Briefings in Bioinformatics, 2023, **24(5)**, 1–22
https://doi.org/10.1093/bib/bbad236
Advance access publication date 21 July 2023
Problem Solving Protocol

Explainable AI for Bioinformatics: Methods, Tools and Applications

Md. Rezaul Karim, Tanhim Islam, Md. Shajalal, Oya Beyan, Christoph Lange, Michael Cochez, Dietrich Rebholz-Schuhmann and Stefan Decker

Corresponding author. Md. Rezaul Karim. E-mail: rezaul.karim@rwth-aachen.de

Abstract

Artificial intelligence (AI) systems utilizing deep neural networks and machine learning (ML) algorithms are widely used for solving critical problems in bioinformatics, biomedical informatics and precision medicine. However, complex ML models that are often perceived as opaque and *black-box* methods make it difficult to understand the reasoning behind their decisions. This lack of transparency can be a challenge for both end-users and decision-makers, as well as AI developers. In sensitive areas such as healthcare, explainability and accountability are not only desirable properties but also legally required for AI systems that can have a significant impact on human lives. Fairness is another growing concern, as algorithmic decisions should not show bias or discrimination towards certain groups or individuals based on sensitive attributes. Explainable AI (XAI) aims to overcome the opaqueness of black-box models and to provide transparency in how AI systems make decisions. Interpretable ML models can explain how they make predictions and identify factors that influence their outcomes. However, the majority of the state-of-the-art interpretable ML methods are domain-agnostic and have evolved from fields such as computer vision, automated reasoning or statistics, making direct application to bioinformatics problems challenging without customization and domain adaptation. In this paper, we discuss the importance of explainability and algorithmic transparency in the context of bioinformatics. We provide an overview of model-specific and model-agnostic interpretable ML methods and tools and outline their potential limitations. We discuss how existing interpretable ML methods can be customized and fit to bioinformatics research problems. Further, through case studies in bioimaging, cancer genomics and text mining, we demonstrate how XAI methods can improve transparency and decision fairness. Our review aims at providing valuable insights and serving as a starting point for researchers wanting to enhance explainability and decision transparency while solving bioinformatics problems. GitHub: https://github.com/rezacsedu/XAI-for-bioinformatics.

Keywords: bioinformatics, explainable AI, interpretable machine learning, deep learning, machine learning, NLP

Md. Rezaul Karim is a Senior Data Scientist at ALDI SÜD - Global Data & Analytics Services and Visiting Researcher at RWTH Aachen University Germany. Before joining ALDI SÜD, he worked as a Senior Data Scientist at Fraunhofer FIT and a Postdoctoral Researcher at RWTH Aachen University, Germany. Previously, he worked as a Machine Learning Engineer at Insight Centre for Data Analytics, University of Galway, Ireland. Before that, he worked as a Lead Software Engineer at Samsung Electronics, South Korea. He received his PhD from RWTH Aachen University, Germany; MSc. degree from Kyung Hee University, South Korea and BSc. degree from University of Dhaka, Bangladesh. His research interests include machine learning, NLP and explainable AI (XAI) with a focus on bioinformatics and healthcare.
Tanhim Islam is a Data Scientist at CONET, Germany. He received his MSc. degree in Data Science from RWTH Aachen University, Germany and BSc. degree in Computer Science and Engineering from BRAC University, Bangladesh. His research interests include data science, applied machine/deep learning, NLP and explainable AI.
Md. Shajalal is a Marie Skłodowska-Curie Research Fellow at Fraunhofer FIT, Germany and a PhD candidate at University of Siegen, Germany. Previously, he worked as an Assistant Professor of Computer Science at Hajee Mohammad Danesh Science and Technology University, Bangladesh. He received his BSc. in Computer Science from University of Chittagong, Bangladesh and MSc. degree in Computer Science from Toyohashi University of Technology, Japan. His research interests include explainable AI (XAI), NLP, information retrieval (IR) and applied machine/deep learning.
Oya Beyan is a Professor of Medical Informatics at the University of Cologne, Faculty of Medicine, and University Hospital Cologne. Prof. Beyan is the director of the Institute for Biomedical Informatics, which aims to support data-driven medicine and digital transformation in healthcare. Prof. Beyan and her research team focuses on health data reusability, semantic interoperability and data science towards continuous improvement of healthcare through innovation and creating new knowledge. Prof. Beyan co-leads the MEDIC data integration center at the Medical Faculty of Cologne, where multi-modal health data are integrated and reused for research. She is also affiliated with Fraunhofer FIT, where she leads a research and innovation group for FAIR data and distributed analytics.
Christoph Lange is the Head of Data Science and Artificial Intelligence department at Fraunhofer FIT, Germany, and a Senior Researcher at RWTH Aachen University. He received his PhD in Computer Science from Jacobs University Bremen, Germany (now: Constructor University). His research is broadly concerned with Knowledge Engineering and Data Infrastructures: choosing the right formalism to represent domain knowledge in an expressive and scalable way following Linked Data and FAIR principles, thus enabling and facilitating machine support with sharing, quality assessment, publishing and collaborative authoring.
Michael Cochez is an Assistant Professor at Vrije Universiteit Amsterdam, the Netherlands and Lab Manager of the Elsevier Discovery Lab, Amsterdam. He was formerly a Postdoctoral Researcher at Fraunhofer FIT, Germany. He received his PhD and MSc. degrees in Mathematical Information Technology from the University of Jyväskylä, Finland, and BSc. degree in Information Technology from the University of Antwerp, Belgium. His research interests include knowledge representation, machine learning and question answering on knowledge graphs.
Dietrich Rebholz-Schuhmann is a Professor of Medicine at the University of Cologne and the Scientific Director of the Information Center for Life Sciences, German National Library of Medicine (ZB MED), Germany. He was formerly the director of the Insight Centre for Data Analytics, and Professor of Informatics at the University of Galway, Ireland. Before that, he was the director of Healthcare IT at a Heidelberg, Germany-based bioscience company and group leader of Semantic Data Analytics at European Bioinformatics Institute (EMBL-EBI). His research interests include data science, semantics-driven data analytics, biomedical text mining and bioinformatics.
Stefan Decker is the Chair and Professor for Information Systems and Databases at RWTH Aachen University and managing director of Fraunhofer FIT, Germany. He was formerly the Director of the Insight Centre for Data Analytics, and Professor of Informatics at the University of Galway, Ireland. Even before that, he worked as Research Assistant Professor at the University of Southern California and as a Postdoctoral Researcher at Stanford University, USA. His research interests include Semantic Web and linked data and knowledge representation.
Received: February 10, 2023. **Revised:** May 10, 2023. **Accepted:** May 26, 2023

