ADVANCED REVIEW

WIREs
DATA MINING AND KNOWLEDGE DISCOVERY   WILEY

Interpretability of machine learning-based prediction models in healthcare

Gregor Stiglic[1,2] | Primoz Kocbek[1] | Nino Fijacko[1] | Marinka Zitnik[3] |
Katrien Verbert[4] | Leona Cilar[1]

[1]Faculty of Health Sciences, University of Maribor, Maribor, Slovenia

[2]Faculty of Electrical Engineering and Computer Science, University of Maribor, Maribor, Slovenia

[3]Department of Biomedical Informatics, Harvard University, Cambridge, Massachusetts

[4]Department of Computer Science, KU Leuven, Leuven, Belgium

Correspondence
Gregor Stiglic, Faculty of Health Sciences, University of Maribor, 2000 Maribor, Slovenia.
Email: gregor.stiglic@um.si

Abstract

There is a need of ensuring that learning (ML) models are interpretable. Higher interpretability of the model means easier comprehension and explanation of future predictions for end-users. Further, interpretable ML models allow healthcare experts to make reasonable and data-driven decisions to provide personalized decisions that can ultimately lead to higher quality of service in healthcare. Generally, we can classify interpretability approaches in two groups where the first focuses on personalized interpretation (local interpretability) while the second summarizes prediction models on a population level (global interpretability). Alternatively, we can group interpretability methods into model-specific techniques, which are designed to interpret predictions generated by a specific model, such as a neural network, and model-agnostic approaches, which provide easy-to-understand explanations of predictions made by any ML model. Here, we give an overview of interpretability approaches using structured data and provide examples of practical interpretability of ML in different areas of healthcare, including prediction of health-related outcomes, optimizing treatments, or improving the efficiency of screening for specific conditions. Further, we outline future directions for interpretable ML and highlight the importance of developing algorithmic solutions that can enable ML driven decision making in high-stakes healthcare problems.

This article is categorized under:
Application Areas > Health Care

KEYWORDS

interpretability, machine learning, model agnostic, model specific, prediction models

1 | INTRODUCTION

There is a widespread usage of artificial intelligence (AI) due to tremendous progress in technology and industrial revolution (Adadi & Berrada, 2018). The machine learning (ML) systems have shown a great success in analysis of complex patterns (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2018) and are present in a wide range of applications using structured data in different fields, including healthcare. Nowadays, complex ML models are outperforming traditional models in healthcare. Those are often difficult to understand because of lack of intuitiveness, difficult interpretation, and lack of explanation of model predictions. ML models are often difficult to interpret due to the Complexity the lack

of transparency in the process that was used to produce the final output. The lack of explanation regarding the decisions made by models represents a major shortcoming in critical decision-making processes. It is important that ML models have two main characteristics, understandability and explainability. Understandability is related to the question of how an explanation is comprehended by the observer. Interpretability and explainability are similar concepts, often used interchangeably (Carvalho, Pereira, & Cardoso, 2019). As stated by Adadi and Berrada (2018): "Interpretable systems are explainable if their operations can be understood by humans." Interpretability of highly complex prediction models are needed in healthcare because of the nature of the work. In recent years, more emphasis was given to the recognition of healthcare solutions supported by ML (Ahmad, Eckert, & Teredesai, 2018). ML models are applicable, for example, in predicting patient risk of disease, patient's likelihood of readmission, and predicting the need for care. To understand and accept or reject predictions end-user and healthcare workers must understand reasoning behind the prediction models (Ahmad, Eckert, & Teredesai, 2018).

Thus, the need for interpretable ML systems has increased as well (Carvalho et al., 2019). Further, the lack of interpretability is a key factor that limits wider adoption of ML in healthcare. This is because healthcare workers often find it challenging to trust complex ML models because the models are often designed and rigorously evaluated only on specific diseases in a narrow environment and depend on one's technical knowledge of statistics and ML. Obermeyer, Powers, Vogeli, and Mullainathan (2019) find the evidence of racial bias in US healthcare algorithms that guide health decisions. Bias of prediction model can undermine the trust of the healthcare experts and other end-users of such models. Furthermore, applying those models on the larger systems may not perform well because of the complexity of the data and diversity of patients and diagnoses. Moreover, most of the models focus on accuracy prediction and rarely explain their predictions in a meaningful way (Ahmad, Eckert, Teredesai, & McKelvey, 2018; Elshawi, Al-Mallah, & Sakr, 2019). This is especially problematic in healthcare applications, where achieving high predictive accuracy is often as important as understanding the prediction. Interpretable ML has thus emerged as an area of research that aims to design transparent and explainable models and develop means to transform black-box methods into white-box methods whose predictions are accurate and can be interpreted meaningfully.

Interpretable ML is fundamentally complex and represents a rapidly developing field of research (Gilpin et al., 2019; Hall & Gill, 2018). Several terms can be found describing the interpretability and related concepts, such as the multiplicity of good models or model locality (Breiman, 2001; Hall & Gill, 2018), comprehensibility, or understandability (Piltaver, Luštrek, Gams, & Martinčić-Ipšić, 2016), and mental fit or explanatory ML (Bibal & Frenay, 2016).

Interpretability is frequently defined as a degree to which a human can understand the cause of a decision from an ML model (Miller, 2019). Gilpin et al. (2019) describe the main goal of interpretability as the ability to describe system internals to humans in an understandable way. Many terms are related to interpretability of ML models but refer to different concepts. Interpretability has a subjective nature, thus there is no consensus about the definition of how to measure or define it. An interpretable model is one that can be understood and is related to accuracy, explainability, and efficiency. On the other hand, efficiency must be taken into account where time needed for the user to grasp the model is of high importance (Bibal & Frenay, 2016). The term explainable artificial intelligence was described by van Lent, Fisher, and Mancuso (2004) to explain the behavior of artificial intelligence-controlled entities in game applications. The aim of explainable artificial intelligence is to ensure easily understandable reasoning for the end-user through the artificial intelligence knowledge and inference. Moreover, interpretability is often mentioned as a key area in the new wave of ML research. In this case, it is used to ensure the confidence in the process of development of ML systems for the end-users of such systems (Carvalho et al., 2019). The interpretation of ML models is often governed by various motivations, such as interpretability requirements, high-stakes decisions impact, societal concerns, and ML desiderata, regulations, and so on (Carvalho et al., 2019). Our operational definition of interpretable models in healthcare is end-user focused, more precisely, a model is interpretable if it can be evaluated by the end-user, where the reasoning behind the prediction is explained giving the end-users reasons to accept or reject predictions and recommendations (Ahmad, Eckert, Teredesai, & McKelvey, 2018). Based on this operational definition basic examples of interpretable models include short decision trees, linear and logistic regression with more examples presented in the next section, non-interpretable or black-box models usually include complex models with a focus on outcomes, for example, ensembles of classifiers or deep neural networks.

