Neural Computing & Applications manuscript No.
(will be inserted by the editor)

The importance of interpretability and visualization in Machine Learning for applications in medicine and health care

Alfredo Vellido

Received: date / Accepted: date

Abstract In a short period of time, many areas of science have made a sharp transition towards data-dependent methods. In some cases, this process has been enabled by simultaneous advances in data acquisition and the development of networked system technologies. This new situation is particularly clear in the life sciences, where data overabundance has sparked a flurry of new methodologies for data management and analysis. This can be seen as a perfect scenario for the use of machine learning and computational intelligence techniques to address problems in which more traditional data analysis approaches might struggle. But this scenario also poses some serious challenges. One of them is model interpretability and explainability, specially for complex nonlinear models. In some areas such as medicine and health care, not addressing such challenge might seriously limit the chances of adoption, in real practice, of computer-based systems that rely on machine learning and computational intelligence methods for data analysis. In this paper, we reflect on recent investigations about the interpretability and explainability of machine learning methods and discuss their impact in medicine and health care. We pay specific attention to one of the ways in which interpretability and explainability in this context can be addressed, which is through data and model visualization. We argue that, beyond improving model interpretability as a goal in itself, we need to integrate the medical experts in the design of data analysis interpretation strategies. Otherwise, machine learning is unlikely to become part of routine clinical and health care practice.

This work was funded by the Spanish TIN2016-79576-R project.

Alfredo Vellido
Intelligent Data Science and Artificial Intelligence (IDEAI) Research Center. Computer Science Dept. Universitat Politècnica de Catalunya. Barcelona, Spain.
ORCID: 0000-0002-9843-1911
Tel.: +34-93-4137787
E-mail: avellido@cs.upc.edu

Keywords Interpretability · Explainability · Machine Learning · Visualization · Medicine · Health Care

1 Introduction

In a very short period of time, many areas of science have made a sharp transition towards data-dependent methods. Examples of this might include astronomy [1] and high-energy physics [2, 3]. This epochal change was of course heralded by the widespread adoption and integration of computers in all aspects of scientific research. In some cases, this process has been enabled by simultaneous advances in data acquisition and the development of networked system technologies. Probably in no other field this new situation is so clear as in the life sciences, where data overabundance in some of their areas has become the main driver behind the development of whole new methodologies for data management.

Research in the life sciences has turned to significantly rely on data acquisition and analysis [4]. One of the main reasons for this is the central role that genetics has come to play over the last few decades. The revolution started by genetics and genomics in the life sciences, product of the coalescence of advances in sequencing techniques (data acquisition) and computer-based data processing and analysis, can today be seen replicated in all members of the extended –omics family, including proteomics, transcriptomics and metabolomics [5].

Part of the research challenges faced in this field have thus been transferred to the computer science domain. Note that these now include issues so basic as the limitations and barriers for the storage of exponentially growing very large genomic databases [6]; the data transfer bottlenecks caused by millions of daily data requests in the form of database queries; or the potential privacy issues caused by trusting private information technology companies with data and software storage [7]. In fact, the challenges for the omics sciences now involve the four elements of data management: acquisition, storage, distribution, and analysis. Genomics data have been forecasted to become the most extreme case of big data over the next decades, surpassing astronomy and the internet [8].

This can be seen as a perfect scenario for the use of machine learning (ML) and computational intelligence (CI) techniques to address problems in which more traditional data analysis approaches might struggle. But this scenario also poses some far from trivial challenges. One of them is model interpretability and explainability, especially for complex nonlinear models. In some areas, such as medicine and health care, where explainability is paramount, such challenge might seriously limit the chances of adoption, in real practice, of computer-based systems that rely on opaque ML and CI methods for data analysis.

In this paper, we reflect on recent investigations regarding the interpretability and explainability of ML and CI methods and discuss their differential impact on medicine and health care. We also pay specific attention to one of the ways in which interpretability and explainability can be addressed in this context, which is through techniques for data and model visualization. By doing so, we aim to stress the importance of considering the human factor when attempting to enhance model interpretability in general and the importance of integrating the medical expert in the process of developing strategies to guarantee the interpretability and explainability of medical data models.

The remaining of the paper is structured as follows. The general problem of interpretability in ML and CI is discussed, in its many facets, in section 2. This is followed in section 3 by a more focused discussion of the role played by data and model visualization strategies in enhancing ML and CI interpretability. The rationale for these two sections is replicated in sections 4 and 5 for the fields of medicine and health care. The first reflects on how interpretable ML in these fields has very specific requirements and may arguably become a key to adoption. The second, again, focuses on the opportunities created by visualization in this context, given the importance of involving external experts in its knowledge generation cycle.

2 Interpretability and explainability in Machine Learning: a many-faceted problem

At the heart of ML and CI, in the end, we have families of algorithmic methods and mathematical models for data analysis. Over the decades, these methods have demonstrated their many benefits and capabilities, but they have also been seen to suffer from shortcomings that endanger or, at the very least, limit their use in a host of practical applications.

One of the latter is the focus of this paper and is the direct result of the design characteristics of many of these methods: the potential lack of interpretability and/or explainability of the data models they generate. Interpretability and explainability have become central issues in ML and CI research over the last few years [9] and at least part of that interest is caused by the resurgence of artificial neural networks (ANN) in the form of deep learning (DL) and the fact that DL risks becoming an extreme case of the *black box model syndrome* that was also a problem and a bottleneck for the application of shallow ANN methods. Being reported as such a dramatic success in ML, the lack of interpretability of DL models becomes one of the most pressing concerns in the area and recent literature reflects that. Some examples include the proposal [10] of an adversarial training scheme where model neurons "are endowed with human-interpretable concepts" and interpretable representations can trace outcomes back to influential neurons, providing an explanation of how models make predictions; the description of interpretability criteria based on analysis of deep networks in the information plane is the result of another recent study [11]. From a different but related perspective, several studies base their proposal in visual interpretations of the deep models, as we will address in the next section.

The first issue to reckon with is that interpretability in ML is by no means a fully formalized problem. Intuitively, it would seem to require that our data models can be explained (thus making it overlap with the problem of model explainability [12–14]). In a real-world application context, interpretability might be judged according only to the specific requirements of the application area (for example, the requirements for diagnosis in oncology and for anomaly detection in industrial production have little in common) in acknowledgement that different applications usually have different interpretability and explainability needs. In general terms, though, we might want to consider standardized interpretability metrics that allow us to quantitatively assess this characteristic of our models. A timely outline of requirements for the formalization of a "rigorous science of interpretable ML" has recently been described in [15]. A key idea in this study is that the main reason for an ML or CI-based system to require interpretability is some form of *incompleteness* in the way the problem addressed by the system is formulated. This incompleteness may take different forms, including a limited understanding of the problem (so that interpretation is a tool to acquire new knowledge), or a mismatch between the modeling objectives and the goals from an application viewpoint.

Interpretability and explainability might be more than just a desired property of ML and CI methods. These are concepts that can in fact take us way off data modeling technical issues. A currently popular example of this are how the recent swift advances on the application of AI in autonomous driving have raised the question of who is legally liable for accidents caused by decisions made by the autonomous system. The issue of AI legal accountability has recently been broached by Doshi-Velez *et al.* [16]. Interpretability and explainability of the system would come to the forefront of AI requirements in such a circumstance. AI-controlled autonomous weapon systems are even harsher remainders of the importance of interpretability and explainability as problems of legal accountability [17], as they pertain to both international humanitarian law [18] and military law [19]. Even further off from data modeling technical issues, ML model interpretability has also recently been scrutinized from a philosophical standpoint. In [20], Kroll discusses that lack of interpretability might be the not always unintended outcome of power dynamics surrounding software systems development, where lack of interpretability could be used as an excuse to avoid the scrutiny that is the guarantee of accountability. The author goes as far as to argue that the idea that some systems are "of necessity inscrutable" should not be accepted and that, in certain areas of application, the system designer has the latitude to avoid using unexplainable algorithms, what could be considered as malpractice.

