. [CrossRef]
137. Bruckert, S.; Finzel, B.; Schmid, U. The Next Generation of Medical Decision Support: A Roadmap Toward Transparent Expert Companions. *Front. Artif. Intell.* **2020**, *3*, 75. [CrossRef]
138. Carvalho, D.V.; Pereira, E.M.; Cardoso, J.S. Machine learning interpretability: A survey on methods and metrics. *Electronics* **2019**, *8*, 832. [CrossRef]
139. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2019**, *267*, 1–38. [CrossRef]
140. Antoniadi, A.M.; Galvin, M.; Heverin, M.; Hardiman, O.; Mooney, C. Development of an explainable clinical decision support system for the prediction of patient quality of life in amyotrophic lateral sclerosis. In Proceedings of the 36th Annual ACM Symposium on Applied Computing, Gwangju, Korea, 22–26 March 2021; pp. 594–602.
141. Zhou, J.; Gandomi, A.H.; Chen, F.; Holzinger, A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics* **2021**, *10*, 593. [CrossRef]
142. Holzinger, A.; Carrington, A.; Müller, H. Measuring the quality of explanations: The system causability scale (SCS). *KI-Künstliche Intell.* **2020**, *34*, 193–198. [CrossRef]
143. Kenny, E.M.; Ford, C.; Quinn, M.; Keane, M.T. Explaining Black-Box classifiers using Post-Hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artif. Intell.* **2021**, *294*, 103459. [CrossRef]
144. Jacobs, M.; Pradier, M.F.; McCoy, T.H.; Perlis, R.H.; Doshi-Velez, F.; Gajos, K.Z. How machine-learning recommendations influence clinician treatment selections: The example of the antidepressant selection. *Transl. Psychiatry* **2021**, *11*, 108. [CrossRef] [PubMed]