The European General Data Protection Regulation (GDPR) policy on the rights of citizens states that researchers must explain algorithmic decisions that have been made in ML models (Greene, Shmueli, Ray, & Fell, 2019; Wachter, Mittelstadt, & Floridi, 2017; Wallace & Castro, 2018). Consequently, this means that there must be a possibility to make the results re-traceable or in other words, the end-user has the right to explanation of all decisions made by the computer. On the other hand, one can also find literature questioning the importance of the model interpretability. For

example, in a recent review paper on the state of AI in healthcare Wang and Preininger (2019) quote Geoff Hinton, one of the pioneers in the field of AI, saying: "Policymakers should not insist on ensuring people understand exactly how a given AI algorithm works, because people can't explain how they work, for most of the things they do." On the other hand, Rudin (2019) suggests the following: "People have hoped that creating methods for explaining these black-box models will alleviate some of these problems, but trying to explain black-box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practices and can potentially cause catastrophic harm to society." To solve this problem Rudin suggests using inherently interpretable models and provides several examples of replacing the black-box models with interpretable models like sparse scoring systems (Ustun & Rudin, 2017) that are particularly suitable for the healthcare domain. The above-mentioned arguments have been the source of debates at the recent conferences and in the literature. Some interesting opinions on positive and negative aspects can also be found in a recent paper by Jia, Ren, and Cai (2020) that presents the experiences of some respected researchers working in the ML field. Healthcare professionals are often overwhelmed with the number of patients, the amount of data, and associated tasks. Thus, providing ML models that are easy to interpret and are explainable is necessary (Ahmad, Eckert, Teredesai, & McKelvey, 2018). Moreover, interpretable ML models are needed to support healthcare workers in decision making and to ensure quality and better patients' outcomes (Carvalho et al., 2019). ML techniques are effective in identifying risk factors and various diseases. Predictive models allow healthcare professionals in healthcare resources allocation, better care, and patient outcomes. Although complicated models may have high accuracy, they are not easily explainable (e.g., deep neural networks; Michalopoulos et al., 2020). On the other hand, Vellido (2019) states that improving model interpretability is very important in the field of healthcare for the adoption in practice. However, there is a need for integrating healthcare experts in the design of data analysis and interpretation. However, there are also negative aspects in ML models that we need to take into account, including dataset shift, accidental confounders fitting, unintended discriminatory bias, the challenges of generalization of findings, and the unintended negative consequences on health outcomes (Kelly, Karthikesalingam, Suleyman, Corrado, & King, 2019). Although deep learning is an advanced ML technique, deep learning algorithms are often considered as black boxes with no clear interpretation. Ren (2020) explains that AI models could address this problem. However, this should not be a requirement for the implementation of AI technologies, because AI can be developed as tool with human approval and oversight and can be made both effective and safe through training and verification. The authors also state that patients often find effectiveness more important than interpretability (Jia et al., 2020).

This article outlines the basic taxonomy of interpretability approaches in ML-based prediction models in healthcare. Therefore, it aims to provide simple definitions that can be used by readers beginning to use ML methods as well as healthcare professionals to get better understanding of the fast-emerging field.

It must be noted that approaches focusing on the interpretability of ML models built on nonstructured data such as different types of medical images, text, or other signal-based data are not included in this study. To obtain more information on a more general view of interpretability and explainable AI in healthcare, the reader is referred to works by Holzinger, Langs, Denk, Zatloukal, and Müller (2019) or London (2019). Recently, a broader concept of responsible ML in healthcare was presented (Wiens et al., 2019), which proposes that an interdisciplinary team of engaged stakeholders (policy makers, health system leaders, and individual researchers) systematically progress from problem formulation to widespread deployment, where also the level of required transparency and thus interpretability is addressed.

2 | CATEGORIZATION OF ML MODELS INTERPRETABILITY

There are different criteria for classifying methods for ML interpretability such as intrinsic or post hoc classification, premodel, in-model, or postmodel and classification based on the model outcome (Carvalho et al., 2019; Molnar, 2020). Intrinsic interpretability refers to a process of selecting and training a ML model that is intrinsically interpretable due to its simple structure (e.g., simple decision tree or regression model). However, it needs to be noted that both, decision trees and regression models offer only a limited interpretability, especially in case of capturing the nonlinearity in data. Post hoc (postmodel) interpretability usually describes application of interpretability methods after the training of the model (Molnar, 2020). Interpretability methods can also be classified based on the time of building the ML model. Premodel methods are independent of the model and can be used before the decision on which model will be used is taken. These methods are, for example, descriptive statistics, data visualization, Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and clustering methods. Even though Molnar (2020) classifies PCA, t-SNE, and clustering methods under interpretable methods it needs to be noted that interpretability of attributes

transformed using PCA, embeddings or clusters cannot offer comprehensible medical interpretation, but can be used in visualization of the results where patterns of interest from the interpretability point of view might be seen. So-called in-model prediction models have inherent interpretability integrated into the model itself. Postmodel interpretability refers to improving interpretability after building a model (post hoc; Carvalho et al., 2019). Moreover, Murdoch et al. (2018) define post hoc interpretability methods as methods that "provide insight into what a trained model has learned, without changing the underlying model".

Interpretability methods can also be classified by the results of the interpretability model. Examples of such approaches include the following methods (Molnar, 2020):

- Feature summary statistics which refer to summary of each feature statistic that affects the model predictions.
- Feature summary visualization describing the methods where the results are only meaningful if we present them using visualization techniques, as they are difficult or impossible to interpret when presented in the form of a table.
- Model internals approach is model specific and refers to the interpretation of intrinsically interpretable models. Weights of the linear regression models or features and thresholds used for the splits in the decision trees represent examples of such approaches.
- The data point interpretability is based on data instances and often uses similar instances to an instance for which we are trying to obtain a prediction model-based interpretation. Such approach to local interpretability is useful in image and text classification, but less in high-dimensional tabular data.
- Intrinsically interpretable models aim to interpret black box models by approximating them with an interpretable model (see next section for more details).