Much more immediate is the implementation of the European Union directive for General Data Protection Regulation (GDPR). Enforced in May, 2018, it mandates a right to explanation of all decisions made by automated or artificially intelligent algorithmic systems [21]. More explicitly, such right to explanation involves providing the individual with "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing [automated decision making] for the data subject".

Needless to say, this directly involves AI in the form of ML or CI, which is illustrated by explicit reference in the legal affairs briefings of the European Parliament [22]. This directive is of compulsory application at the national level, but it is fair to say, though, that there is no consensus just yet about its true reach and implications. It has been argued that the practical implementation of such a *right to explanation* might have limited impact [23] because the GDPR might be too vague about what type of actual information subjects would be entitled to receive. It has also been argued [20] that an individual right to explanation might be of limited interest if it fails to account for aggregated population effects (e.g. decisions made on a segment of population), or if explanations are limited to a mechanistic unraveling of the algorithm operation, failing to engage the context of the problem. In any case, this directive places model interpretability and explainability right at the center of many decisions in areas in which ML and CI are bound to have an impact on the individual, such as health care and medical decision making as we argue further in the following sections.

The fact remains that interpretability is a matter to be dealt with as a human cognitive problem. As such, it could be addressed as a challenge for the design of a proper *interface* between artificial pattern recognition (APR) methods (those algorithmic techniques and statistical models that aim to express patterns from data in the expectation that they *make sense*, that is, that they can be interpreted by a human) and natural pattern recognition (NPR) human abilities (that is, the own internal pre-cognitive and cognitive brain processing of the input information, about which we have, at best, limited knowledge). Note that this is related to another concern raised about the way interpretability and explainability are currently being investigated in AI in general. This concern has been expressively stated as the risk of "inmates running the asylum" [24]. In this work, Miller and co-workers argue that it is not necessarily convenient to leave decisions on how to articulate model interpretability to the own data scientists, because this might result in models that are only interpretable for the own data scientists. Instead, it is suggested that models should be built with the assistance of concepts and expertise from the fields of philosophy, psychology and cognitive science. This is an idea for which some philosophers have recently provided support [25].

This way of addressing interpretability as a problem of interaction between APR and NPR has been described [9] as an opportunity to create *cycles of interpretation*, in which human appraisal of APR results enables the design of a formal rationale for the modification of all the elements of data analysis that have a potential impact on interpretability and comprehensibility, namely data selection and preprocessing and modeling technique choice, as graphically sketched in Fig.1.

In this cycle, the ML and CI interpretation tools (a broad palette of approaches from model sparsity and feature relevance determination to rule extraction and visualization) play the key role of being intermediaries between potentially opaque models and a human expert who needs this interpretation in order to comprehend the problem and, as a result, make decisions about

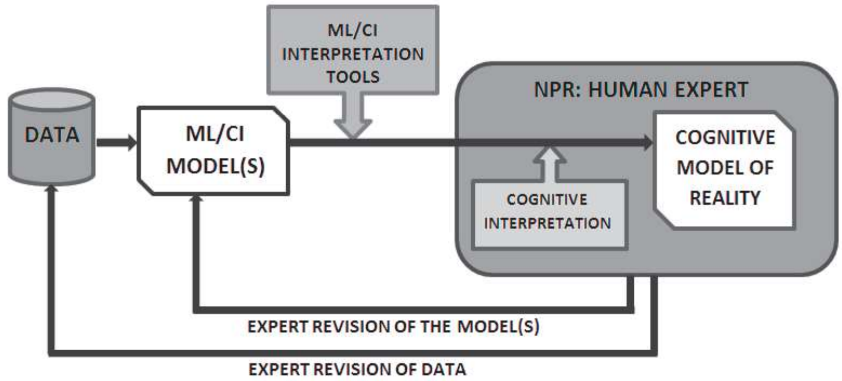

Fig. 1 Human analyst-computer ML interpretability cycle. Adapted from [9].

the data and analytical models leading to the optimization of the data analysis process. This type of framework, contemplating the interplay of APR and NPR, has recently been formalized in greater detail for one of the main tools of interpretation of ML and CI methods, namely data visualization, as we describe in the following section.

3 Visualization as a tool for interpretable Machine Learning

Visualization is a central human cognitive ability, making it the perfect conduit for interpretation of algorithmic data models. It can also be seen as a powerful tool for exploratory data analysis and one that enables inductive reasoning in a natural, seamless manner. That means that visualization can itself be a knowledge generator as it intuitively leads the analyst from observed model outcomes to potential hypothesis about the observed data. Visual analytics have, in fact, become a research field on its own [26, 27]. The transparency of ML and CI can enormously benefit from the use of visualization [28] and, once again, DL has become a hotbed for the use of visual analytics as a tool to improve interpretability. Examples of research in this area include [29–32]. Much of this work concerns Deep Convolutional Neural Networks (CNN).

From the onset, visualization calls for synergies between APR and NPR, and these can be mediated by the concept of interactive visualization. This interaction has recently been characterized [27] as leading towards three goals in the ML domain, namely understanding, diagnosis, and refinement, all of them related to the problems of interpretability of the ML results and comprehensibility of the obtained models. *Understanding* would refer to one of forms of incompleteness in the way the problem addressed with ML or CI techniques is formulated in [15], namely incompleteness defined as the limited understanding about the problem itself. Interpretation in this case is used to acquire new knowledge through visualization. *Diagnosis* refers to the interpretation of the reasons behind ML model performance, be it good, bad, or not according to expectations, using visual analytics. Interpretation would lead here to the design of better models, which would be the final goal described in [27], namely *refinement*.

As stressed in [15], interpretable ML and CI requires a formal framework to encompass and give coverage to the different ways in which this problem can be addressed. One instance of such a framework has recently been proposed for the problem of ML and CI interpretability through interactive visualization [33]. It encompasses the three goals of understanding, diagnosis and refinement outlined in [27] and it emphasizes the interactive aspect of such visualization as the main element of the interface between human cognition and algorithmic learning. As such, it is defined as a human-centered framework. Again, its main strength lies on its cyclical nature, which should make it fit any data mining formal framework in a natural manner. We graphically summarize it in Fig.2.

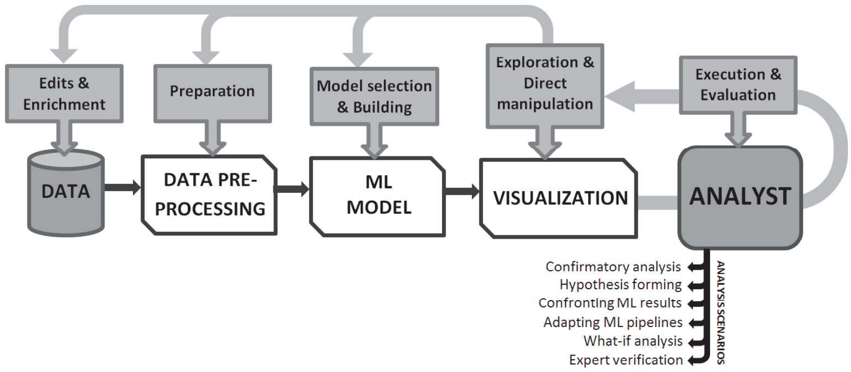

Fig. 2 Human analyst-computer ML interpretability cycle through interactive visualization. Adapted from [33]. The top row boxes describe the actions through which the analyst can adapt and modify, mediated by interactive visualization, each and every component of the analysis (data, pre-processing, models, and the own visualization process).

