biomedinformatics

MDPI

Article

Explainable Artificial Intelligence (XAI) in Biomedicine: Making AI Decisions Trustworthy for Physicians and Patients

Jörn Lötsch [1,2,*], Dario Kringel [1] and Alfred Ultsch [3]

[1] Institute of Clinical Pharmacology, Goethe-University, Theodor Stern Kai 7, 60590 Frankfurt am Main, Germany; kringel@med.uni-frankfurt.de
[2] Fraunhofer Institute for Translational Medicine and Pharmacology ITMP, Theodor-Stern-Kai 7, 60596 Frankfurt am Main, Germany
[3] DataBionics Research Group, University of Marburg, Hans-Meerwein-Straße, 35032 Marburg, Germany; ultsch@Mathematik.Uni-Marburg.de
* Correspondence: j.loetsch@em.uni-frankfurt.de

Abstract: The use of artificial intelligence (AI) systems in biomedical and clinical settings can disrupt the traditional doctor–patient relationship, which is based on trust and transparency in medical advice and therapeutic decisions. When the diagnosis or selection of a therapy is no longer made solely by the physician, but to a significant extent by a machine using algorithms, decisions become nontransparent. Skill learning is the most common application of machine learning algorithms in clinical decision making. These are a class of very general algorithms (artificial neural networks, classifiers, etc.), which are tuned based on examples to optimize the classification of new, unseen cases. It is pointless to ask for an explanation for a decision. A detailed understanding of the mathematical details of an AI algorithm may be possible for experts in statistics or computer science. However, when it comes to the fate of human beings, this "developer's explanation" is not sufficient. The concept of explainable AI (XAI) as a solution to this problem is attracting increasing scientific and regulatory interest. This review focuses on the requirement that XAIs must be able to explain in detail the decisions made by the AI to the experts in the field.

Keywords: data science; artificial intelligence; machine learning; patient–doctor relationship; digital medicine

Citation: Lötsch, J.; Kringel, D.; Ultsch, A. Explainable Artificial Intelligence (XAI) in Biomedicine. Making AI Decisions Trustworthy for Physicians and Patients. *Biomedinformatics* **2022**, *2*, 1–17. https://doi.org/10.3390/biomedinformatics2010001

Academic Editor: David Barlow

Received: 26 November 2021
Accepted: 16 December 2021
Published: 22 December 2021

1. Introduction

The terms artificial intelligence and machine learning are sometimes used interchangeably, although this is incorrect. In fact, artificial intelligence is a branch of computer science that deals with the automation of human activities that are normally considered intelligent human behavior [1]. These activities include understanding human language, representing and using knowledge, reasoning, planning, problem solving, and risk assessment, including guessing and learning from experience. Machine learning is currently by far the most popular method used in artificial intelligence and can be referred to in two different forms [2]: first, approaches in which a class of very general algorithms (artificial neural networks, classifiers, predictors, associative memories, etc.) are tuned based on examples to optimize the prediction or classification of new, unseen cases. This is skill learning. Second, methods that recognize structures, such as subgroups, in the data and describe these structures in such a way that such a description (knowledge) can be used to correctly classify new cases and can be understood by humans. This is the deduction of knowledge from data [3,4].

The growing importance of artificial intelligence or machine learning algorithms in biomedical research is reflected in their increasing influence on clinical decision-making processes. This, in turn, has a direct impact on medical practice and communication between physicians and patients. The traditional doctor–patient relationship is based

on personal trust and transparency of medical advice and therapeutic decisions. If the diagnosis or the selection of the most promising therapy is no longer made solely by the physician, but to a considerable extent by a machine with learning algorithms and artificial intelligence, the decisions become nontransparent. Physicians cannot be assumed to have the computer science skills necessary to understand the decision-making process of an algorithm and should be able to communicate this process in all relevant details to their patients in an understandable way.

The European Union (EU) has recognized the problem that algorithm-based medical decision making poses to the information rights of affected patients and has published a landmark paper highlighting the need for explanations of computerized decisions so that they can be communicated to affected patients in an understandable manner [5]. The solution is found in the concept of explainable AI (XAI), which is attracting increasing scientific interest [6]. This is consistent with the U.S. military's efforts to obtain explainable models that make decisions made by autonomous systems transparent (https://www.darpa.mil/program/explainable-artificial-intelligence [7], accessed on 15 December 2021). Without a deep understanding, machine learning relies on trial and error and has been compared to medieval alchemists [8]. Along the same line of reasoning, this review focuses on the requirement for XAI to be able to explain in detail the decisions made by an AI in a biomedical setting to the expert in the domain, e.g., the physician in the case of AI-based clinical decisions related to diagnosis, treatment, or prognosis of a disease.

2. An Example Case of XAI versus Standard AI

A common clinical situation is the communication of a diagnosis by the physician to the patient. The diagnosis is made on the basis of sound decision criteria, such as the presence of pathognomonic signs or the excess of a laboratory value over the generally accepted limit for healthy individuals. For example, the diagnosis of lymphoma can be made by observing specific cell types in a patient's blood sample. With increasing automation, the assessment of the many cell subpopulations in a sample is increasingly performed by algorithms, including the analysis of microscopic images for cell type separation and counting [9].

A typical but small data set, freely available with the R library "opdisDownsampling" (https://cran.r-project.org/package=opdisDownsampling [10], accessed on 15 December 2021), consists of d = 6 cytological markers measured by fluorescence-activated cell sorting (FACS) in a total of n = 111,686 cells obtained from 100 patients with chronic lymphocytic leukemia and 100 healthy controls, using the seed value of seed = 42, which is reported here for reproducibility of the results with the referenced R libraries and example data included there. After class-proportional downsampling to 3000 instances to speed up subsequent computations, the data space consisting of data space $D = \{(x_i, y_i) \mid x_{i,d} \in \mathbb{R}^X, y_i \in Y\{1,2\},\ i = 1 \ldots n\}$ with input space X consisting of vectors $x_i = <x_{i,1}, \ldots x_{i,d}>$ with d = 6 different cytological markers and the output classes y_i consisting of the diagnoses of healthy versus diseased, two different algorithms were trained to map the cell marker pattern x_i to the diagnosis classes y_i, i.e., to automatically perform the diagnosis of leukemia. Since the present analyses were intended for the demonstration of an introductory example and not to discuss techniques of classifier training and tuning, the interactive R data mining tool rattle (https://cran.r-project.org/package=rattle [11,12], accessed on 15 December 2021) was used with the default parameter setting. The reasons for the choice of implementation details were explained in the cited publications and are not challenged here.

Thus, the downsampled data set was split into training/test/validation subsets sized 70%/15%/15% of the total data set as advised in "rattle". Using the seed of 42, a standard classification and regression tree (CART) [13] and a support vector machine (SVM) [14] were trained by calling the respective R libraries "rpart" (https://cran.r-project.org/package=rpart [15], accessed on 15 December 2021) and "kernlab" (https://cran.r-project.org/package=kernlab [16], accessed on 15 December 2021). The algorithms were trained

and tuned on the training and test data sets, and the accuracy of class assignment was assessed in the validation subsample by calculating accuracy and the area under the receiver operator characteristic (AUC-ROC). The default settings of "rattle" do not include cross-validation or repeated calculations; this was considered sufficient for the present exemplary demonstration purpose and therefore was not changed or refined. For the same reason, no grid search for hyperparameter tuning and similar standard classifier tuning procedures were performed. The two algorithms provided a nearly similar accuracy of assigning a cell sample to "healthy" or "diseased" of 0.7711 for CART ("rpart") and 0.7689 for the SVM ("ksvm"). The AUC-ROC values were also nearly identical (Figure 1A). However, the transparency of the class assignment decision was completely different for the two algorithms, as explained below.