Another approach to interpretability of ML models categorizes the interpretability into global and local. Traditionally, the focus in ML research was on global interpretability in an effort to help users understand the relationship between all possible inputs to an ML model and the space of all predictions made by the model (Bratko, 1997; Martens, Huysmans, Setiono, Vanthienen, & Baesens, 2008). In contrast, local interpretation helps users to understand a prediction for a specific individual or about a small, specific region of the trained prediction function (Hall, Gill, Kurka, & Phan, 2019; Stiglic, Mertik, Podgorelec, & Kokol, 2006). Both types of approaches have been successfully used in different healthcare domains and are still being advanced in parallel with novel methodological approaches in development of ML-based prediction models.

In the clinical setting ML prediction models, we are frequently interested in different evaluation metrics as in other fields where estimation of area under the ROC curve (AUC) is often enough to assess the performance of the prediction model. Use of sensitivity or positive predictive value might be of higher importance in some applications of prediction models in clinical environment (Simon, Shortreed, & Coley, 2019; Steyerberg, 2019) and can also influence the interpretability of the results from prediction models. We propose a categorization of interpretability that is simple and can be used by experts and nonexperts in the field. Therefore, we categorize interpretability approaches in two nonexclusive groups—that is, model-specific or model-agnostic and local or global.

3 | MODEL-SPECIFIC OR MODEL-AGNOSTIC INTERPRETABILITY

Model-agnostic and model-specific models are used to either interpret the decision directly from the model (e.g., extracting all rules from the decision tree) or use some specific technique like knowledge distillation (Hinton, Vinyals, & Dean, 2015) to build a simple model that can be interpreted (e.g., a decision tree, set of rules, or a regression function). Model-specific interpretation methods are limited to specific models and derive explanations by examining internal model parameters (Du, Liu, & Hu, 2019). Model-agnostic methods can be used on any ML model and are usually applied post hoc (Molnar, 2020), where internal model parameters are not inspected as the model is treated as a black box (Du et al., 2019), thus creating a distinction with model-specific methods where this is the case. Typically, to achieve model-agnostic interpretability, one can use a surrogate or a simple proxy model to learn a locally faithful approximation of a complex, black-box model based on outputs returned by the black-box model. One of the earliest approaches of approximating the black-box model using a simple interpretable model can be found in a study where Craven and Shavlik (1994) extracted rules from the neural networks. Another similar approach was introduced in a paper on model compression by Bucilă, Caruana, and Niculescu-Mizil (2006) and was recently revived and updated under the name of knowledge distillation (Hinton et al., 2015).

Here we will briefly mention another specific interpretability method, both model-specific and model-agnostic, Graph Neural Network Explainer (GNNExplainer; Ying, Bourgeois, You, Zitnik, & Leskovec, 2019), which uses graph representation and requires a special ML framework for the complexity of data representation, such as the current state-of-the-art Graph Neural Network (GNN), which is experiencing a surge of interest (Xu, Hu, Leskovec, & Jegelka, 2018).

The details are outside the purview of this article, the reader is referred to works by Hamilton and colleagues (Hamilton, Ying, & Leskovec, 2017), but since GNNExplainer provides both global and local interpretability it might prove useful for practitioners of GNN in the future, as it provides the ability to visualize relevant structures to interpretability and gives insights into errors of faulty GNNs (Ying et al., 2019).

4 | LOCAL OR GLOBAL INTERPRETABILITY

Methods can explain a single prediction (local interpretability) or the entire model behavior (global interpretability; Molnar, 2020). Local interpretation of the models can be achieved by designing justified model architectures that explains why a specific decision was made. It can also be achieved by providing similar examples of instances to the target instance. For example, by emphasizing specific characteristics of a patient that are similar to characteristics of a smaller group of patients but different in other patients. Local interpretation techniques were not so frequently used until recently, but in the last 10 years many novel techniques were introduced that allow at least feature importance estimation for prediction at the personalized level suitable for the models with no or weak interpretability (Lundberg & Lee, 2017; Ribeiro, Singh, & Guestrin, 2016; Ribeiro, Singh, & Guestrin, 2018). On the other hand, globally interpretable models offer transparency about what is going on inside a model on an abstract level (Du et al., 2019). In order to explain global output of the model one needs a trained model, knowledge about the algorithm and the data (Lipton, 2016). Some authors (Ahmad, Eckert, Teredesai, & McKelvey, 2018) also argue about the specific group of interpretability approaches called cohort-specific interpretability, where they focus on characteristics of population subgroups in relation to the predicted outcome. However, we can also classify such cases as either global if the subgroup is treated as the subpopulation or as local if single prediction interpretations for the subgroup are grouped together (Molnar, 2020). An example of such model is Model Understanding through Subspace Explanations (MUSE), where it uses the behavior of subspaces characterized by certain features of interest for explanation (Lakkaraju, Kamar, Caruana, & Leskovec, 2019).

5 | CATEGORIZATION OF INTERPRETABLE ML MODELS

Some examples of prediction models that are interpretable using global/local or model-specific/model-agnostic interpretability techniques are shown in Table 1. Historically, global interpretability was widely used to extract knowledge from the prediction models such as decision trees (difficulty of interpretability depends on depth and number of terminal nodes), linear and logistic regression models (prediction as a weighted sum or weighted sum transformed by the logistic function) or naive Bayes (models based on the independence, thus conditional probability can be interpreted and for each feature, the contribution can be assessed). Figure 1 presents four representative examples of interpretability approaches based on the examples given in Table 1.

Each of the four approaches to interpretability can be applied to real-world problems based on the characteristics of the environment where we would like to apply a specific approach. The global specific models are the most appropriate in cases where the user is interested in global interpretation and can afford to sacrifice some prediction performance to get a directly interpretable model.

One can overcome the weakness of lower prediction performance mentioned for the global model-specific interpretation approaches by employing more complex and usually more accurate models such as ensembles of classifiers or deep neural networks. These models can in most cases only be interpreted by model-agnostic approaches, where some exceptions exist, such as variable importance in Random Forests, but one has to be careful depending on the method used since bias could be possibly present making the ranking unreliable (Nebrini, 2019). In the case of global model-agnostic approaches where surrogate models are most often used as a summarization with a more interpretable approximation, for example, linear regression, decision tree, it can be expected that some critical elements of the more complex model are missing.