This conformity with the requirements of data mining methodological frameworks is guaranteed by the choice of key elements used to describe the visual interpretation interactive cycle. As seen in Fig.2, they include, in succession, the data under analysis, the data pre-processing, the ML models, the visualization strategies and the human expert (the analyst). The interaction with the analyst, mediated by visualization, allows the former to feed the acquired knowledge back into each of the elements of the cycle: visualization may induce the analyst to modify the data sample under analysis through edits and enrichment. It may also suggest modifications in the preparation of the data by fine-tuning the data pre-processing (say, by dealing with visually detected outliers, anomalous data, or data artifacts, as well as by implementing alternative forms of feature selection and/or extraction as a basis for refined visualization). Visualization may also guide model selection and building and, in an even closer loop, can provide the analyst, through interactive exploration, with clues that might advice about the quality, usability and adequacy of the own visualization techniques.

The role of the analyst as described in [33] is eminently proactive, and it is seen as the hub through which visualization can be used in different analysis scenarios, including *confirmatory analysis*, *hypothesis forming*, *confronting ML results*, *adapting ML pipelines*, performing *what-if analysis* and seeking *expert verification*. These analysis scenarios are compatible with and extend those proposed in [27] for interpretation with visual analytics. The *expert verification* listed in [33] is paramount in medical and health applications, as we argue and discuss in the following sections.

4 Interpretable Machine Learning in health care and medicine: a key to adoption

As mentioned in the introduction, we are witnessing a radical and extremely swift transition towards data-dependency in the life sciences. This process is putting much pressure on the development of novel strategies for biological data management, curation and, ultimately, analysis.

These challenges have also been acknowledged in the particular domains of health care and medicine, in which they may have had comparatively less repercussion due to the fact that, in most cases, they do not generate data in the sheer quantities that are becoming commonplace in, for instance, bioinformatics. Even here, though, the potential complexity and heterogeneity of medical data implies that "it is not yet possible to create a comprehensive model capable of considering all the aspects of health care systems" [34].

This is not to say that there are no areas in these domains that have quickly evolved to become rich information ecosystems. An example can be found in the widespread adoption of electronic health records (EHR) in medical practice, allowing clinicians networked access to multimodal medical data including image, signal and text about patients' history, hospital admissions, drug prescriptions, visiting notes and the like. As a result, EHR and medical databases have become an analytical target for natural language processing and medical text mining, to the extent that one of the currently most fruitful applications of ML in medicine is, precisely, the mining of EHR text [35].

Given the overabundance of this type of information, the use of mining techniques for the automated extraction of knowledge might seem like an impeccable idea. It has been argued, though, that, unless properly designed and implemented, these methods might lead to a reduction of skills among medical experts. The pressure put on physicians to make use of EHRs on a routine basis may also lead to content-impoverished reports due to time constrains and to an increase in the difficulty of striking a balance between an appropriate personal engagement with patients and the compliance with EHR use guidelines [36]. Feeding information-poor EHRs to data analysis methods will yield poor results. This problem may also occur in the opposite direction: Due to the limitations on the type of data that analytical methods can handle and model, we might end up with EHRs that are artificially impoverished only to fit the modeling needs instead of the medical requirements. This negative consequence of the use of ML methods in medicine has been formulated as ML methods' "focus on text and the demise of context" [37].

Data-dependence is only likely to increase in health care and medical practice, given the prominent place occupied by the concept of evidence-based medicine in the current health care agenda . A paradigmatic example of this is the area of critical care. An intensive care unit (ICU) cares for acutely ill patients, many of whom, and particularly those at a surgical ICU, are technologically dependent on life-sustaining devices such as infusion pumps, mechanical ventilators, catheters, etc. The assessment of the patient's clinical needs may change depending on the conditions present at the point of care, while their status influences the medical team's requests for further data (flow sheets, EHR, demographic information, laboratory blood tests, medical images, etc.).

This situation should again be seen as an opportunity for data science in general, and ML and CI in particular. These methods aim to extract knowledge from observational data, and this knowledge aims to be novel as well as useful and actionable in the sense that, beyond shedding light on medical issues, medical decisions could be made on its basis. This was from the onset the premise on which medical decision support systems (MDSS), often based on ML and CI methods, are developed. Although still far from common medical practice at the point of care, MDSS have made significant inroads in specific domains [38–40]. On the other hand, the advantages and possible barriers to the adoption of MDSS based on ML or similar methods have been investigated for over a decade [41].

The simultaneous creation of an information-rich medical environment and the development of techniques for knowledge extraction tailored to this domain, would seem to be a win-win situation for ML and CI. The fact, though, that these methods have not yet been universally accepted and adopted in health care and medicine should warn us about possible difficulties of adoption and non-trivial implementation challenges. Arguably, and despite the existence of plenty of evidence supporting their usefulness, ML and CI methods are likely not to be adopted in routine medical practice beyond a limited number of niche applications unless those challenges are addressed.

Three main challenges for the application of ML in medicine were recently described in [37] and one of them is precisely interpretability, expressed there as "the need to open the machine learning black box". This is indeed not an unknown challenge for ML and CI in the medical domain, because the *black box problem* was already being discussed decades ago for ANNs [42] in this context. As previously mentioned, at least part of the new interest in interpretability and explainability is driven by the new ANN models described as DL, a family of successful methods that have also found their way into the life sciences [43], as well as in biomedicine and health care [44–49]. In their review of DL applications in health informatics, Ravì and co-workers [47] rightly point out that one of the reasons that hinder the adaptation of ML methods (and DL methods in particular) in medical settings is precisely the lack of interpretability they suffer. This view is shared by Che *et al.* in [50], where gradient boosting decision trees were used to extract interpretable knowledge from a trained deep network. In a somehow related study [51], deep models are explicitly regularized so that their class-probability predictions can be modeled with minimum loss by decision trees with few nodes, amenable to intuitive interpretation.

Beyond these particular examples, self-contained tables of key bibliographic references pertaining to the use of DL in biomedicine and health care and to early attempts to imbue DL models with interpretability in medical and health care applications are included in an appendix to the paper. A few conclusions may be drawn from the studies compiled in these tables. The first is that almost all review works in Table 1 identify interpretability and explainability as key challenges to address in medical and health care applications of DL methods. From Table 2, it is also clear that the problem of interpretability can addressed in very different manners. There is a commonality, though: all methods somehow replicate human interpretation procedures. This includes mimicking the performance of DL using simpler more interpretable models such as decision trees [51,52], visual analytics on their own or combined with attention models [53–56], motif analysis [57], or semantic representations [56, 58]. All of these emphasize the need to treat interpretability at the human cognition level, beyond technical detail.

In the health care and medical contexts, this is a problem with obvious implications: if an ML or CI-based MDSS churns out decisions that cannot straightforwardly be described in comprehensible terms, a potentially insurmountable barrier is raised between the MDSS and the human subjects. For instance, the medical expert could not trust to implement a decision that she or he cannot explain to either the patient or to other medical experts, whereas the patient might not trust an expert that bases her or his judgement on unexplainable outcomes of a computer-based algorithmic method. Efforts have been made to generate underlying knowledge representations that are comprehensible to the human expert. Examples of this include rule-based representations, which are usually compatible with medical reasoning [59]; and nomograms, commonly used by clinicians because they allow visualizing the relative weight of each symptom on a diagnosis or prognosis [60]. At a higher level, and on the basis of legal safeguards such as the GDPR described in previous paragraphs, a health care system might not be willing to implement an opaque MDSS in clinical practice, in order to avoid litigation costs. Even so, there is increasing evidence that doctors welcome the assistance of MDSS in medical practice [41]. Interestingly, there is also evidence that doctors are less likely to accept MDSS recommendations if they are confident about their own decisions and the other way around. Note that this might be the cause of a negative feedback cycle in which less confident medical personnel would tend to rely more on MDSS assistance, behaviour that might in turn lead to further deskilling of that personnel [36].