INTRODUCTION

Artificial intelligence (AI) systems that are built on machine learning (ML) and deep neural networks (DNNs) are increasingly deployed in numerous application domains such as military, cybersecurity, healthcare, etc. Further, ML and DNN models are applied to solving complex and emerging biomedical research problems: from text mining, drug discovery and single-cell RNA sequencing to early disease diagnosis and prognosis. Further, the paradigm of evidence-based precision medicine has evolved toward a more comprehensive analysis of disease phenotype and their induction through underlying molecular mechanisms and pathway regulation. Another common application of AI in precision medicine is predicting what treatment protocols are likely to succeed on a patient based on patient phenotype, demographics and treatment contexts [1].

Biomedical data science encompasses a range of data types, such as genome sequences, omics, imaging, clinical and structured/unstructured biomedical texts [2], where ML methods are typically used for the analysis and interpretation of multimodal data (e.g. multi-omics, imaging, clinical, medication, disease progression), in a multimodal data setting. Further, datasets, including bioimaging and omics are of increasing dimensionality. The surge of these massive amounts of data not only brings unprecedented progress in bioinformatics and opportunities to perform predictive modeling at scale [2], it also introduces challenges to existing AI methods and tools such as data heterogeneity, high-dimensionality and volume [3]. Principal component analysis (PCA) and isometric feature mapping (Isomap) are extensively used as dimensionality reduction techniques [4]. However, the representations learned by these methods often lose essential properties [5], making them less effective against a known phenomenon called *curse of dimensionality*, particularly for high-dimensional datasets [4].

A complex DNN model can effectively handle complex problems, thanks to its ability to extract features and to learn useful representations from high-dimensional datasets. For instance, autoencoders (AEs) are used for unsupervised learning tasks, where their multi-layered, non-linear architecture enables the learning of complex and higher-order feature interactions. By transforming input feature space into a lower-dimensional latent space, AEs capture important contextual information of the underlying data, which can be used for various downstream tasks. However, the latent factors learned by AEs are not easily interpretable. Disentangling them can provide insights into the features captured by the representations and the attributes of the samples that the tasks are based on [3].

AI has already surpassed human experts in specific areas, such as detecting tumors and analyzing disease progression. However, the widespread use of AI in healthcare is hindered by the lack of models capable of handling vast amounts of data. Although DNN models can address complex problems, their *black-box* nature raises concerns about transparency and accountability. With their increasing complexity, complex DNN models tend to be less interpretable and may end up as *black-box* methods. Although mathematical certainty means it should be possible to transparently show that no hidden logic is influencing the behavior of a model [6], predictions made by these models cannot be traced back, making it unclear how or why they arrived at a certain outcome. This lack of explainability can lead to issues of trust in the AI system and the inability to provide users with human-interpretable explanations for its decisions.