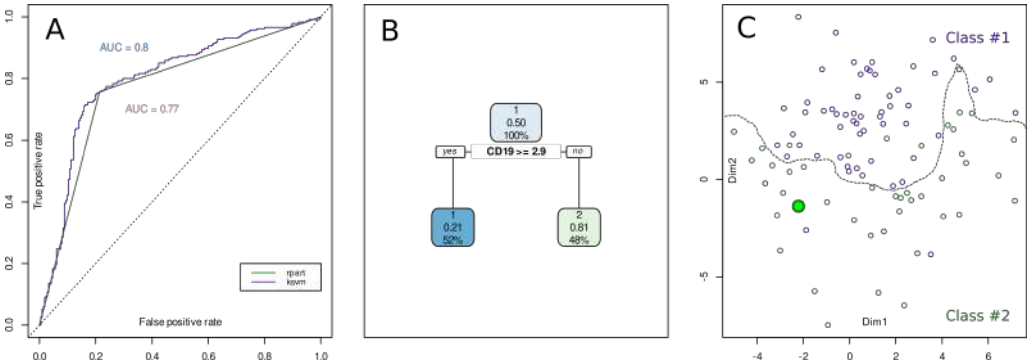

Figure 1. Classification performance of two different types of classifiers, comprising hierarchical decision rules implanted as classification and regression trees ("rpart") and hyperplanes as used in support vector machines "ksvm". Panel (**A**): Receiver operator characteristic of the two classifiers for the classification of cell samples as obtained from healthy subjects (class #1) or subjects with leukemia (class #2). The figure corresponds to the original output of the "rattle" R package with the curve for rpart (= CART) composed of only 3 points since a single decision rule was used in just one iteration for the present demonstration purpose. (**B**): Decision rule by which the hierarchical classifier made the assignment to class #1 or #2. (**C**): Schematic drawing of an SVM decision hyperplane between healthy and diseased samples. For illustrative purposes, the number of data points is reduced to n = 200, and the figure is purely schematic, without performing real calculations and SVM training. In contrast to panels A and B, which show results of computations, this is a schematic drawing. The plots were created using the R software package (version 4.1.2 for Linux; https://CRAN.R-project.org/ [17]) and the R library "rattle" (https://cran.r-project.org/package=rattle [11,12], and the vector drawing software "Inkscape" (https://inkscape.org/de/ [18], all accessed on 15 December 2021).

CART delivers a single simple rule as explanation for the decision, namely that the sample belongs to the disease if the expression of the CD19 marker has a value of 2.9 or more; otherwise, it is from a healthy subject (Figure 1B). This can be communicated to the physician, who understands the role of CD19. Such a biomedical expert will know that the expression level of CD19 on cell surfaces plays an important role for the functioning of B cells [19]. With this information, transparency is transferred from the field of informatics back to medicine, where the physician has to explain the meaning of CD19 to the patient, while the decision-making process of the algorithm is made transparent to both the physician and the patient. Thus, transparency of XAI does not necessarily mean transparency for the patient, but emphasizes the compressibility of a biomedical decision based on machine learning first for the (biomedical) field expert, who then takes over the establishment of comprehensibility for the layperson.

SVM are machine learned classifiers, which use kernel functions to assign data to given classes [14]. Kernel functions represent distances from a hyperplane in a space, where the original data are mapped to. This decision space is typically much higher in dimension (up to infinity) than the feature space of the data. This may have the result that the classification is easier in that space. However, a representation of the decision surface in the data's

space is typically torn and may contain holes and bumps, i.e., senseless [20]. The SVM explanation of why a patient receives a diagnosis of leukemia based on his/her blood sample would be that because the patient's blood sample (thick green dot in (Figure 1C) contains cell marker patterns that place it on the "sick" side of the decision surface (black line in Figure 1C), which separates healthy and sick cells in a unintelligible projection of marker expressions (Dim1/Dim2 in Figure 1C) .

3. Historical Origins of the Need for Explainable AI

3.1. Knowledge Representation in Expert Systems

Explainable AI is not a new field [21]. AI systems were extensively researched in 1980/1990. These systems were based on a precise and formal representation of human knowledge using predicate logics, graphs, e.g., directed acyclic graphs (DAG), and a type of approximate reasoning, such as Bayes [22], fuzzy reasoning, and Dempster–Shafer theory [23]. For example, one of the world's highly successful systems of these knowledge-based systems is the GeneOntology knowledge base [24].

However, a serious bottleneck of these systems is that the AI needs a knowledge representation created by hand before it can start working. Algorithms that can learn to appear to act intelligently seemed to be a solution to this problem. However, most of the machine learning models in use today are neither knowledge based nor knowledge producing. From the perspective of knowledge-based AI, these systems sacrifice understandability and explainability in favor of performance. A better name for most machine learning systems and many "AI" systems would, therefore, be artificial skills-based systems (AS), which is elaborated in the next chapter.

3.2. Knowledge-Based Systems

A transparent decision based on AI could ideally be achieved when a sound scientific theory is available as a basis for how the underlying ML system works. Then, the trustworthy system can draw logical inferences based on this theory to reach its conclusions. This is like knowing Kepler's laws for predicting the positions of planets in astronomy. In systems based on sound scientific theory, the scope and accuracy of a prediction can be estimated. In addition, a rationale (explanation) for the result can be given [25]. In astronomy, for example, Kepler's three laws can be derived from Newton's law of gravity. From these laws, it can be deduced why a particular planet is in a particular position.

3.3. Skill-Based Systems

Machine learning systems are often used for tasks where a scientific theory is not given or even known. For example, machine learning systems are developed to diagnose patients based on various measurements of gene expression, even when the exact molecular processes involved in the disease are only partially known or understood. In such situations, the machine learning literature is content to measure the "quality" of a diagnostic system by measuring the accuracy of its predictions on a limited data set that was not used during the development (training, learning, adaptation, and tuning) of the system, i.e., the so-called "test data" [26,27]. That is, the algorithm is trained on a carefully selected training and test data set to develop the ability to perform a specific task, such as making a clinical diagnosis. In this way, the ability to generalize to new, unknown cases is evaluated. In this approach, confidence is determined by a measure of performance on unseen data. However, in most cases, these "unseen" data were already available when the model was developed.

The limitations of skill-based ML systems are obvious: for data that are very similar in structure to the training data, the system will perform well. For data that have a different structure, the skill-based ML system will fail, and may not even recognize that such data do not fall within the algorithm's skill domain. In astronomy, this is like the epicycle model of planetary motion. It can be thought of as an empirical Fourier analysis of planetary motion, with a series of larger and smaller circles superimposed [28]. For small periods of time and

under "normal" circumstances, the epicycle model can predict the position of a planet to some accuracy [29]. However, it is not known when the prediction is correct or incorrect.

For skill-based ML systems, it is pointless to ask for an explanation for a decision. Systems of the "associative memory" type, for example, store all cases and their diagnoses in a memory (database). The diagnosis of a new case is determined by searching for the most similar cases and assigning the majority of the diagnoses of the most similar case. An example of this type of algorithm is the k-nearest neighbor classification algorithm [30]. Attempts to analyze skill-based algorithms in detail only lead to an understanding of the mathematical model used in such a system. For example, patient A's diagnosis is D because A is most similar to patient X, who had a diagnosis of D in the past. Moreover, fairness and nondiscrimination against minorities, as well as other ethical requirements, such as not harming people, cannot be guaranteed or enforced in skill-based systems (see https://digital-strategy.ec.europa.eu/en/library/communication-artificial-intelligence-europe, accessed on 15 December 2021).

4. Transition from AI to XAI in Biomedical Data Science

4.1. Methods to Identify the Decision Processes of Subsymbolic "Black-Box" Algorithms

The main types of classifiers used in machine learning are symbolic [31] or subsymbolic [32] classifiers. For symbolic classifiers, the decision of how a classification is arrived at can be interpreted by a domain expert as a combination of conditions on the features. For example, a symbolic classifier may consist of a set of rules that are hierarchical in a decision tree or non-hierarchical. This is consistent with what is currently required for an XAI. In contrast, subsymbolic algorithms do not make transparent the exact criteria used to assign a subject to a particular class, e.g., healthy or sick. An example of subsymbolic classifiers are random forest classifiers [33,34], which are based on a number of different, uncorrelated and simple decision trees. Class assignment is done by majority voting on many well-behaved trees.