One of the justifications for seeking interpretability in ML-based systems, listed in [15] and mentioned in previous sections, especially resonates with problems in the medical and health care domains. It is the idea that interpretability is needed when there is incompleteness in the formulation of a problem. This incompleteness may be due to a limited understanding of the problem. An example of this from the field of oncology (detailed later on in the paper) is the limited knowledge about what tissue metabolic fingerprint can better discriminate different tumours in diagnosis. Such limited knowledge makes the definition of sparse models (those able to yield maximum discrimination from the minimum number of metabolites) almost compulsory [61]. The incompleteness may also be due to a mismatch between the modeling objectives and the goals from an application viewpoint. Arguably, this might be one of the most common situations in medicine and health care; ML and CI-based systems may have quantifiable goals in the form of prediction, accuracy, robustness, etc., but they might be rendered useless if the obtained models do not conform to clinical guidelines. Note that computer-based systems such as the MDSS we are discussing here are often seen as an extra burden for the day-to-day practice of clinicians [62] and that clinical guidelines, even if necessary to standardize clinical practice, often conflict with the objective of personalizing medical practice [36]. In this scenario, interpretability might be seen as the way to make model performance and guidelines compliance compatible.

Arguably, interpretability and explainability are necessary to fill the gaps between raw information and human decision making. The role of ML in health care should be acting “as a tool to aid and refine specific tasks performed by human professionals” [63]. Note that this adds a key element to the issue of interpretability: the fact that it cannot be dissociated from the cognitive abilities of the human interpreter itself. In other words, that even though we need to address the technical problem of opening those ML black boxes, the problem by no means ends there and the human role must be factored in the interpretability problem. Furthermore, the human factor is key in the implementation of MDSS because, beyond the need to extract novel knowledge from data, the justification for the use of computer-based MDSS can be found in the need to counter-balance human experts’ diagnostic over-confidence [64]. This is clearly expressed in [41] when discussing the weak levels of adoption of MDSS at the point of care. Authors argue there that one of the reasons for such situation is that, by focusing too much on MDSS performance *per se*, researchers often sidestep practical questions pertaining the interface between the system and the medical expert, including, for instance, whether adequate “explanations [are] given for the system’s diagnosis”, “the form of explanation [is] satisfactory for the physicians using the system”, or “how intuitive is its use”.

There is yet another way to look at this matter: it has been argued [65] that many of the existing ML and CI approaches to biomedical data analysis do not make the effort to integrate the often available expert knowledge into the models, or use prior expert knowledge to improve model interpretability. All this means that formal frameworks for machine-human interaction in the pursue of interpretability and explainability, such as those described in previous sections, are even more important in health care and medicine than in other ambits of science. They should almost be considered as a pre-requisite in the development of ML and CI-based MDSS and medical decision making in general. In the framework described in [33], we have mentioned the particularly proactive role played by the human analyst and we have stressed the analyst's role in seeking *expert verification*. In medical and health care applications this verification is not a one-way process and we illustrate this in the following section.

Although many of the interpretability and comprehensibility challenges for ML and CI in medicine are related to the own characteristics of the methods, it is also true that the challenges can be amplified by legal issues such as the implementation of the European Union GPDR directive, enforced in 2018 and described in previous sections. The reason is clear: it would affect any ML or CI-based MDSS that could not guarantee "the right to explanation" it grants to individuals affected by automated algorithmic decisions made about them.

5 Visualization as a problem in health care and medicine

Unsurprisingly, visualization has been mentioned to play a central role as an interpretability tool for medicine in recent research [65]. In this study, we find a list of relevant open questions on this topic such as the choice of the most suitable visualization techniques for the heterogeneous and structured data to be commonly found in the biomedical context; the assessment of the most relevant features to be visualized in order to ease human experts' interpretation; or the integration of domain experts' requirements/limitations in the ML model, amongst others.

In a previous section, we have stressed the importance of visualization as a tool to assist interpretation and comprehensibility of ML and CI models in general, as well as the importance to provide a formal framework for the use of visualization with these goals. The human analyst has an active role in the interactive visualization framework proposed by Sacha and co-workers [33], acting as a bridge between visual pattern discovery (using ML tools) and knowledge validation by external experts. The importance of appraising the possible benefits of putting the "human-in-the-loop" in applications of ML to medical problems has been persuasively argued in [66]. In real-world use cases in medicine and health care, visual discovery is not always purely exploratory and, therefore, potentially interesting patterns obtained through visualization must be cross-checked and validated against expert knowledge from the domain. Quite often, this external assessment requires a committee of domain experts who, in turn, will provide feedback to the analyst that can be reinvested in the redesign of visualization experiments.

Note that this adds an extra layer of human subjectivity to the process of interpretation through visualization. As a result, the framework must care not only about a cycle involving computer-based visual techniques and a human analyst, but also about a coupled cycle involving two human parts: the data analyst and the experts from the medical and health care domains, providing *expert verification*. Finding a shared language for knowledge exchange between these human experts may, in practice, be more difficult than the task of knowledge extraction from the visual models generated by the computer systems carried out by the analyst.

This new coupled cycle is not even just about verification, but should be seen as a whole extra sub-cycle in which the medical expert may seek several things from the data analyst, including: a) guarantees of interpretability and explainability that are adapted to the specific requirements of the medical problem; b) model compliance with clinical protocols and guidelines for a given problem; c) model compliance with system-human interaction workflows at the point of care. In turn, the data analyst may seek from the medical expert: a) a clear statement of the medical requirements concerning interpretability and explainability; b) a realistic understanding of the interpretability limitations and possibilities of the analytical models; c) a clear description of the real medical decision making process in place at the point of care; and d) a guarantee of verification of the data analysis results.

This sub-cycle is represented in Fig.3 as an add-on to the cycle depicted in Fig.2. Note that, although the sub-cycle is here integrated in a visualization-based approach to interpretability, it could naturally be inserted in any systematic process that aimed to achieve model interpretability and explainability beyond visualization. The next section illustrates the importance of taking this extra loop into account using a case study in the area of neuro-oncology.

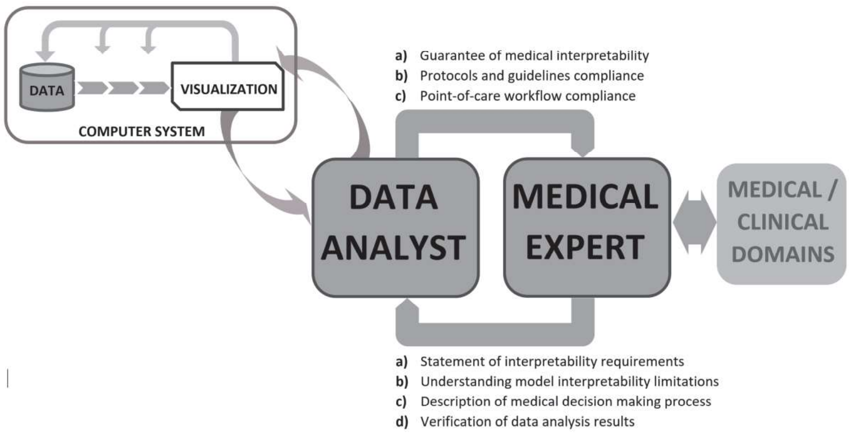

Fig. 3 Extension of the human analyst-computer ML interpretability cycle through interactive visualization proposed in [33] to account for a new sub-cycle of importance to the medical and health care domains. This new sub-cycle covers the necessary interaction between the human analyst, who must deliver data models that are interpretable and/or explainable from a medical viewpoint, and the medical expert, who must ensure that the data analyst is informed of the requirements that make interpretability valid from a medical standpoint. Arrows in the graphical depiction of the interaction between these two agents point from the agent that can deliver the interpretability item to the agent that requires it. Note that (medical) expert verification (item *d)* of the data analyst's list of requests to the medical expert) was one of the analysis scenarios considered in [33] (see Fig.2).

6 Interpretability: a case study from the neuro-oncology domain

If ML, CI and related methods are to find real application in the medical and health care domains, a *transparent* bridge of communication must be built between the analysts extracting knowledge from the available data and the medical experts making sense of that knowledge and putting it in practice. Transparency in this context works at two levels: The results of data analytics must be interpretable and the resulting interpretation must in turn be made explainable by the medical expert, sometimes transferring the explanation to the agent at the end of the chain: the patient.