T A B L E 1   Examples of approaches to interpretability of prediction regression models

	Global	Local
Model-specific	- Decision trees (depends on depth and number of terminal nodes; Hastie, Tibshirani, & Friedman, 2009; Stiglic, Kocbek, Pernek, & Kokol, 2012), - Linear and logistic regression models (Harrell Jr, 2015), - Generalized linear models (GLM) and generalized additive models (GAM; Hastie et al., 2009), - Naive Bayes classifier (Kononenko, 1993), - GNNExplainer (Ying et al., 2019)	- Set of rules (for specific individual; Visweswaran, Ferreira, Ribeiro, Oliveira, & Cooper, 2015), - Decision trees (by tree-decomposition; Visweswaran et al., 2015), - Visual analytics-based approaches (interactive visualization techniques for interpretation focusing on individual prediction), - k-Nearest neighbors (k-NN; depends on the number of important features, retrieving k-nearest neighbors for interpretation; Yuwono et al., 2015), - GNNExplainer (Ying et al., 2019)
Model-agnostic	- Different variants of model compression/knowledge distillation/global surrogate models (Elshawi, Al-Mallah, & Sakr, 2019), - Partial Dependence Plots (PDP; Elshawi, Al-Mallah, & Sakr, 2019), - Individual Conditional Expectation (ICE) plots (Elshawi, Al-Mallah, & Sakr, 2019), - Black Box Explanations through Transparent Approximations (BETA; Lakkaraju, Kamar, Caruana, & Leskovec, 2017) - Model understanding through subspace explanations (MUSE; Lakkaraju et al., 2019).	- Local interpretable model-agnostic explanations (LIME; Ribeiro et al., 2016), - Shapley additive explanations (SHAP; Lundberg & Lee, 2017), - Anchors (Ribeiro et al., 2018), - Attention map visualization, - Model understanding through subspace explanations (MUSE; Lakkaraju et al., 2019).

On the other side, there are specific cases where interpretation of the models will demand local interpretation. One of the strengths in local model-specific approaches is their ability to present the reasons for a decision in a very comprehensible and intuitive way (a small number of rules, part of a decision tree, or a few examples representing the nearest neighbors of the example of interest). Again, the weakness of such approach is in the prediction performance which is usually lower than in the case of model-agnostic approaches often employing more complex models. Unfortunately, most of the approaches in this group only allow very limited, usually variable importance based, reasoning about the outcome for an example of interest.

6 | APPLICATIONS OF INTERPRETABLE ML IN HEALTHCARE

In this section, we provide examples of real-world studies that employed at least one interpretability approach mentioned in the previous section to demonstrate their use in different fields of healthcare.

Model-specific approaches focused on global interpretability of ML-based models in healthcare have been in use for more than two decades. Due to their high level of interpretability and simple use in practice, the approaches like linear regression or naive Bayes models are still used in different fields of healthcare like urology (Otunaiya & Muhammad, 2019; Zhang, Ren, Ma, & Ding, 2019), toxicology (Zhang et al., 2019; Zhang, Ma, Liu, Ren, & Ding, 2018), endocrinology (Alaoui, Aksasse, & Farhaoui, 2019), neurology (Zhang & Ma, 2019), cardiology (Doshi-Velez & Kim, 2018; Feeny et al., 2019; Salmam, 2019), or psychiatry (Guimarães, Araujo, Araujo, Batista, & de Campos Souza, 2019; Obeid et al., 2019). However, even linear regression or naive Bayes models are only interpretable to some extent as it is difficult to interpret the results of such models in case of nonlinearity or nonhomogeneous attributes. Model-specific approaches focusing on local interpretation that can be based, for example, on k-NN or decision trees were recently used for interpretation in prediction of health-related conditions, including occupational diseases (Di Noia, Martino, Montanari, & Rizzi, 2020) or knee osteoarthritis (Jamshidi, Pelletier, & Martel-Pelletier, 2019), cancer (e.g., breast cancer or prostate cancer; Aro, Akande, Jibrin, & Jauro, 2019; Seker, Odetayo, Petrovic, Naguib, & Hamdy, 2000), severity of a disease, including chronical diseases (e.g., diabetes or Alzheimer's disease; Bucholc et al., 2019; Karun, Raj, & Attigeri, 2019), and mortality rates (e.g., myocardial infarction or perinatal stroke; Gao et al., 2020; Prabhakararao &
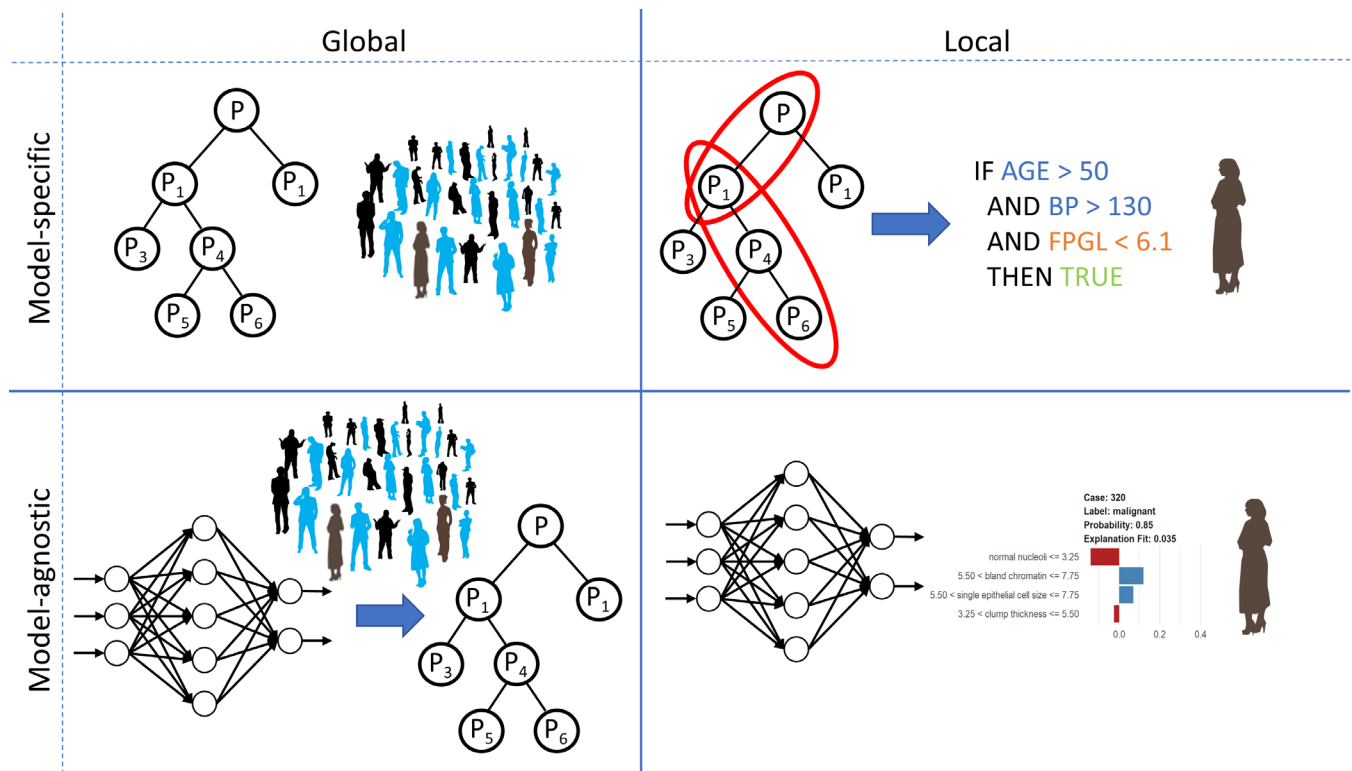
FIGURE 1 Visual representation of interpretability approaches for machine learning-based predictive modeling in healthcare