However, this research focuses on combining the advantages of both types of classifiers, i.e., the high performance of the subsymbolic algorithms and the transparency and, hence, trustworthiness of the symbolic classifiers. It is readily possible to narrow down the number and type of features of the subject on which the decision is based and to rank their importance, but the exact process remains a black box. Further analysis is needed to uncover the exact decision process. For example, one method developed for random forests is to analyze representative trees in the forest [35] (Figure 2). The result is a small selection of well-functioning prototypic trees out of a total of 1500 trees in the random forest on which possible decision processes can be traced. An alternative method for extracting non-hierarchical decision rules from the decision process in random forests, and also in other subsymbolic classifiers, is the so-called LIME method (local interpretable model-agnostic explanations) [36]. This learns an interpretable model locally around the single prediction of a trained AI, e.g., a random forest. This is achieved by changing the assignment rules for a single data instance, e.g., a single patient, by changing the feature values and then observing the resulting impact on the classification. The result of LIME is a set of rules representing the contribution of each feature to a prediction for a single data instance, which is a form of local interpretability.

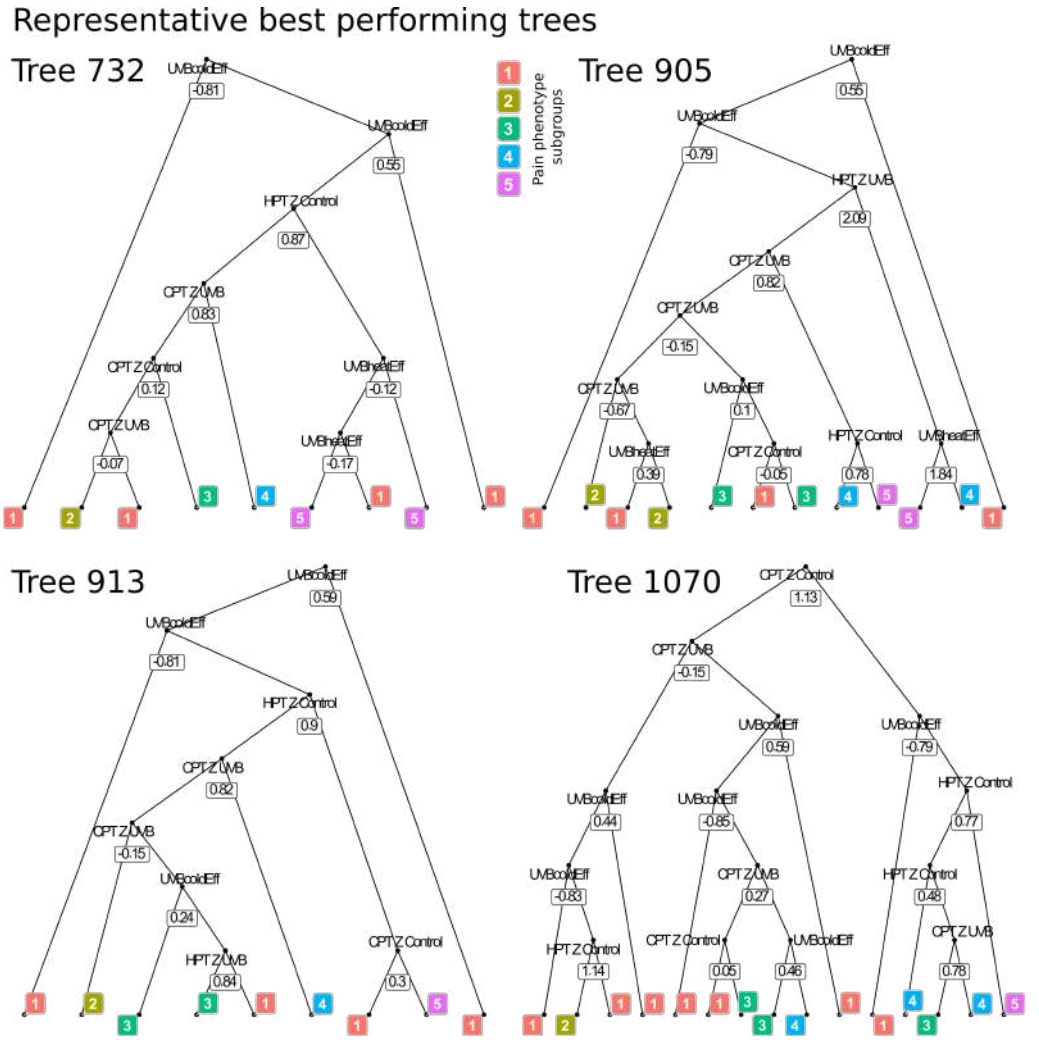

Figure 2. Example of an attempt to make subsymbolic classifiers transparent in terms of the decision structure along which class assignment occurs by extracting well-behaved and representative trees from a random forest classifier. A subsymbolic random forest classifier with a size of 1500 trees was created that contained up to d = 3 pain-related variables by setting hyperparameters, using the R library "randomForest" (https://cran.r-project.org/package=randomForest [26], accessed on 15 December 2021). The pain-related variables are from [37] and consist of thresholds for different stimuli recorded in quantitative sensory tests in a study of experimentally induced pain in humans, namely, pain thresholds for noxious heat and cold. The variables used included heat pain thresholds (HPT) and cold pain thresholds (CPT). The pain data included the z-transformed pain thresholds for heat or cold stimuli recorded under control conditions and after UV-B irradiation, and the UV-B effects recorded as the difference between the z-transformed thresholds ($zHPT_{baseline}$, $zHPT_{UVB}$, $zCPT_{baseline}$, $zCPT_{UVB}$, $UVBEff_{Heat}$, and $UVBEff_{cold}$) acquired from 84 healthy subjects. Analysis of representative trees in the forest resulted in the four trees shown in panel C. This analysis used the trained random forest and the data to run predictions while identifying representative trees based on the d2 metric [35] using the Euclidean distance. The result was trees number 732, 905, 913 and 1070 of the 1500 trees in the forest. These calculations and plots were performed using the R libraries "reprtree" (https://github.com/araastat/reprtree/blob/master/R/ReprTree.R [38], accessed on 15 December 2021) and "ggraph" (https://cran.r-project.org/package=ggraph [39], accessed on 15 December 2021). The figure shows the representative trees, with the class assignments as colored leaves at the respective bottoms. The figure was created using the R software package (version 4.1.2 for Linux; https://CRAN.R-project.org/ [17]) and the R package "ggplot2" (https://cran.r-project.org/package=ggplot2 [40], all accessed on 15 December 2021).

4.2. XAI Designed for Non-Developers

The detailed understanding of the mathematical details of an AI algorithm may be possible for experts in statistics or computer science. However, when it comes to the fate of humans, this "developer's explanation" is not enough. For example, the World Bank requires of AI systems for credit scoring "the ability of humans to interpret, understand, explain, and justify decisions made with methods that use a large number of variables" [41]. Ultimately, a human must be able to take responsibility for the consequences of an AI system's decision.

Linear models, especially structurally simple ones, are assumed to be understood by mathematicians, statisticians, or computer scientists (i.e., the developers). However, these models are limited in what they can do. The development of "parallel distributed processing" models has attempted to overcome the limitations of linear systems [42]. Such models consist of a very large number of nonlinear functions, often referred to as neural networks or forests of decision trees. By adjusting many parameters (e.g., "synaptic weights"), such a model can "learn" to reproduce given input–output situations. Due to the large number of interacting elementary processes (neurons), understanding the details of such a system in finite time is neither intended nor possible. Such systems are referred to as "black boxes". The comprehensibility of the system is sacrificed in favor of efficiency and simplicity of development. For example, it took many hundreds of man years of acoustic and statistical specialists to develop the first speech recognition program [43]. Modern so-called deep learning neural networks require only computational power to optimize a standard algorithm to achieve even better quality [44]. The comprehensibility of such systems was traded for the capability (accuracy) of their performance [41].

XAI requires knowledge discovery methods that are machine-usable and explainable to a domain expert or even a layperson. The most precise definitions of XAI [6] go back to research on knowledge-based AI [45,46]. A truly explainable AI (XAI) system is one that draws its conclusions based on a model that is understood and accepted in depth by a human expert in the field in which the XAI is used (domain expert). This understanding and acceptance of the AI inference model must be such that the expert is willing to ultimately assume legal responsibility for the AI's decisions. Such XAI systems cannot rely on their skills alone. Instead, they must make their decisions using scientific logical reasoning based on recognized expertise. XAI systems must be able to explain each decision and its derivations in a way that can be understood and comprehended by the domain expert (domain intelligibility).