Those two levels have been instantiated in the previous section in the form of a further interpretation sub-cycle, depicted in Fig.3. It includes several things that the medical expert (ME) may seek from the data analyst (DA); to simplify the description of the case study presented in this section, we will refer to these as DA2ME. Likewise, it includes things the DA may seek from the ME and we will refer to these as ME2DA.

As mentioned, the case study with which the elements of this cycle will be illustrated belongs to the area of neuro-oncology. This is an especially sensitive area of oncology dealing with tumours of the central nervous system at large and of the brain in particular. Several studies concerning the application of ML and related methods to the analysis of brain tumour data will be covered here. From the point of view of ML, several problems were investigated, including classification, clustering, visualization, feature selection and outlier analysis in different combinations. Analyses focused on Magnetic Resonance Spectroscopy (MRS) as data acquisition modality. MRS, unlike the more commonly used MR Imaging (MRI), provides a detailed and spatially-located biochemical and metabolical fingerprint of the brain tissue composition. This technique can shed light on cases that remain ambiguous after clinical investigation and can provide a more precise diagnosis of some tumour types.

The analyzed data belonged to databases resulting from the INTERPRET and eTUMOUR European research projects. Details of data acquisition and processing procedures, as well as of further database characteristics can be found in [67, 68]. These multi-centre, international databases gathered just a few hundred cases, but they are still to date amongst the largest available databases of their type. Only a number of tumour pathologies are represented by a sizeable number of cases amenable of automated analysis. They include low-grade astrocytomas, oligoastrocytomas and oligodendrogliomas (sometimes bundled as low-grade gliomas), glioblastomas and metastases (sometimes bundled as high-grade malignant tumours), and meningiomas.

One of the first ML-based studies we carried out with these data involved outlier analysis. A simple exploratory visualization of the data using Sammon's Mapping was first performed, followed by a quantification of the atypicality of the MR spectra using a generative manifold learning model [69]. Note that these are quality controlled, curated databases. Each of the MR spectra had to conform to several non-trivial selection criteria. These criteria included that "the spectrum had not been discarded because of acquisition artefacts".

Furthermore, class labelling of each case was performed "according to the World Health Organization (WHO) system for diagnosing brain tumours by histopathological analysis of a biopsy sample". It was thus unexpectedly that the methods identified clearly atypical data of two types: clearly abnormal data, way outside the main data distributions, and class-abnormal data residing within the main data distributions, but outside the distribution of their own class. The latter are most likely to be misclassified by any ML method, but are a lesser concern. Individual inspection of the former, though, revealed the existence of data acquisition artifacts that were not meant to exist.

This turned out to be a flagrant case of both ME2DA and DA2ME items. First, of ME2DA item a): (*A clear statement of the interpretability and explainability medical requirements*). It happens that the tumour types (class labels) were agreed among a committee of neuropathologists. Usually, a majority voting of two out of three experts was enough to assign the hystopathological diagnosis to a case. A human diagnosis from the MRS could be reached even if the spectrum was *partially* affected by one or more artifacts, provided that enough relevant information remained in the rest of the spectrum. The existence of artifacts will easily fool an automated classification (decision making) system, but will probably not fool a trained radiologist (the human decision maker). This situation revealed that the medical experts had not appropriately informed the data analysts of the medical terms of interpretability for this particular problem. This situation also works the other way around as a DA2ME item a) (*guarantees of interpretability and explainability that are adapted to the specific requirements of the medical problem*), because the analysts incorrectly assumed that their concept of data outlier was consistent with that of the medical experts when this was clearly not the case. The identification of both items led to address ME2DA item d) (*verification of the data analysis results*), as a thorough bipartite inspection of outlier candidates revealed an unexpectedly nuanced variety of atypicalities and combinations of artifacts that had hitherto not been characterized in the database.

A second study [70] focused on a more specific (and difficult) problem: the discrimination between two types of high-grade malignant tumours, namely glioblastomas and metastases. This problem is difficult because differentiating one from the other from their images is almost impossible, but even their MRS biochemical signatures are quite similar when taken as a whole, because they are dominated by the presence of lipids, which are the result of the anaerobic metabolism associated to their aggressive proliferation. The goal of the study was replacing classification based on the whole MRS by classification based on an intensive feature selection process (where each feature was one of the discrete frequencies of the spectrum). The first experiments were fairly successful, yielding better than the state-of-the-art classification accuracy with a parsimonious selection of frequencies. Discussion of the results with medical experts made us realize that we had failed on two accounts: first, because some of the features (frequencies) selected as important did not correspond to any known metabolites in the tissue (only a limited number of frequencies were metabolically interpretable); second, because the nonlinear classifier of choice did not allow a straightforward (i.e. interpretable and explainable) visualization of the decision surface.

In the first failure, we were not complying with any of the DA2ME items: not with a) (*guarantees of interpretability and explainability that are adapted to the specific requirements of the medical problem*); not with b) (*model compliance with clinical protocols and guidelines for a given problem*) because a decision as sensitive as discriminating between these two types of tumours cannot be supported by a doctor if based on an automated result without any biochemical/metabolical explanation; and not with c) (*model compliance with system-human interaction workflows at the point of care*) because an uninterpretable and unexplainable decision such as this cannot become the basis for a diagnostic and prognostic decision and, even less, for a decision on treatment, given that glioblastomas and metastases have completely different courses of treatment.

In the second motive for failure, it is worth highlighting ME2DA item a) (*A clear statement of the interpretability and explainability medical requirements*). It was first assumed by the analysts that the model success criterion was balanced accuracy in the discrimination between tumour types. It turned out that medical experts were happy to trade off part of such accuracy in exchange for an increase on interpretability. Such increase entailed a linear-on-the-parameters classifier and the use of only three features (the three most relevant spectrum frequencies) for straightforward visualization of the data in relation to the decision surface. In consequence, the choice of classifier was changed to a single-layer perceptron, allowing to fulfill ME2DA item d) (*verification of the data analysis results*).

The third and last of the studies reviewed in this section [71] builds on the experience gained in previous work. The important difference in this case is that the study preventively tackled DA2ME item c) (model compliance with system-human interaction workflows at the point of care) from inception. Here, both the analysts and the medical experts started from the assumption that the database might include bad data cases, even after expert assessment and database curation. The problem is that the medical definition of a "good quality" MR spectrum from the point of view of diagnostics is not yet well established: the gold standard is human-dependent, that is, despite the existence of some guidelines, the standard may well vary from expert to expert. The study proposed an ML pipeline that involves source extraction using Convex Non-negative Matrix Factorization (cNMF) and a variety of classifiers, including Logistic Regression, Linear Discriminant Analysis, AdaBoost, and Random Forests. Importantly, this pipeline was built according to what medical experts required in practice. cNMF was used for artifact detection and characterization, and the system was designed to flag potentially bad cases for radiologists' consideration.

7 Conclusions

The life sciences are at the avant-garde of an irreversible trend that, in a very short time, is placing data at the heart of scientific discovery. Medicine and health care, at their own pace, are following suit. This is an unprecedented opportunity for ML, CI and related techniques for knowledge extraction from data. Unsurprisingly, big IT companies are swiftly veering towards AI while simultaneously showing their interest in entering the medical domain and even directly becoming health care solution providers [72], sometimes with unexpected and undesired results [73].

In this paper, we have argued that there are still many barriers to overcome before these techniques become mainstream in real applications. One of them is model interpretability and explainability, which must be guaranteed before ML and CI-based MDSS are trusted by the final users, who are the medical practitioners and the health care systems for which they work. Model interpretability has become a central issue for ML in recent times, and this renewed (not new) interest can at least be partially explained by the success of DL models, which, despite their abilities, are paradigmatic examples of lack of interpretability. We have tried to convey the message that, although technical advances that aim to increase ML models' interpretability are important, medical data analysts must widen their scope to ensure the interpretability of the complete analytical process by involving medical experts in it. We have also paid special attention to one of the approaches to achieve interpretability that best encompasses the necessary integration of technical developments and human judgement, namely interactive visualization.