Dandapat, 2019). Local and model-agnostic interpretability can be used in interpretability of complex models, such as deep learning models. For example, SHAP was used in interpretation of predictions for the prevention of hypoxaemia during surgery (Lundberg et al., 2018), which increased the anaesthesiologists anticipation of hypoxemia events by 15%.

As already mentioned in the previous section, some recent approaches to interpretability cannot be simply classified in one of the four groups displayed in Figure 1. Model interpretability focusing on feature subspaces defined by the domain experts was proposed by Lakkaraju et al. (2019). MUSE was used to help in explaining decisions from the three-level neural network trained on depression diagnosis dataset. It generates sets of if-then rules that describe the model decisions on a global level, but it also provides a separate set of rules for a subspace based on the features selected by a healthcare expert working with patients. More specifically, authors demonstrate the effectiveness of MUSE to produce rules optimized for fidelity, unambiguity, and interpretability for a subspace focusing on excursive and smoking as those two features might be chosen by the expert as actionable features. A typical global interpretability approach would not consider the fact that there are features in the datasets that are more interesting than others as it is possible for the patient and healthcare expert to influence their values by introducing some interventions. As such, MUSE could be classified as a global model-agnostic interpretability approach, but it also demonstrates characteristics of approaches focusing on personalized interpretation by narrowing the subspace of search by end-user input.

In addition to technical challenges related to the development of interpretable models, we also need to address a myriad of ethical, legal, and regulatory challenges, for example, the GDPRs right to explanation (Wachter et al., 2017; Wallace & Castro, 2018). The latter type of interpretability is usually represented as a user-centric approach that enables users to find an appropriate model for a specific problem. For example, when the interpretation of general predictions at the population level is required, then global and model-specific models might be an appropriate choice. However, most practical applications of prediction models in healthcare are focused on the individual and would therefore require local model-specific interpretability approaches to allow the highest possible level of interpretability. On the other hand, most of the local model-specific prediction models (e.g., personalized rule sets or k-NN) cannot achieve the level of prediction performance of the more advanced models where model-specific interpretability is not possible (e.g., most deep learning approaches).

7 | DISCUSSION

This article provides a current and practical overview of interpretability methods for prediction models in healthcare, with a focus on usability from an end-user perspective. In contrast to many other fields, decisions in healthcare are high-stake decisions as they can directly influence a treatment outcome or even survival of a patient. From the technical perspective, this article focuses on structured data where a lot of emphasis is on feature (variable) importance. In contrast to structured data, the interpretability represents an important research topic in medical computer vision (CV) and natural language processing (NLP) as well (Lei, 2017). In CV, the focus is on identification of the image sections that are responsible for successful recognition of critical, health condition-specific regions in the image that might be relevant for medical experts (Escalante et al., 2018; Hosny, Parmar, Quackenbush, Schwartz, & Aerts, 2018; Mazurowski, Buda, Saha, & Bashir, 2019; Razzak, Naz, & Zaib, 2018). Although this type of approaches to interpretability is specific, they could be adapted to unstructured data by transforming the data to a format similar to images. In the NLP, the focus in interpretability is on marking the sections of text representing the content of interest. An example of such approach is classification of documents where the ML-based interpretability approaches are used to mark sections of textual documents that explain why they were categorized to a specific group (Arras, Horn, Montavon, Müller, & Samek, 2017). In case of healthcare applications, the textual parts of electronic healthcare records could be used to obtain more information on a patient by extracting marked information from the clinical text.

In recent years, a shift in the research of interpretability methods can be observed from model-specific and global interpretable models to model-agnostic and local interpretable models, where one of the reasons is the availability of massive datasets in healthcare (Esteva et al., 2019). It needs to be noted that this overview does not provide an in-depth

---

BOX 1 Visual analytics and interpretability

In recent years, researchers are increasingly relying on visual analytics (VA) techniques to support interpretability of ML models within healthcare fields (Simpao, Ahumada, Gálvez, & Rehman, 2014). Visual analytics extends the interaction paradigm of traditional information visualization techniques to support both the interpretation of ML models and model steering with feedback from end-users (Endert et al., 2017). Vellido (2019) argues that medical experts need to be integrated into the design of these data analysis interpretation strategies, so as to become a part of routine clinical and health care practice. The idea can be traced back to Kovalerchuk, Vityaev, and Ruiz (2001).

Both model-specific and model-agnostic approaches have been researched in VA prototypes. RetainVis (Kwon et al., 2018) is a prominent example of a model-specific VA approach that visualizes a recurrent neural network to support interpretation and diagnosis of the model. RetainVis uses t-SNE (Maaten & Hinton, 2008) to project patients on a 2D space and explains the model's interpretation of data by showing which patients are closely located in the latent space. The approach has been used to predict future diagnosis of heart failure and cataract. Although the approach is promising, van der Maaten (2018) also points out limitations, as t-SNE will not help assign meaning to point densities in clusters.