Consequently, XAI systems must be based on (machine-processable) knowledge oriented to human language (symbolic systems). The so-called "expert" or "knowledge-based" systems [47] fulfill this requirement. They are typically based on a representation of the concepts, facts, rules, relationships, and theories in a given domain [48]. The GeneOntology knowledge base [24] is an example of such a knowledge representation in the field of cell biology and genetics. An XAI system arrives at conclusions (decisions/diagnoses) by applying formal methods of scientific reasoning, e.g., predicate calculus [49]. There are machine learning AI systems that use this type of reasoning, i.e., the individual decision steps are provided directly by so-called "symbolic" machine learning methods that base the class assignment of a case on a set of hierarchically or non-hierarchically organized rules [50–53]. Examples include hierarchical classification and regression trees (CART [13]) or non-hierarchical repeated incremental clipping for error reduction (RIPPER [54]). Among the symbolic tree-based algorithms, the so-called "Fast and Frugal Trees" (FFTs [55,56]) provide particularly simple decision trees, usually consisting of 1 to 5 pieces of information, which makes them particularly suitable for biomedical problems, as they mimic the processes of making a clinical diagnosis [57].

However, one important requirement for machine-learned AI systems is often overlooked: the comprehensibility of the knowledge used in the system to a domain expert. An important requirement for comprehensibility is simplicity. Machine-learned symbolic systems often lack this property. For example, decision trees may consist of hundreds

of conditions. Identical subtrees may be used repeatedly in different branches of such a decision tree (Figure 2). It is acceptable for a computer algorithm to base a decision on hundreds of conditions. Humans, on the other hand, have a limited capacity in terms of the complexity and redundancy of models or explanations. According to Miller's law, the typical limit of human information processing capacity is 7 ± 2 elements [58]. XAI explanations must, therefore, be as simple as possible (Occam's razor) and use abstractions (generalizations) from example situations.

5. Main Biomedical Goals of XAI

Approaches to explain the decisions of deep learning algorithms for biomedical tasks have their main focus on visualizing the elements that contributed to each decision [59]. For example, one of these methods is interactive heat maps [60]. There are several ways in which such mechanical explanations can highlight which input is relevant to an output obtained, using gradients as a multivariable generalization of the byproduct, where the neural network is viewed as a function and the explanation is based on the gradient of the function available from the backpropagation algorithm [59,61]. The volume of studies on machine learning interpretability methods in recent years demonstrates the growing interest in this research area. However, despite the rapid growth, the goal of understandability for experts (statisticians) is sacrificed for the understandability for professionals [62].

An attempt was made to define the position of XAI in the biomedical context in general. The main goal of achieving explainability and traceability of machine-learning-based decisions is inherent, and a further breakdown was proposed based on a study of the terms frequently used in the XAI context [6]. Accordingly, XAI should serve the following goals, namely (i) trustworthiness, (ii) causality, (iii) transferability, (iv) informativeness, (v) trust, (vi) fairness, (vii) accessibility, (viii) interactivity, and (ix) privacy awareness.

5.1. Trustworthiness

Machine learning (ML) methods for classification tasks decide which class (e.g., a clinical diagnosis) is appropriate for a given case. When a person's fate depends on the outcome of such a decision made by an algorithm, the trustworthiness of the ML system is of particular importance.

The term trustworthy AI is increasingly used as an alternative to the term XAI in clinical research and AI-assisted decision making when the concept of XAI is used in the context of patient–physician interaction. The idea behind trustworthiness of an AI is to gain the trust of individuals or organizations in the AI model by explaining the characteristics and reasons for the AI output, which helps to achieve the full potential of the AI. For example, if neither physicians nor patients trust an AI-based recommendation for a clinical diagnosis, it is unlikely that any of them will follow the recommendation [63]. A solution is provided by the above-mentioned LIME method. A recent example of its use was the transparent assignment to specific pain phenotype clusters based on random forest [37]. However, in the same report, it was also shown that the LIME method is not perfect, and rules can only be expected for a subset of the instances in the data set.

5.2. Further and Related Goals of AI in Biomedicine

The following explanations and examples attempt to capture more of the main goals identified in [6], with a focus on biomedical research and, in particular, clinical decision making based on AI or machine-learned algorithms (Figure 3).

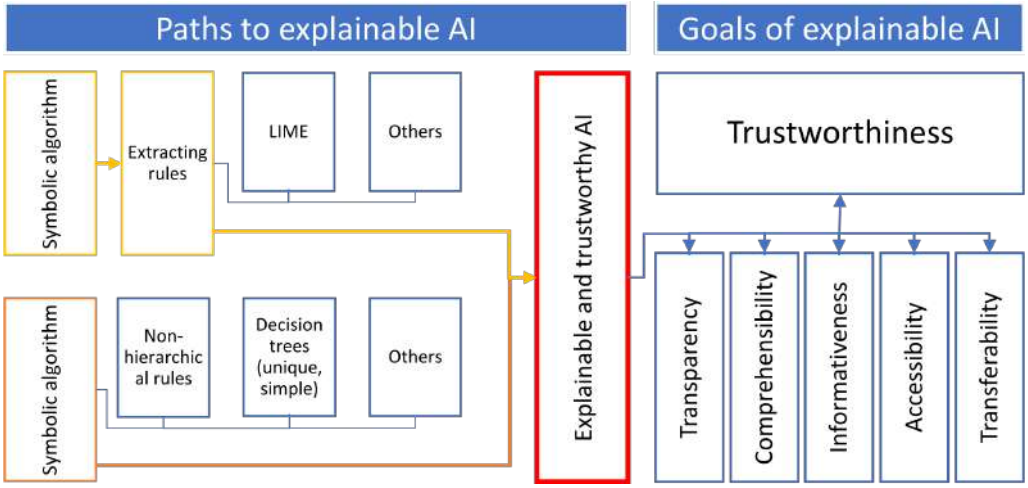

Figure 3. Schematic representation of achieving trustworthy AI in biomedicine, including clinical decision making, and the requirements for such XAI in this environment. The left part shows paths to trustworthy AI. The AI-based decisions can be implemented as symbolic algorithms, which often use rules or small rule sets for classification, or a sub-symbolic and often more powerful type of machine learning algorithm is used, to which further methods are subsequently applied, such as local interpretable model-agnostic explanations (LIME [36]) to extract comprehensible rules for class assignment. The right part shows the main objectives assigned to an XAI in the biomedical and clinical context, as proposed in [6], with the main goal of making AI-based clinical decisions trustworthy by being comprehensible to both the physician and the patient. The figure was created using Microsoft PowerPoint® 365 (Redmond, WA, USA) on Microsoft Windows 11 running in a virtual machine powered by VirtualBox 6.1 (Oracle Corporation, Austin, TX, USA) on a computer running Ubuntu Linux 20.04.03 LTS 64-bit (Canonical, London, UK).

5.2.1. Transparency

Transparency is also grouped under the terms black box to white box, intrinsic explanations, understandability, or comprehensibility. All of these terms refer to a precise description of the mathematical/statistical/algorithmic details of how the AI model works internally [6]. This type of explanation may be understandable to statisticians, mathematicians, and/or computer scientists. However, it is usually useless to the physician or the patient. In banking, for example, the "transparency" of an AI deciding whether a customer is eligible for a loan is based on the equation $ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1$ [64]. The success of today's "subsymbolic" AI is based on trading the understandability of the model for its performance in terms of the accuracy of the AI's predictions or class assignments, as described in more detail below in the History section. Such models, e.g., artificial neural networks (ANN [65]) or random forests [33,34], use a large number of nonlinear functions as neurons or decision trees coupled by thousands of coupling factors, such as synapses and weights, respectively, which in deep learning can be organized in many layers. Thanks to the power of modern computers, up to current PCs, the many thousands or millions of internal parameters can be optimized for these models so that those functions can be approximated that map the high-dimensional multivariate input data to multivalued outputs, such as different diagnostic classes. In this way, the individual elements (neurons, trees) can be accurately described, but the resulting collective behavior of the system cannot be understood. This can be compared to the impossibility of explaining thoughts or ideas by the firing of neurons in the brain. In systems theory, this is considered one of the central properties of emergent systems [66–68].