Sometimes, in our efforts to formalize the problems of interpretability and explainability in the application of ML and related methods, we run the risk of remaining at an excessively abstract level. This risk is especially clear in the case of medical and health care applications, where the human factor, as we have seen, extends beyond the data analyst to require the proactive involvement of the medical experts, who must convey the interpretation of the obtained data analysis results towards the patient. Here, we have tried to move from the abstract to the specific by using a case study in neuro-oncology as illustration. The research covered in this case study illustrates the often overlooked and often unexpected interpretability and explainability issues hampering the real application of (semi-)automated ML methods to medical and health care problems in general. These experiences lead us to conclude that one way to increase the interpretability and explainability of our ML models when applied to medical problems is by involving the medical experts in the analytical process. As part of that involvement, we should make sure that the interaction between the data analysts and the medical experts adheres to a formal protocol in which the specific requirements of each of these parties, as detailed in the previous sections, are clearly and unambiguously laid out. This form of interactive ML makes methodically correct experiments more difficult to implement, evaluate and replicate, as correctly pointed out in [66]. These difficulties, though, could be offset by the advantages of adhering to such protocol, which would maximize the chances of ML and CI-based MDSS being integrated in the routine of clinical practice.

Acknowledgements The author declares that he has no conflict of interest.

References

1. Wu Q, Zhu Y, Wang X, Li M, Hou J, Masoumi A (2017) Exploring high efficiency hardware accelerator for the key algorithm of Square Kilometer Array telescope data processing. In: Proceedings of the IEEE 25^{th} Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM) pp 195-195
2. Britton D, Lloyd SL (2014) How to deal with petabytes of data: the LHC Grid project. Rep Prog Phys 77(6):065902
3. Adam-Bourdarios C, Cowan G, Germain-Renaud C, Guyon I, Kégl B, Rousseau D (2015) The Higgs machine learning challenge. J Phys Conf 664(7):072015
4. Leonelli S (2016) Data-centric biology: a philosophical study. University of Chicago Press
5. Kashyap H, Ahmed HA, Hoque N, Roy S, Bhattacharyya DK (2015) Big data analytics in bioinformatics: a machine learning perspective. arXiv preprint arXiv:1506.05101
6. Marx, V (2013) Biology: The big challenges of big data. Nature 498(7453):255-260
7. Stein LD (2010) The case for cloud computing in genome informatics. Genome Biol 11(5):207
8. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE (2015) Big data: astronomical or genomical?. PLoS Biol 13(7):e1002195
9. Vellido A, Martín-Guerrero JD, Lisboa PJG (2012) Making machine learning models interpretable. In: Proceedings of the 20^{th} European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), Bruges, Belgium, pp 163-172
10. Dong Y, Su H, Zhu J, Bao F (2017) Towards Interpretable Deep Neural Networks by Leveraging Adversarial Examples. arXiv preprint arXiv:1708.05493
11. Schwartz-Ziv R, Tishby N (2017) Opening the black box of deep neural networks via information. arXiv preprint arXiv:1703.00810v3
12. Biran O, Cotton C (2017) Explanation and justification in machine learning: A survey. In: IJCAI-17 Workshop on Explainable AI (XAI) p 8
13. Pereira-Fariña M, Reed C (2017) Preface to Proceedings of the 1st Workshop on Explainable Computational Intelligence (XCI 2017)
14. Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable Artificial Intelligence (XAI). IEEE Access 6:52138-52160
15. Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv preprint. arXiv:1702.08608
16. Doshi-Velez F, Kortz M, Budish R, Bavitz C, Gershman S, O'Brien D, Schieber S, Waldo J, Weinberger D, Wood A (2017) Accountability of AI under the law: the role of explanation. arXiv preprint arXiv:1711.01134
17. Vignard K (2014) The weaponization of increasingly autonomous technologies: considering how meaningful human control might move discussion forward. UNIDIR Resources, 2
18. Davison N (2018) A legal perspective: Autonomous weapon systems under international humanitarian law. United Nations Office of Disarmament Affairs (UNODA) Occasional Papers:5-18
19. Press M (2016) Of robots and rules: autonomous weapon systems in the law of armed conflict. Geo J Int'l L (Georgetown J of Int Law) 48:1337
20. Kroll JA (2018) The fallacy of inscrutability. Philos Trans Royal Soc A 376(2133):20180084
21. Goodman B, Flaxman S (2017) European Union regulations on algorithmic decision making and a "right to explanation". AI Mag 38(3)

22. Rossi F (2016) Artificial Intelligence: potential benefits and ethical considerations, European Parliament: Policy Department C: Citizens' Rights and Constitutional Affairs Briefing PE 571.380
23. Wachter S, Mittelstadt B, Floridi L (2017) Why a right to explanation of automated decision-making does not exist in the general data protection regulation. International Data Privacy Law 7(2):76-99
24. Miller T, Howe P, Sonenberg L (2017) Explainable AI: beware of inmates running the asylum. In: IJCAI-17 Workshop on Explainable AI (XAI) p 36
25. Cath C (2018) Governing artificial intelligence: ethical, legal and technical opportunities and challenges. Philos Trans Royal Soc A 376(2133):20180080
26. Keim DA, Mansmann F, Schneidewind J, Thomas J, Ziegler H (2008) Visual analytics: scope and challenges. In Visual Data Mining (pp.76-90), LNCS vol.4404, Springer
27. Liu S, Wang X, Liu M, Zhu J (2017) Towards better analysis of machine learning models: A visual analytics perspective. Visual Informatics 1(1):48-56
28. Vellido A, Martín JD, Rossi F, Lisboa PJ (2011) Seeing is believing: The importance of visualization in real-world machine learning applications. In: Proceedings of the 19^{th} European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN) Bruges, Belgium, pp 219-226
29. Liu M, Shi J, Li Z, Li C, Zhu J, Liu S (2017) Towards better analysis of deep convolutional neural networks. IEEE T Vis Comput Gr 23(1):91-100
30. Selvaraju RR, Das A, Vedantam R, Cogswell M, Parikh D, Batra D (2017) Grad-CAM: Why did you say that? visual explanations from deep networks via gradient-based localization. In: Proceedings of the International Conference on Computer Vision (ICCV 2017), pp.618-626
31. Simonyan K, Vedaldi A, Zisserman A (2014) Deep inside convolutional networks: Visualising image classification models and saliency maps. In: Workshop Proceedings of the International Conference on Learning Representations (ICLR)
32. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: Proceedings of the European Conference on Computer Vision (ECCV) pp 818-833
33. Sacha D, Sedlmair M, Zhang L, Lee JA, Peltonen J, Weiskopf D, North SC, Keim DA (2017) What you see is what you can change: Human-centered machine learning by interactive visualization. Neurocomputing 268:164-175
34. Reza SM (2016) Transforming big data into computational models for personalized medicine and health care, Dialogues Clin Neurosci 18(3):339-343
35. Jensen PB, Jensen LJ, Brunak S (2012) Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 13(6):395-405
36. Hoff T (2011) Deskilling and adaptation among primary care physicians using two work innovations. Health Care Manage R 36(4):338-348
37. Cabitza F, Rasoini R, Gensini GF (2017) Unintended consequences of machine learning in medicine. JAMA-J Am Med Assoc 318(6):517-518
38. Safdar S, Zafar S, Zafar N, Khan NF (2017) Machine learning based decision support systems (DSS) for heart disease diagnosis: a review. Artif Intell Rev 50(4):597-623
39. Pombo N, Araújo P, Viana, J (2014) Knowledge discovery in clinical decision support systems for pain management: A systematic review. Artif Intell Med 60(1):1-11
40. Vellido A, Ribas V, Morales C, Ruiz-Sanmartín A, Ruiz-Rodríguez JC (2018) Machine Learning for critical care: state-of-the-art and a sepsis case study. BioMed Eng OnLine 17(S1):135
41. Dreiseitl S, Binder M (2005) Do physicians value decision support? A look at the effect of decision support systems on physician opinion. Artif Intell Med 33(1):25-30
42. Tu JV (1996) Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J Clin Epidemiol 49(11):1225-1231
43. Angermueller C, Pärnamaa T, Parts L, Stegle O (2016) Deep learning for computational biology. Mol Syst Biol 12(7):878
44. Mamoshina P, Vieira A, Putin E, Zhavoronkov A (2016) Applications of deep learning in biomedicine. Mol Pharm 13(5):1445-1454
45. Miotto R, Li L, Kidd BA, Dudley JT (2016) Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. Sci Rep 6:26094
46. Jackups R (2017) Deep Learning makes its way to the clinical laboratory. Clin Chem 63(12):1790-1791