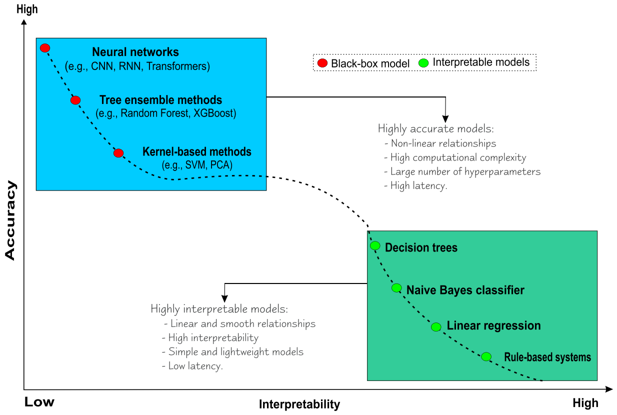

High
Accuracy
Neural networks
(e.g., CNN, RNN, Transformers)
Tree ensemble methods
(e.g., Random Forest, XGBoost)
Kernel-based methods
(e.g., SVM, PCA)
● Black-box model ● Interpretable models
Highly accurate models:
- Non-linear relationships
- High computational complexity
- Large number of hyperparameters
- High latency.
Decision trees
Naive Bayes classifier
Linear regression
Rule-based systems
Highly interpretable models:
- Linear and smooth relationships
- High interpretability
- Simple and lightweight models
- Low latency.
Low
Interpretability
High

Figure 1. Accuracy versus interpretability trade-off [12].

The field of explainable artificial intelligence (XAI) aims to make AI systems more transparent and understandable to humans [7] by interpreting how *black-box* models should make decisions [8]. XAI strives to enhance the human comprehensibility, transparency and accountability of AI systems [2]. An interpretable ML model can reveal the factors that impact (e.g. statistically significant features) its outcomes and explain the interactions among [6]. Since how a prediction is made should be as transparent as possible in a faithful and interpretable manner, model-specific and model-agnostic approaches have emerged, covering local and global interpretability [9]. While local explanations focus on explaining individual predictions, global explanations explain entire model behavior in the form of plots or more interpretable approximations such as decision sets or lists.

Linear models, decision trees (DTs) and rule-based systems are less complex and inherently interpretable. However, they are less accurate compared to tree-based ensembles like random forests (RFs) and DNNs. This phenomenon, which is shown in Figure 1, is called the *accuracy vs. interpretability trade-off*. XAI has recently gained widespread interest from both academia and industry [10]. This has led to the development of numerous model-specific and model-agnostic interpretable ML methods on which an XAI system can be built in order to enhance its local and global interpretability [9]. Despite the advancements in the development of interpretable ML methods to explain the decisions of *black-box* models in recent years, many of those methods are seldom used beyond visualization [11]. Many of these methods are not designed in a domain-agnostic manner hence may not be suitable for every bioinformatics problem and data type [2]. Therefore, they require customization or extension to fit the specific needs of bioinformatics and diverse data types [2].

The majority of existing reviews of XAI that focus on interpretable ML methods cover general challenges in biomedical data science, thereby not emphasizing concrete bioinformatics research problems. In this paper, we outline the importance of XAI in bioinformatics by taking into consideration transparency, fairness, reasoning, causal inference, contrastive explanations and human-level interpretability. We comprehensively reviewed and discussed various interpretable ML techniques with mathematical details. We provide a brief overview of interpretable ML tools and libraries for diverse data types (e.g. from tabular data, texts and knowledge graphs (KGs) to bioimaging). Besides, we show, through several case studies in bioimaging, cancer genomics and biomedical text mining, how bioinformatics

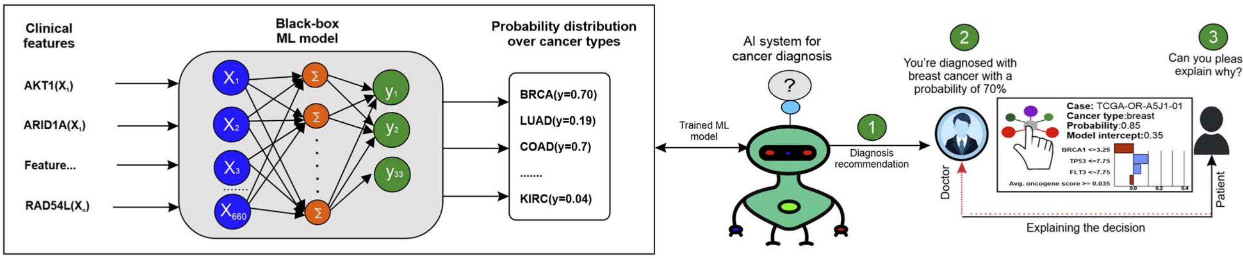


Figure 2. Example of practical consequence: a black-box model cannot explain diagnosis decisions [12].

research can benefit from XAI methods and improve decision fairness. In the next section, we discusses the importance of interpretability and key challenges of black-box ML.