5.2.2. Comprehensibility

Being also referred to as interpretability, the comprehensibility of an explanation means the provision of a causal and logical deduction of the results (decisions) from the

given facts (input data) using the terms, formulations and methods of decision making in the respective subject area. Comprehensible explanations are formulated using deductive logic, considering approximations (fuzziness) and risks. This is the very meaning of explainability of an AI [6] and is used as explainable AI (XAI) in the rest of this paper. A common definition of XAI is that it is AI in which the results and the derivation of the solution can be explained in a way that is understandable to humans. The term "explainable" is often used and defined very differently by researchers, as there is no concrete mathematical definition [69]. Most importantly, the human who needs to understand the AI's decision is controversial: is it the patient, the doctor, or a computer scientist or statistician? Another problem is the interchangeable misuse of interpretability and explainability in the literature, as there are significant differences between these concepts, but in all existing definitions, the term "understandability" emerges as the most essential concept in XAI [6].

However, there is a consensus that XAI must ensure that computational decisions are transparent so that they can be communicated to affected patients in an understandable way when it comes to biomedical and especially clinical decisions. In this regard, the goal of XAI research is to define the specific interests, goals, expectations, needs, and requirements for artificial systems and to drive their implementation.

5.2.3. Informativeness

The informativeness of AI is required, for example, in clinical decision support systems that assist physicians in diagnostic or therapeutic tasks. Such systems are based on AI and are increasingly used in clinical practice, with a current focus on medical imaging. Alzheimer's disease can be diagnosed from magnetic resonance images by training deep neural networks to identify abnormal brain regions [70]. Similarly, deep neural networks were introduced in clinical imaging to facilitate decision making using these data [71]. However, while the results of these analyses appear reasonable to a medical expert because they are consistent with medical knowledge and, as such, could be communicated to the patient, the exact mechanisms of how a diagnosis is made for a particular patient remain vague. Deep neural networks are subsymbolic classifiers in the above sense, as are random forests, whose hundreds or thousands of decision trees also cannot be grasped in full detail by the physician and communicated in an understandable way to the patient. Informativeness aims at simpler models of what an AI does internally such that this abstraction provides more information to a user [6].

5.2.4. Accessibility

Accessibility can mean involving end users in the process of improving and developing an AI algorithm, as previously suggested [6]. Furthermore, accessibility can be considered the ability to make machine-learning-based decisions without deep programming and AI knowledge. This can be done via interactive pre-packaged software, such as the R package "rattle" used in the second chapter of this report or other interactive tools such as the R libraries "AdaptGauss" https://cran.r-project.org/package=AdaptGauss, which provides interactive fitting of Gaussian mixture models [72], "pguIMP" for preprocessing biomedical laboratory data sets, including imputations of missing values by machine learning (https://cran.r-project.org/package=pguIMP [73], all accessed on 15 December 2021), and many others that would require separate review exceeding the present scope.

However, accessibility can also be understood in terms of intellectual accessibility. This goal is often realized in symbolic forms of AI, often in simple machine learning algorithms that take the form of hierarchical or non-hierarchical rules. A symbolic rule-based classifier that included 21 individual or aggregate parameters, including demographic characteristics, psychological, and pain-related parameters recorded early after breast cancer surgery, predicted the subsequent development of persistent pain with a cross-validated accuracy of 86% and a negative predictive value of about 95% when most non-hierarchical rules were used [74]. Another example from pain research is how subsymbolic AI can be subjected

to further analysis to extract the individual decision process for each case, or how it can be complemented by symbolic AI that provides understandable explanations for group assignment, although the exact decision process may differ from that of subsymbolic AI [37]. This has been elaborated in more detail in a visual analysis system for multi-model comparison of predictions for clinical data [75]. The system allows comparison and evaluation of different AI models based on their interpretable information, with the goal of assisting clinicians in decision making. The different models are compared in terms of the predictive criteria used, and the consistency of their application is evaluated.

5.2.5. Transferability

Transferability was cited as the second most common reason for using XAI in research [76]. Transferability means that explaining how an AI model works serves to better understand the underlying problem so that the solution can be more easily applied to a different application or problem. In the breast cancer cohort mentioned above [74], the initial set of rules for predicting pain persistence included information collected from patients through repeated use of comprehensive psychological questionnaires. Once this proved informative, supervised machine learning could be used to reduce the questionnaires to items relevant to the pain context. This was accomplished by creating a shorter form of questionnaires that contained only seven items, representing 10% of the original psychological questions, but yielded the same predictive performance for pain persistence as the full questionnaires [77]. Certainly, this short questionnaire is much easier for clinicians and patients to understand than the more general full questionnaires.

6. Concerted AI Interpretation between Informatics and Biomedical Domain Experts

Computer science is described as a rapidly growing multidisciplinary field that uses advanced computing capabilities to understand and solve complex problems [78]. It inherently requires collaboration that involves sharing and collaboration on information and methods between professionals of different domains, such as physicians and machine learning experts [79]. This sets XAI, i.e., explaining AI to non-mathematicians, into the center of the classifier development workflow rather than placing it at its end. Concerted model building, involving a variety of experts from different fields, is necessary to identify and eliminate machine learning pitfalls, such as confounding variables and surrogate markers, commonly referred to as data-leaking covariates. An example of this is the identification of a protective effect of the 5-HT_3 serotonin receptor antagonist ondansetron, an antiemetic routinely used to treat nausea and vomiting, against hospitalization-related venous thromboembolism [79]. Whereas an initial classifier achieved a ROC-AUC for risk prediction of 0.92, after a concerted effort by biomedical and computer scientists to exclude data-leaking covariates such as specific pharmacological prophylaxis or treatment of thromboembolism, the ROC-AUC decreased to 0.87, which seems to more realistically capture the benefit of ondansetron in this context. A purely statistical approach without consulting biomedical expertise may not be sufficient here, as it has been emphasized that it is often difficult to distinguish between confounding and mediating variables in statistical analyses [80]. It appears that expertise is required to deal with confounding variables so that an expert can decide which variable can potentially be considered a confounding variable (rather than a mediating variable) or a surrogate marker [81]. As stated elsewhere, prediction only requires correlation, but understanding requires significant knowledge underlying the causal mechanisms [79,82]. Other problems in machine learning model building include the shortage of data points relative to the number of available variables to select from and sparse data sets where many of the labels are missing. Again, one way to address these shortcomings is proposed to consist of the involvement of domain human experts in various steps of data set construction, model training and evaluation and, in particular, the integration of prior medical knowledge [83]. An example for these effects is the identification of olfactory effects of various drugs from a data set with many candidate drugs applied to a limited number of patients, which has also been assessed

by both biomedical and computer science experts [84]. It is noteworthy that this report is also the result of a collaborative project between authors whose original fields of study are medicine/data science, biology, or computer science.

7. XAI in Biomedical Publishing

Artificial intelligence and its most popular application, machine learning, increasingly permeate many areas of daily life and science, including biomedical research. An automated search of the PubMed database on 25 September 2021, using the R library "RISmed" (https://cran.r-project.org/package=RISmed [85], accessed on 15 December 2021) yielded 166,938 hits with the search terms ("machine-learning OR artificial intelligence OR explainable artificial intelligence") and 138,556 hits with the search terms ("artificial intelligence OR explainable artificial intelligence"). When excluding reviews by adding "NOT review[PT]" to the search terms, 153,868 and 127,438 hits were obtained, respectively. The earliest hit using the MeSH term "artificial intelligence" was a 1951 report on a neurological research robot [86]. Publications per year were infrequent until the 1980s and did not exceed 100 per year until 1986 (Figure 4A). Since then, publication activity has accelerated and reached a temporary peak in 2020, when the above searches, which included all types of publications, yielded 25,622 and 19,302 hits, respectively.

In biomedical research, the concept of XAI was only recently mentioned in publications. XAI accounts for only a small portion of the hits in the second search above. An automated search of the PubMed database as above, using only the term ("explainable artificial intelligence"), yielded 340 hits on 25 September 2021, with the earliest publication dating from 1990 [87]. However, XAI is increasingly included in publications, and most publications are from the last three years, with 113 articles from 2020 and 172 articles already from 2021 (Figure 4B), which fits well with the publication dates of the seminal articles mentioned in the above chapters on concepts of AI and XAI.