47. Ravì D, Wong C, Deligianni F, Berthelot M, Andreu-Pérez J, Lo B, Yang GZ (2017) Deep learning for health informatics. IEEE J Biomed Health 21(1):4-21
48. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, Ferrero E, Agapow PM, Zietz M, Hoffman MM, Xie W (2018) Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface 15(141):20170387
49. Bacciu D, Lisboa PJ, Martín JD, Stoean R, Vellido A (2018) Bioinformatics and medicine in the era of deep learning. In: Proceedings of the 26^{th} European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2018), Bruges, Belgium, pp 345-354
50. Che Z, Purushotham S, Khemani R, Liu Y (2015) Distilling knowledge from deep networks with applications to healthcare domain. arXiv preprint arXiv:1512.03542
51. Wu M, Hughes M, Parbhoo S, Doshi-Velez F (2017) Beyond sparsity: Tree-based regularization of deep models for interpretability, In: Neural Information Processing Systems (NIPS) Conference. Transparent and Interpretable Machine Learning in Safety Critical Environments (TIML) Workshop, Long Beach (CA) USA
52. Che Z, Purushotham S, Khemani R, Liu Y (2016) Interpretable deep models for icu outcome prediction. In AMIA Annual Symposium Proceedings (Vol. 2016:371). American Medical Informatics Association.
53. Choi E, Bahadori MT, Sun J, Kulas J, Schuetz A, Stewart W (2016) Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In Advances in Neural Information Processing Systems (NIPS):3504-3512
54. Ma F, Chitta R, Zhou J, You Q, Sun T, Gao J (2017) Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In Procs. of the 23^{rd} ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD):1903-1911
55. Sha Y, Wang MD (2017) Interpretable predictions of clinical outcomes with an attention-based recurrent neural network. In: Proceedings of the 8^{th} ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB):233-240
56. Zhang Z, Xie Y, Xing F, McGough M, Yang L (2017) MDNet: A semantically and visually interpretable medical image diagnosis network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR):6428-6436
57. Nguyen P, Tran T, Wickramasinghe N, Venkatesh S (2017) Deepr: a convolutional net for medical records. IEEE J Biomed Health Inform 21:22–30
58. Hicks SA, Eskeland S, Lux M, de Lange T, Randel KR, Jeppsson M, Pogorelov K, Halvorsen P, Riegler M (2018) Mimir: an automatic reporting and reasoning system for deep learning based analysis in the medical domain. In Procs. of the 9^{th} ACM Multimedia Systems Conference (MMSys):369-374
59. Rögnvaldsson T, Etchells TA, You L, Garwicz D, Jarman I, Lisboa PJ (2009) How to find simple and accurate rules for viral protease cleavage specificities. BMC Bioinformatics 10(1):149
60. Van Belle V, Van Calster B, Van Huffel S, Suykens JAK, Lisboa P (2016) Explaining Support Vector Machines: a color based nomogram. PLoS ONE 11(10):e0164568
61. Vellido A, Romero E, Julià-Sapé M, Majós C, Moreno-Torres À, Arús C (2012) Robust discrimination of glioblastomas from metastatic brain tumors on the basis of single-voxel proton MRS. NMR Biomed 25(6):819–828
62. Ash JS, Berg M, Coiera E (2004) Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. JAMA-J Am Med Assoc 11(2):104-112
63. Reid MJ (2017) Black-box machine learning: implications for healthcare, Polygeia, April 6
64. Berner ES, Graber ML (2008) Overconfidence as a cause of diagnostic error in medicine. Am J Med 121(5):S2-S23
65. Bhanot G, Biehl M, Villmann T, Zühlke D (2017) Biomedical data analysis in translational research: Integration of expert knowledge and interpretable models. In: Proceedings of the 25^{th} European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN) pp 177-186

66. Holzinger A (2016) Interactive machine learning for health informatics: when do we need the human-in-the-loop?. Brain Inform 3(2):119-131
67. Julià-Sapé M, Acosta D, Mier M, Arús C, Watson D, The INTERPRET Consortium (2006) A multi-centre, web-accessible and quality control-checked database of in vivo MR spectra of brain tumour patients, Magn Reson Mater Phy 19(1):22-33
68. Julià-Sapé M, Lurgi M, Mier M, Estanyol F, Rafael X, Candiota AP, Barceló A, García A, Martínez-Bisbal MC, Ferrer-Luna R, Moreno-Torres À (2012) Strategies for annotation and curation of translational databases: the eTUMOUR project. Database bas035
69. Vellido A, Romero E, González-Navarro FF, Belanche-Muñoz LA, Julià-Sapé M, Arús C (2009) Outlier exploration and diagnostic classification of a multi-centre ^{1}H-MRS brain tumour database. Neurocomputing 72(13-15):3085-3097
70. Vellido A, Romero E, Julià-Sapé M, Majós C, Moreno-Torres À, Pujol J, Arús C (2012) Robust discrimination of glioblastomas from metastatic brain tumors on the basis of single-voxel ^{1}H MRS. NMR Biomed 25(6):819-828
71. Mocioiu V, Kyathanahally SP, Arús C, Vellido A, Julià-Sapé M (2016) Automated quality control for proton magnetic resonance spectroscopy data using convex non-negative matrix factorization, In: Proceedings of the 4^{th} International Conference on Bioinformatics and Biomedical Engineering (IWBBIO), LNCS/LNBI Vol 9656:719-727
72. Rajkomar A et al (2018) Scalable and accurate deep learning for electronic health records. npj Digital Medicine 1(1):18
73. Shah H (2017) The DeepMind debacle demands dialogue on data. Nature 547:259
74. Miotto R, Wang F, Wang S, Jiang X, Dudley JT (2017) Deep learning for healthcare: review, opportunities and challenges. Brief Bioinformatics 19(6):1236-1246
75. Litjens G et al (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60-88
76. Chartrand et al (2017) Deep learning: a primer for radiologists. Radiographics 37(7):2113-2131
77. Shickel B, Tighe PJ, Bihorac A, Rashidi P (2018) Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform 22(5):1589-1604
78. Zaharchuk G, Gong E, Wintermark M, Rubin D, Langlotz CP (2018) Deep learning in neuroradiology. AJNR Am J Neuroradiol 39(10):1776-1784
79. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T (2018) The rise of deep learning in drug discovery. Drug Discov Today 23(6):1241-1250
80. Kwon BC, Choi MJ, Kim JT, Choi E, Kim YB, Kwon S, Sun J, Choo J (2019) RetainVis: visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. IEEE Trans Vis Comput Graph 25(1):299-309
81. Wu J, Peck D, Hsieh S, Dialani V, Lehman CD, Zhou B, Syrgkanis V, Mackey L, Patterson G (2018) Expert identification of visual primitives used by CNNs during mammogram classification. In SPIE Medical Imaging 2018: Computer-Aided Diagnosis 10575:105752T

Appendix

This appendix includes a self-contained summary of research publications in the form of two tables. Table 1 covers a selection of general references on DL methods applied to biomedicine, while Table Table 2 focuses only on studies that deal with the problem of interpretability of DL methods applied in the field.