Most other approaches are model-agnostic. RuleMatrix (Ming, Qu, & Bertini, 2018) is a prominent example that relies on extracted rule-based knowledge from the input–output behavior of a model. The system helps domain experts understand and inspect classification models using rule-based explanations and was used to improve cancer and diabetes classification. Prospector (Krause, Perer, & Bertini, 2016) is a second model-agnostic example that leverages the concept of partial dependencies that communicate how individual features or multiple input variables affect the prediction. The system also supports instance-level explanations, enabling users to understand why certain data results in a specific prediction. Zintgraf, Cohen, Adel, and Welling (2017) present a method that visualizes which pixels of an input image are evidence for or evidence against a node in a deep neural network. Moreover, Li, Fujiwara, Choi, Kim, and Ma (2020) focus on visualizing features to explain the reasoning of a model. The approach is used to explain and compare different models for clinical data beyond their accuracy scores. The VA system visualizes an overview of the similarities of local feature contributions of different models using t-SNE. The system also incorporates a model summary view to analyze the differences in consistencies of the features across models.

analysis of some topics that are still open (Ahmad, Eckert, Teredesai, & McKelvey, 2018), such as simplification of complex models for explanations that may result in suboptimal results, thus auditing output for fairness and bias might be a better solution (Tomasello, 1999), scalability of interpretable ML models (e.g., LIME or SHAP values can be computationally expensive) and methods for evaluation of explanation models (e.g., different models with similar error rate but different explanations). Another current research topic related to interpretability of ML models in healthcare is causality (Holzinger et al., 2019), which is defined as causal understanding with effectiveness, efficiency, and satisfaction in a specified context of use (Box 1).

8 | CONCLUSION

To trust, maintain fairness and transparency of a specific model and its predictions, it is important that we understand different approaches to model interpretability. Interpretability can be categorized in terms of model-specific and model-agnostic or global and local interpretability. There are various new techniques and approaches available for interpretable ML. However, the key challenges are still unsolved, and future research is needed to find new and reasonable solutions for the progress in this field. In many cases, the challenges of interpreting the results of prediction models are related to more and more complex prediction models. For example, recent solutions incorporating the medical knowledge in the form of knowledge graphs represent a specific problem also in the interpretability and representation of the results. Some speculations on possible improvements that would allow more effective interpretability solutions by combining approaches presented in this article are presented in the remainder of this section.

In the future, we expect more approaches like MUSE (Lakkaraju et al., 2019) where a global interpretability approach is supplemented by specific interpretations that can explain predictions either for a single individual or a smaller group of individuals. Although it is difficult to predict the exact direction of future research in this field, it is certain that interpretability techniques represent an important concept that needs to be taken into account when developing prediction models for healthcare. Another avenue of research might lie with GNN, which combines node feature information and graph structure by using neural networks (Ying et al., 2019) and the aforementioned GNN Explainer as a tool for post hoc explanation of GNN. However, there is a limited understanding of GNN properties and limitations (Xu et al., 2018). Moreover, it is necessary to understand and additionally improve the generalization properties of GNNs. Future research should focus on developing algorithmic solutions that enable ML driven decision-making for various healthcare problems that influence disease course and outcome.

CONFLICT OF INTEREST
The authors declare that there is no conflict of interest.

AUTHOR CONTRIBUTIONS
Gregor Stiglic: Writing-original draft; writing-review and editing. **Primoz Kocbek:** Writing-original draft; writing-review and editing. **Nino Fijacko:** Writing-original draft; writing-review and editing. **Katrien Verbert:** Writing-original draft; writing-review and editing. **Marinka Zitnik:** Writing-original draft; writing-review and editing. **Leona Cilar:** Writing-original draft; writing-review and editing.

ORCID
Gregor Stiglic https://orcid.org/0000-0002-0183-8679

RELATED WIREs ARTICLE
Causability and explainability of artificial intelligence in medicine

FURTHER READING
Craven, M. W., & Shavlik, J. W. (1994). Using sampling and queries to extract rules from trained neural networks. In *Machine learning proceedings 1994* (pp. 37–45). Morgan Kaufmann.
Elshawi, R., Sherif, Y., Al-Mallah, M., & Sakr, S. (2019). *Interpretability in HealthCare a comparative study of local machine learning interpretability techniques.* In 2019 IEEE 32nd International Symposium on Computer-based Medical Systems (CBMS). pp. 275–280.

Katuwal, G. J., & Chen, R. (2016). Machine learning model interpretability for precision medicine. *arXiv preprint*, *arXiv:1610.09045*.

Kwon, B. C., Choi, M. J., Kim, J. T., Choi, E., Kim, Y. B., Kwon, S., ... & Choo, J. (2018). Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *IEEE transactions on visualization and computer graphics*, *25*(1), 299–309.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell 1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x

REFERENCES

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: Survey on explainable artificial intelligence (XAI). In *IEEE access* (pp. 52138–52160). New York, NY: IEEE.

Ahmad, A. M., Eckert, C., Teredesai, A., & McKelvey, G. (2018). Interpretable machine learning in healthcare. In *IEEE intelligent informatics bulletin* (pp. 1–7). New York, NY: IEEE.

Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). *Interpretable machine learning in healthcare*. In Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. pp. 559–560.

Alaoui, S. S., Aksasse, B., & Farhaoui, Y. (2019). *Data mining and machine learning approaches and Technologies for Diagnosing Diabetes in women*. In International Conference on Big Data and Networks Technologies. Springer, Cham. pp. 59–72.

Aro, T. O., Akande, H. B., Jibrin, M. B., & Jauro, U. A. (2019). Homogenous ensembles on data mining techniques for breast cancer diagnosis. *Daffodil International University Journal of Science and Technology*, *14*(1), 9–12.

Arras, L., Horn, F., Montavon, G., Müller, K. R., & Samek, W. (2017). "What is relevant in a text document?": An interpretable machine learning approach. *PLoS One*, *12*(8), e0181142.

Bibal, A., & Frenay, B. (2016). *Interpretability of machine learning models and representations: An Introduction*. In 24th European symposium on artificial neural networks, computational intelligence and machine learning, Bruges. pp. 77–82.

Bratko, I. (1997). Machine learning: Between accuracy and interpretability. In G. Della Riccia, H.-J. Lenz, & R. Kruse (Eds.), *Learning, networks and statistics* (pp. 163–177). Vienna: Springer.

Breiman, L. (2001). Statistical modelling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231.

Bucholc, M., Ding, X., Wang, H., Glass, D. H., Wang, H., Prasad, G., ... KongFatt, W. (2019). A practical computerized decision support system for predicting the severity of Alzheimer's disease of an individual. *Expert Systems with Applications*, *130*, 157–171.