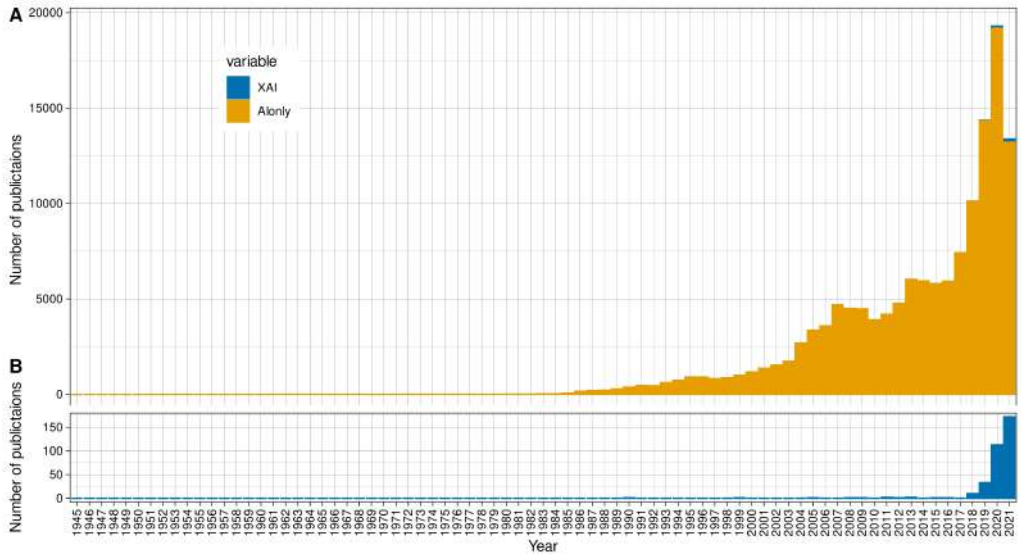

Figure 4. Stacked bar chart of publications listed in PubMed per year, with particular emphasis on publications found with the search term ((("machine-learning OR artificial intelligence OR explainable artificial intelligence") NOT (review[PT])), with the proportion of publications found with the search term ("explainable artificial intelligence" NOT (review[PT])) separated as blue parts of each bar. (**A**) Publications per year were infrequent until the 1980s and did not exceed 100 per year until 1986. (**B**) Enlarged view of the latter search, i.e., for "XAI" only. The figure was created using the R software package (version 4.1.2 for Linux; http://CRAN.R-project.org/ [17]) and the library "ggplot2" (https://cran.r-project.org/package=ggplot2 [40], all accessed on 15 December 2021).

8. Discussion

The purpose of artificial skill-based (AS) algorithms is to use examples to learn how to classify (diagnose) cases in such a way that this can be generalized to unseen cases. This is akin to teaching a child to ride a bicycle in a parking lot with the expectation that he or she will later be able to ride on the street. This is ideal for application areas such as drug repurposing or protein secondary or tertiary structure prediction (for a summary, see, for example, [88]). Deep learning neural networks are the prototypical example of this type of AI. Skill-based algorithms can surpass the current state of the art in in patient categorization. However, they do so by intentionally sacrificing explainability. For those application domains that target performance, this may be appropriate, e.g., for purely technical applications such as AI-based detection and separation of cell types, which are often implemented in close proximity to the laboratory equipment used to collect or generate these biomedical data. The literature does not cover this aspect of different types of application domains: aiming for skill and performance versus aiming at knowledge and explainability.

However, where the decisions made by AI are relevant to people's lives, knowledge-based AI should be used. For this type of application, the AI's decisions must be understandable to the medical or other professional, and the application of AI methods in the medical field should be limited to user-understandable systems. These are models that provide a causal and logical derivation of their decisions from the given multivariate data using the terms, formulations, and methods of medical decision making. This means that such systems should use a formal, i.e., understandable to humans (subject matter experts), knowledge representation. This is the viewpoint taken here for XAI. So-called white-box explanations, which provide intrinsic explanations, are left to mathematicians, statisticians, or computer scientists who deal with the internal workings of the models. XAI should focus on ensuring that computational decisions are made transparently and in a form that can be communicated to medical staff and patients in an understandable way. This approach is likely to accelerate the adoption of XAI in biomedical research and subsequently in clinical practice. This type of XAI can be implemented in a variety of ways, such as transforming subsymbolic AI into symbolic systems using knowledge discovery methods. XAI is an active research topic in computer science. Because of their direct impact on the realization of patients' right to informational self-determination, their results have a direct impact on biomedical research and clinical practice and are rapidly being transferred from the field of theoretical computer science to practical applications in clinical work.

The present XAI approach may differ from alternative approaches in that we explicitly define XAI as an algorithm that makes the decision as to why a particular individual should be assigned to a particular class (diagnosis) in a manner that is accurate and logically comprehensible to those involved. The steps of the decision-making process should be accessible and understandable at least to the expert in the field, who can then explain these rules to the affected individual. Ultimately, it should also be possible for the person affected to directly understand the decisions made by the system. This explicitly goes beyond making the decision-making process understandable to the data scientist who knows, for example, the mathematical background of a regression-based classifier. This background, expressed in equations, can hardly help most patients understand why an algorithm assigned them a particular diagnosis. It also goes beyond the mere plausibility of the features used to make the decision, which may allow a vague association with the classification but not the precise reasons for an individual. It also goes beyond the claim that it is sufficient for an XAI to meaningfully connect the input space, i.e., the available biomedical information, with the output space (clinical diagnosis). In contrast, the exact decision-making process must be made transparent to a medical professional for an algorithm to qualify as an XAI. As outlined above, XAI enables humans to get "into the loop" of machine learned systems, for example, to decide which variables can potentially be considered "surrogate markers". Such markers are effective in predicting the diagnosis because their values are a consequence of the diagnosis. Such XAI can than contribute to (i)

the understanding of the mechanisms of a particular disease and (ii) building trust that the results found by machine learning are not spurious [82].

9. Concluding Remarks

Although the present report emphasized the need for comprehensibility of AI-based biomedical decisions, it should not be ignored that it falls short to require interpretability only from statisticians involved in the medical decision-making process. Biomedical terms and methods may be similarly incomprehensible to a nonbiomedical expert as AI-specific terms and methods often are to the biomedical expert. Although the medical environment is the medical expert's home professional field, with the increasing use of AI in the field, it is not enough to ask incoming disciplines to explain their methods without viewing this task as reciprocal, including the need for both informaticians and medical professionals to learn about each other's disciplines. It is, therefore, the joint responsibility of biomedical and informatics experts to establish a common basis of terms and concepts for discussion, which each expert can then explain to the other expert and both experts to the patient. To return to the introductory example, CD19 is probably as unfamiliar to a computer science expert as SVM is to a medical professional. Both have the task of making themselves understood by the other expert and passing on their mutual understanding to the patient.

Author Contributions: A.U.—Conceptualization of the project, literature recherche, and writing of the manuscript. D.K.—Writing of the manuscript, literature recherche. J.L.—Conceptualization of the project, literature recherche, programming, writing of the manuscript, data analyses and creation of the figures. All authors have read and agreed to the published version of the manuscript.

Funding: This work has received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors have declared that no competing interests exist.