Table 1 Summary of key bibliographic references concerning DL in the (bio-)medical and health care application areas and, particularly, concerning the problem of interpretability. Only key selected references from the 2016-2018 period are included. This list does not aim, by any means, to be a systematic review of publications. Instead, it should be considered as a sample of the possibilities of DL methods in these areas.

Authorship	Area	Main contribution	Ref.
Mamoshina et al (2016) *Mol Pharm*	Pharmaceutical research	This review covers DL application areas with an emphasis on pharmacology, including: biomarkers, omics (genomics, proteomics, transcriptomics and multiomics), structural biology and chemistry, and drug discovery. DL challenges are singled out, including interpretability, data scarcity, task appropriate model selection and computation costs.	[44]
Miotto et al (2016) *Sci Rep*	EHR	This paper presents the "deep patient" as a "general-purpose patient representation from EHR data" obtained through a novel unsupervised deep feature learning method (stacked autoencoders (AE)), facilitating clinical predictive modeling. EHRs corresponding to 78 diseases were transformed into "deep patient representations" prior to disease prediction in two forms: disease classification and patient disease tagging.	[45]
Raví et al (2017) *IEEE J Biomed Health*	General	Comprehensive review of DL in health informatics, focusing on methods description (Deep AE, Recurrent Neural Networks (RNN), Restricted Boltzmann Machines (RBM), and CNN). Several application areas are covered: translational bioinformatics, medical imaging, pervasive sensing for wellbeing and public health. Challenges are also discussed: lack of interpretability, data availability, lack of pre-processing standards and lack of robustness.	[47]
Miotto et al (2017) *Brief Bioinform*	General	General review of DL methods in medicine and health care. Several main areas of application are categorized, including clinical imaging, EHRs, genomics and mobile (m)-health. Several challenges are identified, including data volume availability, medical data quality issues, medical data as time series, domain complexity and interpretability.	[74]
Litjens et al (2017) *Med Image Anal*	Medical image	Thorough review of DL in the specific sub-field of medical imaging, for which it makes a case for CNN as the method of choice. Other methods such as RNN, RBM, Deep Belief Networks, Generative Adversarial Networks and AE are described. Several problems are considered, including image classification, object detection, segmentation, registration, image enhancement, content-based image retrieval, etc. A further literature categorization according to anatomical region is provided.	[75]
Chartrand et al (2017) *Radiographics*	Radiology	Pedagogical introduction of DL for radiologists. Basic concepts of machine learning are explained on the way to DL and, specifically, CNNs. Problems of interest in clinical radiology are outlined, including classification, detection and segmentation, with an emphasis on methodological procedures for each of them. It also outlines limitations, namely data availability, interpretability and lack of contextual awareness in contrast with human reasoning.	[76]
Ching et al (2018) *J R Soc Interface*	General	Arguably the most thorough and authoritative review to date in the area of DL for medicine (also including the life sciences at large). It concludes that, although DL has so far provided promising advances on many medical problems, it is yet to resolve any of the most pressing ones. Three large application areas are considered, namely disease and patient categorization, fundamental biology studies and patients' treatment. Several societal challenges for DL are highlighted, including model interpretation for testable hypothesis making and legal and privacy constraints. Other challenges include scarcity of labelled data.	[48]
Shickel et al (2018) *IEEE J Biomed Health Inform*	EHR	A survey of current research on the application of DL methods to clinical tasks based on EHR data. Shallow and deep architectures are summarily reviewed. Applications are categorized as: information extraction from unstructured notes, representation learning, patient outcome prediction (static and temporal), phenotyping for personalized medicine, and deidentification for anonymization and privacy preservation. The problem of deep model interpretability is given specific attention.	[77]
Zaharchuk et al (2018) *AJNR Am J Neuroradiol*	Neuroradiology	Narrowing the perspective provided by Chartrand et al. [76], this brief review focuses on neuroradiology and, from the DL viewpoint, on CNNs. Authors suggest following the neuroradiology workflow in order to consider application areas; these would start with imaging logistics and protocol conformance and follow with image acquisition and enhancement, image transformation, lesion detection and segmentation and, ultimately, diagnosis. Authors discuss the adaptations of the neuroradiologist workflow that might result from DL-based methods adoption.	[78]
Chen et al (2018) *Drug Discov Today*	Pharmaceutic research	Brief review of the fast growing field of DL applications in pharmaceutical research. It covers the applications to compound property and bioactivity prediction, *de novo* molecular design, reaction and synthesis prediction, ligand-protein interaction prediction, and biological image analysis. DL methods are lightly covered, including CNN, RNN and AE. Specific attention is paid to DL methods capable of dealing with data scarcity.	[79]

Table 2 Summary of bibliographic references concerning DL and addressing the problems of model interpretability and explainability in the (bio-)medical and health care domain. Only references from the 2016-2018 period are included. Again, this list does not aim to be a systematic review, although it intends to cover the widest possibly palette of early attempts to address these problems.

Authorship	Area	Main contribution	Ref.
Che et al (2016) *AMIA*	Critical Care	An approach called *interpretable mimic learning* to explain DL models is proposed. DL methods are used as *teacher* models that are *mimicked* by more simple and interpretable models. The method is applied to pediatric ICU data.	[52]
Choi et al (2016) *NIPS*	EHR/cardiology	Authors present a Reverse Time Attention Model (RETAIN) as a variation on RNN with neural attention models as a way to increase model interpretability in temporal data analysis. The model is illustrated in the analysis of EHR data in a cardiology problem. This approach has been refined to include visual analytics in the form of RetainVis [80].	[53]
Zhang et al (2017) *IEEE CVPR*	Medical Image/ Oncology	CNN (for image analysis), LSTM (for language analysis) and an attention model are combined to put forward a method called MDNet. This method establishes direct multimodal mappings between medical images and diagnostic reports, generating semantically and visually interpretable outputs. The method in illustrated using an oncological image problem.	[56]
Nguyen et al (2017) *IEEE J Biomed Health Inform*	EHR	Deepr is a method for EHR feature engineering using CNN methods as the basis for data transformation prior to classification in diagnosis and treatment problems. Interpretability is guaranteed by motif analysis and visual inspection through PCA and t-SNE.	[57]
Sha et al (2017) *ACM-BCB*	Critical Care	A gated recurrent unit (GRU)-based RNN with hierarchical attention is proposed to address a problem of mortality prediction at the intensive care unit. Interpretability is guaranteed through basic visualization of relative feature relevance.	[55]
Wu M et al (2017) *NIPS*	Infectious Diseases	Authors propose a method to optimize deep models (RNN-GRU) for human-simulatability using tree regularization, which favors models whose decision boundaries can be well-approximated by small decision trees, rendering them interpretable. The method is illustrated with problems related to sepsis at the ICU and HIV infection.	[51]
Ma et al (2017) *ACM SIGKDD KDD*	EHR	Dipole, a model for predicting patients' future health using bidirectional RNNs is proposed. It includes three attention mechanisms that are used to interpret prediction results effectively.	[54]
Hicks et al (2018) *MMSys*	Urology	Authors introduce *Mimir*, an interpretative method that directly adds explainability to DL models in medical problems (here illustrated with a problem in the urology domain) by producing structured and semantically correct reports, composed of text and images. The method is presented for CNNs and relays on visual inspection using *class-activation maps*.	[58]
Wu J et al (2018) *SPIE Medical Imaging*	Oncology	Authors propose a *human-in-the-loop* (radiologist) method to interpret internal representations of CNN models for diagnostic classification of mammograms, by labeling the behavior of internal units. Network Dissection (NetDissect) for quantifying intepretability as a measure of how well individual CNN units align with sets of human-interpretable concepts.	[81]

Click here to access/download
Supplementary Material
reply-to-reviewers.pdf