Bucilǎ, C., Caruana, R., & Niculescu-Mizil, A. (2006). *Model compression*. In KDD '06 Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY. pp. 535–541.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8* (832), 1–34.

Di Noia, A., Martino, A., Montanari, P., & Rizzi, A. (2020). Supervised machine learning techniques and genetic optimization for occupational diseases risk prediction. *Soft Computing*, *24*(6), 4393–4406.

Doshi-Velez, F., & Kim, B. (2018). Considerations for evaluation and generalization in interpretable machine learning. In H. Jair Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, & M. A. J. van Gerven (Eds.), *Explainable and interpretable models in computer vision and machine learning* (pp. 3–17). Cham: Springer.

Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, *63*(1), 68–77.

Elshawi, R., Al-Mallah, M., & Sakr, S. (2019). On the interpretability of machine learning based model for predicting hypertension. *BMC Medical Informatics and Decision Making*, *19*(1), 146.

Endert, A., Ribarsky, W., Turkay, C., Wong, B. W., Nabney, I., Blanco, I. D., & Rossi, F. (2017). The state of the art in integrating machine learning into visual analytics. *Computer Graphics Forum*, *36*(8), 458–486.

Escalante, H. J., Escalera, S., Guyon, I., Baró, X., Güçlütürk, Y., & Güçlü, U. M.. (2018). *Explainable and interpretable models in computer vision and machine learning*. Cham: Springer International Publishing.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, *25*, 24–29.

Feeny, A. K., Rickard, J., Patel, D., Toro, S., Trulock, K. M., Park, C. J., ... Gorodeski, E. Z. (2019). Machine learning prediction of response to cardiac resynchronization therapy: Improvement versus current guidelines. *Circulation. Arrhythmia and Electrophysiology*, *12*(7), e007316.

Gao, Y., Long, Y., Guan, Y., Basu, A., Baggaley, J., & Plötz, T. (2020). Automated general movement assessment for perinatal stroke screening in infants. In F. Chen, R. I. García-Betances, L. Chen, M. F. Cabrera, & C. D. Nugent (Eds.), *Smart assisted living* (pp. 167–187). Cham: Springer.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2019). *Explaining explanations: An overview of interpretability of machine learning*. In Fifth International Conference on Data Science and Advanced Analytics (DSAA). New York, NY: IEEE. pp. 80–89.

Greene, T., Shmueli, G., Ray, S., & Fell, J. (2019). Adjusting to the GDPR: The impact on data scientists and behavioral researchers. *Big Data*, *7*(3), 140–162.

Guimarães, A. J., Araujo, V. J. S., Araujo, V. S., Batista, L. O., & de Campos Souza, P. V. (2019, May). *A hybrid model based on fuzzy rules to act on the diagnosed of autism in adults*. In IFIP International Conference on Artificial Intelligence Applications and Innovations. Cham: Springer. pp. 401–412.

Hall, P., & Gill, N. (2018). *An Introduction to machine learning interpretability: An applied perspective on fairness, accountability, transparency, and explainable AI*. Boston, MA: O'Reilly.

Hall, P., Gill, N., Kurka, M., & Phan, W. (2019). *Machine learning interpretability with H₂O driverless AI*. Mountain View, CA: H2O.ai, Inc.

Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems* (pp. 1024–1034). Cambridge, MA: MIT Press.

Harrell, F. E., Jr. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis*. Cham: Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Berlin: Springer.

Hinton, G., Vinyals, O., & Dean, J. (2015). *Distilling the knowledge in a neural network*. NIPS Deep Learning and Representation Learning Workshop.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainabilty of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, *9*, e1312.

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*, *18*(8), 500–510.

Jamshidi, A., Pelletier, J. P., & Martel-Pelletier, J. (2019). Machine-learning-based patient-specific prediction models for knee osteoarthritis. *Nature Reviews Rheumatology*, *15*(1), 49–60.

Jia, X., Ren, L., & Cai, J. (2020). Clinical implementation of AI technologies will require interpretable AI models. *Medical Physics*, *47*(1), 1–4.

Karun, S., Raj, A., & Attigeri, G. (2019). Comparative Analysis of Prediction Algorithms for Diabetes. In S. K. Bhatia, S. Tiwari, K. K. Mishra, & M. C. Trivedi (Eds.), *Advances in computer communication and computational sciences* (pp. 177–187). Singapore: Springer.

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, *17*, 195.

Kononenko, I. (1993). Inductive and Bayesian learning in medical diagnosis. *Applied Artificial Intelligence an International Journal*, *7*(4), 317–337.

Kovalerchuk, B., Vityaev, E., & Ruiz, J. F. (2001). Consistent and complete data and "expert" mining in medicine. *Studies in Fuzziness and Soft Computing*, *60*, 238–281.

Krause, J., Perer, A., & Bertini, E. (2016). Using visual analytics to interpret predictive machine learning models. *arXiv preprint arXiv: 1606.05685*.

Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2017). Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154*.

Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2019). *Faithful and customizable explanations of black box models*. In AIES '19 Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY: ACM. pp. 131–138.

Lei, T. (2017). *Interpretable neural models for natural language processing (doctoral dissertation)*. Cambridge, MA: Massachusetts Institute of Technology.

Li, Y., Fujiwara, T., Choi, Y. K., Kim, K. K., & Ma, K. L. (2020). A visual analytics system for multi-model comparison on clinical data predictions. *arXiv preprint arXiv:2002.10998*.

Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

London, A. J. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus Explainability. *The Hastings Center Report*, *49*(1), 15–21.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., … Lee, S. I. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, *2*(10), 749–760.

Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Martens, D., Huysmans, J., Setiono, R., Vanthienen, J., & Baesens, B. (2008). Rule extraction from support vector machines: An overview of issues and application in credit scoring. In J. Diederich (Ed.), *Rule extraction from support vector machines* (pp. 33–63). Berlin, Heidelberg: Springer.

Mazurowski, M. A., Buda, M., Saha, A., & Bashir, M. R. (2019). Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. *Journal of Magnetic Resonance Imaging*, *49*(4), 939–954.

Michalopoulos, G., Chen, H., Yang, Y., Subendran, S., Quinn, R., Oliver, M., … Wong, A. (2020). Why do I trust your model? Building and explaining. Predictive models for peritoneal dialysis eligibility. *Journal of Computational Vision and Imaging Systems*, *5*(1), 1.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38.