References

1. Luger, G. *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*, 5th ed.; Pearson Addison Wesley: San Francisco, CA, USA, 2004.
2. Lötsch, J.; Ultsch, A. Machine learning in pain research. *Pain* **2017**, *159*, 623–630. [CrossRef]
3. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; The MIT Press: Cambridge, MA, USA, 2012; p. 1096.
4. Dhar, V. Data science and prediction. *Commun. ACM* **2013**, *56*, 64–73. [CrossRef]
5. Hamon, R.; Junklewitz, H.; Sanchez, I. *Robustness and Explainability of Artificial Intelligence—From Technical to Policy Solutions*; Publications Office of the European Union: Luxembourg, 2020. [CrossRef]
6. Arrieta, A.B.; Díaz-Rodríguez, N.; Ser, J.D.; 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]
7. Turek, M. *Explainable Artificial Intelligence (XAI)*; Defense Advanced Research Projects Agency: Arlington County, VA, USA, 2016.
8. Hutson, M. Has artificial intelligence become alchemy? *Science* **2018**, *360*, 478. [CrossRef]
9. Brasko, C.; Smith, K.; Molnar, C.; Farago, N.; Hegedus, L.; Balind, A.; Balassa, T.; Szkalisity, A.; Sukosd, F.; Kocsis, K.; et al. Intelligent image-based in situ single-cell isolation. *Nat. Commun.* **2018**, *9*, 226. [CrossRef] [PubMed]
10. Lötsch, J.; Malkusch, S.; Ultsch, A. Optimal distribution-preserving downsampling of large biomedical data sets (opdisDownsampling). *PLoS ONE* **2021**, *16*, e0255838. [CrossRef]
11. Williams, G.J. *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery*; Use R; Springer: Berlin/Heidelberg, Germany, 2011.
12. Williams, G.J. Rattle: A Data Mining GUI for R. *R J.* **2009**, *1*, 45–55. [CrossRef]
13. Breimann, L.; Friedman, J.; Olshen, R.; Stone, C. *Classification and Regression Trees*; Chapman and Hall: Boca Raton, FL, USA, 1993.
14. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
15. Therneau, T.; Atkinson, B. *Rpart: Recursive Partitioning and Regression Trees*; R Package Version 4.1-15; 2019. Available online: https://cran.r-project.org/package=rpart (accessed on 15 December 2021).
16. Karatzoglou, A.; Smola, A.; Hornik, K.; Zeileis, A. kernlab—An S4 Package for Kernel Methods in R. *J. Stat. Softw.* **2004**, *11*, 1–20. [CrossRef]

17. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2021.
18. Inkscape Project. Inkscape, Version 0.92.5. 2020. Available online: https://inkscape.org (accessed on 15 December 2021).
19. De Rie, M.A.; Schumacher, T.N.M.; van Schijndel, G.M.W.; van Lier, R.W.; Miedema, F. Regulatory role of CD19 molecules in B-cell activation and differentiation. *Cell. Immunol.* **1989**, *118*, 368–381. [CrossRef]
20. Hastie, T.; Rosset, S.; Tibshirani, R.; Zhu, J. The entire regularization path for the support vector machine. *J. Mach. Learn. Res.* **2004**, *5*, 1391–1415.
21. Goebel, R.; Chander, A.; Holzinger, K.; Lecue, F.; Akata, Z.; Stumpf, S.; Kieseberg, P.; Holzinger, A. Explainable AI: The New 42? In *Machine Learning and Knowledge Extraction. CD-MAKE 2018*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2018; Volume 11015. [CrossRef]
22. Bayes, M.; Price, M. An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, FRS Communicated by Mr. Price, in a Letter to John Canton, AMFRS. *Philos. Trans.* **1763**, *53*, 370–418. [CrossRef]
23. Kyburg, H.E.T.C.M. *Uncertain Inference*; Cambridge University Press: Cambridge, UK, 2001.
24. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef]
25. Murray, C.D.; Dermott, S.F. *Solar System Dynamics*; Cambridge University Press: Cambridge, UK, 2000. [CrossRef]
26. Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18–22.
27. Hryniewska, W.; Bombiński, P.; Szatkowski, P.; Tomaszewska, P.; Przelaskowski, A.; Biecek, P. Checklist for responsible deep learning modeling of medical images based on COVID-19 detection studies. *Pattern Recognit.* **2021**, *118*, 108035. [CrossRef]
28. Murschel, A. The Structure and Function of Ptolemy's Physical Hypotheses of Planetary Motion. *J. Hist. Astron.* **1995**, *26*, 33–61. [CrossRef]
29. Hanson, N.R. The Mathematical Power of Epicyclical Astronomy. *Isis* **1960**, *51*, 150–158. [CrossRef]
30. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theor.* **1967**, *13*, 21–27. [CrossRef]
31. Newell, A.; Simon, H.A. Computer science as empirical inquiry: Symbols and search. *Commun. ACM* **1976**, *19*, 113–126. [CrossRef]
32. Smolensky, P. On the proper treatment of connectionism. *Behav. Brain Sci.* **2010**, *11*, 1–23. [CrossRef]
33. Ho, T.K. Random Decision Forests. In *ICDAR '95: Proceedings of the Third International Conference on Document Analysis and Recognition*; IEEE Computer Society: Washington, DC, USA, August 1995; Volume 1, p. 278.
34. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
35. Banerjee, M.; Ding, Y.; Noone, A.M. Identifying representative trees from ensembles. *Stat. Med.* **2012**, *31*, 1601–1616. [CrossRef]
36. 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. [CrossRef]
37. Lötsch, J.; Malkusch, S. Interpretation of cluster structures in pain-related phenotype data using explainable artificial intelligence (XAI). *Eur. J. Pain* **2021**, *25*, 442–465. [CrossRef]
38. Dasgupta, A. Reprtree: Representative Trees from Ensembles. 2014. Available online: https://github.com/araastat/reprtree/blob/master/R/ReprTree.R (accessed on 15 December 2021).
39. Pedersen, T.L. *Ggraph: An Implementation of Grammar of Graphics for Graphs and Networks*; R package version 2.0.5; 2021. Available online: https://cran.r-project.org/package=ggraph (accessed on 15 December 2021).
40. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2016.
41. Knutson, M.L. *Credit Scoring Approaches Guidelines-Final-Web*; The World Bank Group: Washington, DC, USA, 2020. Available online: https://thedocs.worldbank.org/en/doc/935891585869698451-0130022020/original/CREDITSCORINGAPPROACHESGUIDELINESFINALWEB.pdf (accessed on 15 December 2021).
42. Rumelhart, D.E.; McClelland, J.L. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*; MIT Press: Cambridge, MA, USA, 1986; Volume 1.
43. Huang, X.; Baker, J.; Reddy, R. A Historical Perspective of Speech Recognition. *Commun. ACM* **2014**, *57*, 94–103. [CrossRef]
44. Li, J.; Lavrukhin, V.; Ginsburg, B.; Leary, R.; Kuchaiev, O.; Cohen, J.M.; Nguyen, H.; Gadde, R.T. Jasper: An End-to-End Convolutional Neural Acoustic Model. *arXiv* **2019**, arXiv:1904.03288.
45. Michalski, R.S. A theory and methodology of inductive learning. In *Machine Learning*; Michalski, R.S., Carbonell, J.G., Mitchell, T.M., Eds.; Morgan Kaufmann: San Francisco, CA, USA, 1983; pp. 83–134. [CrossRef]
46. Craven, M.W.; Shavlik, J.W. *Extracting Comprehensible Models from Trained Neural Networks*; Computer Sciences Department, University of Wisconsin-Madison: Madison, WI, USA, 1996.
47. Yanase, J.; Triantaphyllou, E. The seven key challenges for the future of computer-aided diagnosis in medicine. *Int. J. Med. Inf.* **2019**, *129*, 413–422. [CrossRef] [PubMed]
48. Ultsch, A.; Kleine, T.; Korus, D.; Farsch, S.; Guimaraes, G.; Pietzuch, W.; Simon, J. Evaluation of Automatic and Manual Knowledge Acquisition for Cerebrospinal Fluid (CSF) Diagnosis. In *Artificial Intelligence in Medicine*; Keravnou, E., Garbay, C., Baud, R., Wyatt, J., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; Volume 1211. [CrossRef]
49. Hodges, W. Classical Logic I: First Order Logic. In *The Blackwell Guide to Philosophical Logic*; Wiley-Blackwell: Hoboken, NJ, USA, 2001.

50. Quinlan, J.R. Induction of Decision Trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
51. Loh, W.Y.; Vanichsetakul, N. Tree-Structured Classification via Generalized Discriminant Analysis. *J. Am. Stat. Assoc.* **1988**, *83*, 715–725. [CrossRef]
52. Loh, W.Y. Classification and regression trees. *WIREs Data Min. Knowl. Discov.* **2011**, *1*, 14–23. [CrossRef]
53. Loh, W.Y. Fifty Years of Classification and Regression Trees. *Int. Stat. Rev.* **2014**, *82*, 329–348. [CrossRef]
54. Cohen, W.W. Fast Effective Rule Induction. In Proceedings of the 12th International Conference on Machine Learning, Tahoe City, CA, USA, 9–12 July 1995; Morgan Kaufmann: Burlington, MA, USA, 1995; pp. 115–123.
55. Gigerenzer, G.; Todd, P.M. Fast and frugal heuristics: The adaptive toolbox. In *Simple Heuristics That Make Us Smart*; Evolution and Cognition; Oxford University Press: New York, NY, USA, 1999; pp. 3–34.
56. Martignon, L.; Katsikopoulos, K.V.; Woike, J.K. Categorization with limited resources: A family of simple heuristics. *J. Math. Psychol.* **2008**, *52*, 352–361. [CrossRef]
57. Marewski, J.N.; Gigerenzer, G. Heuristic decision making in medicine. *Dialogues Clin. Neurosci.* **2012**, *14*, 77–89. [PubMed]
58. Miller, G.A. The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychol. Rev.* **1956**, *63*, 81–97. [CrossRef]
59. Holzinger, A. Explainable AI and Multi-Modal Causability in Medicine. *i-com* **2020**, *19*, 171–179. [CrossRef]
60. Bach, S.; Binder, A.; Müller, K.R.; Samek, W. Controlling Explanatory Heatmap Resolution and Semantics via Decomposition Depth. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016.
61. Montavon, G. *Gradient-Based vs. Propagation-Based Explanations: An Axiomatic Comparison*; Springer: Cham, Switzerland, 2019; pp. 253–265. [CrossRef]
62. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **2020**, *23*, 18. [CrossRef]
63. Thiebes, S.; Lins, S.; Sunyaev, A. Trustworthy artificial intelligence. *Electron. Mark.* **2021**, *31*, 447–464. [CrossRef]
64. Skantzos, N.; Castelein, N. *Credit Scoring—Case Study in Data Analytics*; Deloitte Touche Tohmatsu Limited: London, UK, 2016.
65. Rosenblatt, F. The perceptron: A probabilist@articleic model for information storage and organization in the brain. *Psychol. Rev.* **1958**, *65*, 386–408. [CrossRef]
66. Ultsch, A. Emergence in Self-Organizing Feature Maps. In Proceedings of the International Workshop on Self-Organizing Maps (WSOM '07), Bielefield, Germany, 3–6 September 2007; Ritter, H., Haschke, R., Eds.; Neuroinformatics Group, Bielefeld University: Bielefeld, Germany, 2007.
67. Kringel, D.; Ultsch, A.; Zimmermann, M.; Jansen, J.P.; Ilias, W.; Freynhagen, R.; Griessinger, N.; Kopf, A.; Stein, C.; Doehring, A.; et al. Emergent biomarker derived from next-generation sequencing to identify pain patients requiring uncommonly high opioid doses. *Pharmacogenomics J.* **2017**, *17*, 419–426. [CrossRef] [PubMed]
68. Stephan, A. *Emergenz: Von der Unvorhersagbarkeit zur Selbstorganisation. 4. Auflage*; Brill | Mentis: Leiden, The Netherlands, 2020. [CrossRef]
69. Gilpin, L.H.; Bau, D.; Yuan, B.Z.; Bajwa, A.; Specter, M.; Kagal, L. Explaining Explanations: An Overview of Interpretability of Machine Learnin. In Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), Turin, Italy, 1–4 October 2018.
70. 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]
71. Papadimitroulas, P.; Brocki, L.; Christopher Chung, N.; Marchadour, W.; Vermet, F.; Gaubert, L.; Eleftheriadis, V.; Plachouris, D.; Visvikis, D.; Kagadis, G.C.; et al. Artificial intelligence: Deep learning in oncological radiomics and challenges of interpretability and data harmonization. *Phys. Med.* **2021**, *83*, 108–121. [CrossRef]
72. Ultsch, A.; Thrun, M.C.; Hansen-Goos, O.; Lötsch, J. Identification of Molecular Fingerprints in Human Heat Pain Thresholds by Use of an Interactive Mixture Model R Toolbox (AdaptGauss). *Int. J. Mol. Sci.* **2015**, *16*, 25897–25911. [CrossRef] [PubMed]
73. Malkusch, S.; Hahnefeld, L.; Gurke, R.; Lötsch, J. Visually guided preprocessing of bioanalytical laboratory data using an interactive R notebook (pguIMP). *CPT Pharmacometrics Syst. Pharmacol.* **2021**, *10*, 1371–1381. [CrossRef] [PubMed]
74. Lötsch, J.; Sipilä, R.; Tasmuth, T.; Kringel, D.; Estlander, A.M.; Meretoja, T.; Kalso, E.; Ultsch, A. Machine-learning-derived classifier predicts absence of persistent pain after breast cancer surgery with high accuracy. *Breast Cancer Res Treat.* **2018**, *171*, 399–411. [CrossRef] [PubMed]
75. Li, Y.; Fujiwara, T.; Choi, Y.K.; Kim, K.K.; Ma, K.L. A visual analytics system for multi-model comparison on clinical data predictions. *Vis. Inform.* **2020**, *4*, 122–131. [CrossRef]
76. 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, New York, NY, USA, 25–30 April 2020.
77. Lötsch, J.; Sipilä, R.; Dimova, V.; Kalso, E. Machine-learned selection of psychological questionnaire items relevant to the development of persistent pain after breast cancer surgery. *Br. J. Anaesth.* **2018**, *121*, 1123–1132. [CrossRef]
78. Benioff, M.R.; Lazowska, E.D.; Bajcsy, R.; Beese, J.C.; Celis, P.; Evans, P.T.; Yang, G. *Report to the President: Computational Science: Ensuring America's Competitiveness*; President's Information Technology Advisory Committee: Washington, DC, USA, 2005.

79. Datta, A.; Matlock, M.K.; Le Dang, N.; Moulin, T.; Woeltje, K.F.; Yanik, E.L.; Joshua Swamidass, S. 'Black Box' to 'Conversational' Machine Learning: Ondansetron Reduces Risk of Hospital-Acquired Venous Thromboembolism. *IEEE J. Biomed. Health Inf.* **2021**, *25*, 2204–2214. [CrossRef]
80. Bhattacharya, J.; Vogt, W.B. Do Instrumental Variables Belong in Propensity Scores? *Int. J. Stat. Econ.* **2012**, *9*, A12. [CrossRef]
81. VanderWeele, T.J. Principles of confounder selection. *Eur. J. Epidemiol.* **2019**, *34*, 211–219. [CrossRef]
82. Datta, A.; Flynn, N.R.; Barnette, D.A.; Woeltje, K.F.; Miller, G.P.; Swamidass, S.J. Machine learning liver-injuring drug interactions with non-steroidal anti-inflammatory drugs (NSAIDs) from a retrospective electronic health record (EHR) cohort. *PLoS Comput. Biol.* **2021**, *17*, e1009053. [CrossRef] [PubMed]
83. Holzinger, A. Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Inf.* **2016**, *3*, 119–131. [CrossRef] [PubMed]
84. Lötsch, J.; Daiker, H.; Hähner, A.; Ultsch, A.; Hummel, T. Drug-target based cross-sectional analysis of olfactory drug effects. *Eur. J. Clin. Pharmacol.* **2015**, *71*, 461–471. [CrossRef] [PubMed]
85. Kovalchik, S. *RISmed: Download Content from NCBI Databases*; R Package Version 2.3.0; 2021. Available online: https://cran.r-project.org/package=RISmed (accessed on 15 December 2021).
86. Fletcher, K.H. Matter with a mind; a neurological research robot. *Research* **1951**, *4*, 305–307. [PubMed]
87. Lanzola, G.; Stefanelli, M.; Barosi, G.; Magnani, L. NEOANEMIA: A knowledge-based system emulating diagnostic reasoning. *Comput. Biomed. Res.* **1990**, *23*, 560–582. [CrossRef]
88. Ching, T.; Himmelstein, D.S.; Beaulieu-Jones, B.K.; Kalinin, A.A.; Do, B.T.; Way, G.P.; Ferrero, E.; Agapow, P.M.; Zietz, M.; Hoffman, M.M.; et al. Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **2018**, *15*, 20170387. [CrossRef] [PubMed]