Ming, Y., Qu, H., & Bertini, E. (2018). Rulematrix: Visualizing and understanding classifiers with rules. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 342–352.

Molnar, C. (2020). *Interpretable machine learning: A guide for making black box models explainable*. Victoria, BC, Canada: Leanpub.com.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2018). Interpretable machine learning: Definitions, methods, and applications. *Proceedings of the National Academy of Sciences*, *116*(44), 22071–22080.

Nebrini, S. (2019). Bias in the intervention in prediction measure in random forests: Illustrations and recommendations. *Bioinformatics*, *35*(13), 2343–2345.

Obeid, J. S., Weeda, E. R., Matuskowitz, A. J., Gagnon, K., Crawford, T., Carr, C. M., & Frey, L. J. (2019). Automated detection of altered mental status in emergency department clinical notes: A deep learning approach. *BMC Medical Informatics and Decision Making*, *19* (1), 164.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453.

Otunaiya, K. A., & Muhammad, G. (2019). Performance of Datamining techniques in the prediction of chronic kidney disease. *Computer Science and Information Technology*, *7*(2), 48–53.

Piltaver, R., Luštrek, M., Gams, M., & Martinčić-Ipšić, S. (2016). What makes classification trees comprehensible? *Expert Systems with Applications*, *62*, 333–346.

Prabhakararao, E., & Dandapat, S. (2019). A weighted SVM based approach for automatic detection of posterior myocardial infarction using VCG signals. In *2019 National Conference on Communications (NCC)*. New York, NY: IEEE. pp. 1–6.

Razzak, M. I., Naz, S., & Zaib, A. (2018). Deep learning for medical image processing: Overview, challenges and the future. In N. Dey, A. Ashour, & S. Borra (Eds.), *Classification in BioApps* (pp. 323–350). Cham: Springer.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *Model-agnostic interpretability of machine learning*. In Proceedings of the 2016 ICML workshop on human interpretability in machine learning (WHI 2016). pp. 91–95.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). *Anchors: High-precision model-agnostic explanations*. In Thirty-Second AAAI Conference on Artificial Intelligence.

Salmam, I. (2019). Heart attack mortality prediction: An application of machine learning methods. *Turkish Journal of Electrical Engineering and Computer Sciences*, *27*(6), 4378–4389.

Seker, H., Odetayo, M. O., Petrovic, D., Naguib, R., & Hamdy, F. (2000). A soft measurement technique for searching significant subsets of prostate cancer prognostic markers. In P. Sincak, J. Vascak, V. Kvasnicka, & R. Mesiar (Eds.), *The state of the art in computational intelligence* (pp. 325–328). Heidelberg: Physica.

Simon, G. E., Shortreed, S. M., & Coley, R. Y. (2019). Positive predictive values and potential success of suicide prediction models. *JAMA Psychiatry*, *76*(8), 868–869.

Simpao, A. F., Ahumada, L. M., Gálvez, J. A., & Rehman, M. A. (2014). A review of analytics and clinical informatics in health care. *Journal of Medical Systems*, *38*(4), 45.

Steyerberg, E. W. (2019). *Clinical prediction models*. Cham: Springer International Publishing.

Stiglic, G., Kocbek, S., Pernek, I., & Kokol, P. (2012). Comprehensive decision tree models in bioinformatics. *PLoS One*, *7*(3), e33812.

Stiglic, G., Mertik, M., Podgorelec, V., & Kokol, P. (2006). *Using visual interpretation of small ensembles in microarray analysis*. In 19th IEEE symposium on computer-based medical systems (CBMS'06). New York, NY: IEEE. pp. 691–695.

Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.

Ustun, B., & Rudin, C. (2017). *Optimized risk scores*. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining (KDD).

van der Maaten L. (2018). Dos and Don'ts of using t-SNE to Understand Vision Models, CVPR 2018 Tutorial on Interpretable Machine Learning for Computer Vision. Retrieved from http://deeplearning.csail.mit.edu/slide_cvpr2018/laurens_cvpr18tutorial.pdf.

van Lent, M., Fisher, W., & Mancuso, M. (2004). *An explainable artificial intelligence system for small-unit tactical behavior*. In Proceedings of the National Conference on Artificial Intelligence, San Jose, CA, 25–29 July 2004; AAAI Press: Menlo Park, CA; MIT Press: Cambridge, MA, pp. 900–907.

Vellido, A. (2019). The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*, 1–15.

Visweswaran, S., Ferreira, A., Ribeiro, G. A., Oliveira, A. C., & Cooper, G. F. (2015). Personalized modeling for prediction with decision-path models. *PLoS One*, *10*(6), e0131022.

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.

Wallace, N., & Castro, D. (2018, March 26). The impact of the EU's new data protection regulation on AI. Retrieved from http://www2.datainnovation.org/2018-impact-gdpr-ai.pdf.

Wang, F., & Preininger, A. (2019). AI in health: State of the art, challenges, and future directions. *Yearbook of Medical Informatics*, *28*(1), 16–26.

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., ... Goldenberg, A. (2019). Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine*, *25*(9), 1337–1340.

Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2018). How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNN explainer: A tool for post-hoc explanation of graph neural networks. *arXiv preprint arXiv:1903.03894*.

Yuwono, T., Setiawan, N. A., Nugroho, A., Persada, A. G., Prasojo, I., Dewi, S. K., & Rahmadi, R. (2015). Decision support system for heart disease diagnosing using K-NN algorithm. *Proceeding of the Electrical Engineering Computer Science and Informatics*, *2*(1), 160–164.

Zhang, H., Ma, J. X., Liu, C. T., Ren, J. X., & Ding, L. (2018). Development and evaluation of in silico prediction model for drug-induced respiratory toxicity by using naïve Bayes classifier method. *Food and Chemical Toxicology*, *121*, 593–603.

Zhang, H., Ren, J. X., Ma, J. X., & Ding, L. (2019). Development of an in silico prediction model for chemical-induced urinary tract toxicity by using naïve Bayes classifier. *Molecular Diversity*, *23*, 381–392.

Zhang, Y., & Ma, Y. (2019). Application of supervised machine learning algorithms in the classification of sagittal gait patterns of cerebral palsy children with spastic diplegia. *Computers in Biology and Medicine*, *106*, 33–39.

Zintgraf, L. M., Cohen, T. S., Adel, T., & Welling, M. (2017). Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*.