ADVANTAGES OF XAI IN BIOINFORMATICS

Handling large-scale biomedical data involves several challenges such as the presence of heterogeneity, high dimensionality, unstructured formats, noise and incompleteness, and the occurrence of high levels of uncertainty. Further, despite its data-driven nature and complexity, the adoption of data-driven approaches in many bioinformatics scenarios is hindered by the lack of efficient ML models capable of tackling these challenges. On the other hand, higher interpretability of the model means easier comprehension for targeted users [13]. Therefore, a model itself needs to be interpretable, even though not all the predictions need to be explained. Weber *et al.* [11] showed that under the right conditions, augmentations based on XAI can provide significant, diverse and reliable benefits advantages over *black-box* counterparts. Giannotti *et al.* [7] argued that XAI is important to build trust, enable auditing and to improve the model. We summarize such key benefits of XAI systems in Figure 5.

Helps avoid practical consequences

One of the critical applications of AI is aiding diagnoses and treatment of cancerous conditions. Early detection and classification of patients into high- or low-risk groups are crucial for effective management of illness [14]. An example of this is a doctor diagnosing a patient having breast cancer, which is a leading cause of death in women. It is therefore important for the diagnosis to be thoroughly investigated. By utilizing omics data, such as genetic mutations, copy number variations (CNVs), gene expression (GE), DNA methylation and miRNA expression, accurate diagnosis and treatment can be recommended. Suppose a deep learning (DL) model trained on multi-omics data can classify cancerous samples from healthy ones with high accuracy. Assume a patient is diagnosed by the model with breast cancer with a probability of 70%. Knowing the decision, the patient may ask follow-up questions like '*why do I have breast cancer?*' or '*how did the model reach this decision?*' or '*which biomarkers are responsible?,*' it may not be possible to clearly explain the model's decision-making process as the representations learned by the model may not be easily interpretable.

The diagnosis may further depend on several distinct molecular subtypes and factors such as estrogen-, progesterone- and human epidermal growth factor receptors. The diagnosis of a breast cancer patient requires careful examination of multiple sources of data, including omics information, bioimaging and clinical records. A multimodal DNN model trained on this data can classify samples with high accuracy. Further, image-guided pathology is required to analyze imaging data (e.g. breast histopathological image analysis), as shown in Figure 3. However, the representations and decision-making process of such a multimodal model may not be easily interpreted. This can make it challenging to explain the diagnosis to the patient and raises concerns about the model's transparency and accountability in a clinical setting, as shown in Figure 2. Further, the use of black-box models is problematic as their internal logic is hidden from users, leading to theoretical, practical and legal consequences [15]. Therefore, it is important for AI systems used in the diagnosis of cancer to have an explainable mode of operation.

An interpretable ML model that emphasizes transparency and traceability of its logic can explain *why and how* it arrived at certain decisions, reducing negative consequences. In the context of our cancer example, *local interpretability* can provide reasons for a decision made for a specific patient or reference to similar cases, allowing identification of unique characteristics of a patient in a small group [13]. In contrast, *global interpretability* shows the overall behavior of the model at a high level; e.g. if an ML model is trained to predict gene up-regulation after treatment based on the presence of regulatory sequences, global interpretability will indicate the significance of the sequences in predicting up-regulation for all genes in the dataset, while local interpretability will reveal the importance of the sequences in predicting up-regulation for a specific gene.

Reduces modeling complexity and improves accuracy

Interpretability not only helps re duces modeling complexity, but also helps improve its accuracy. For instance, the aim of genomics data analysis is to extract biologically relevant information and gain insight into the role of biomarkers, such as genes, in cancer development. However, biological processes in humans are complex systems controlled by the interactions of thousands of genes, not single gene-based mechanisms [2]. The high dimensionality of omics data, with many genes potentially irrelevant to the task of cancer prediction, creates challenges for ML models. In clinical trials involving over 30 000 genes, the feature space for a model becomes very large, resulting in sparse training data and small sample sizes. Including all features not only hinders the model's predictive power by adding unwanted noise, but also increases computational complexity.

Therefore, selecting biologically significant features with a high correlation to the target classes and a low correlation between genes is crucial. Accurately identifying cancer-specific biomarkers: (i) enhances classification accuracy, (ii) enables biologists to study the interactions of relevant genes and (iii) helps understand their functional behavior, leading to further gene discovery [16]. After identifying these biomarkers based on

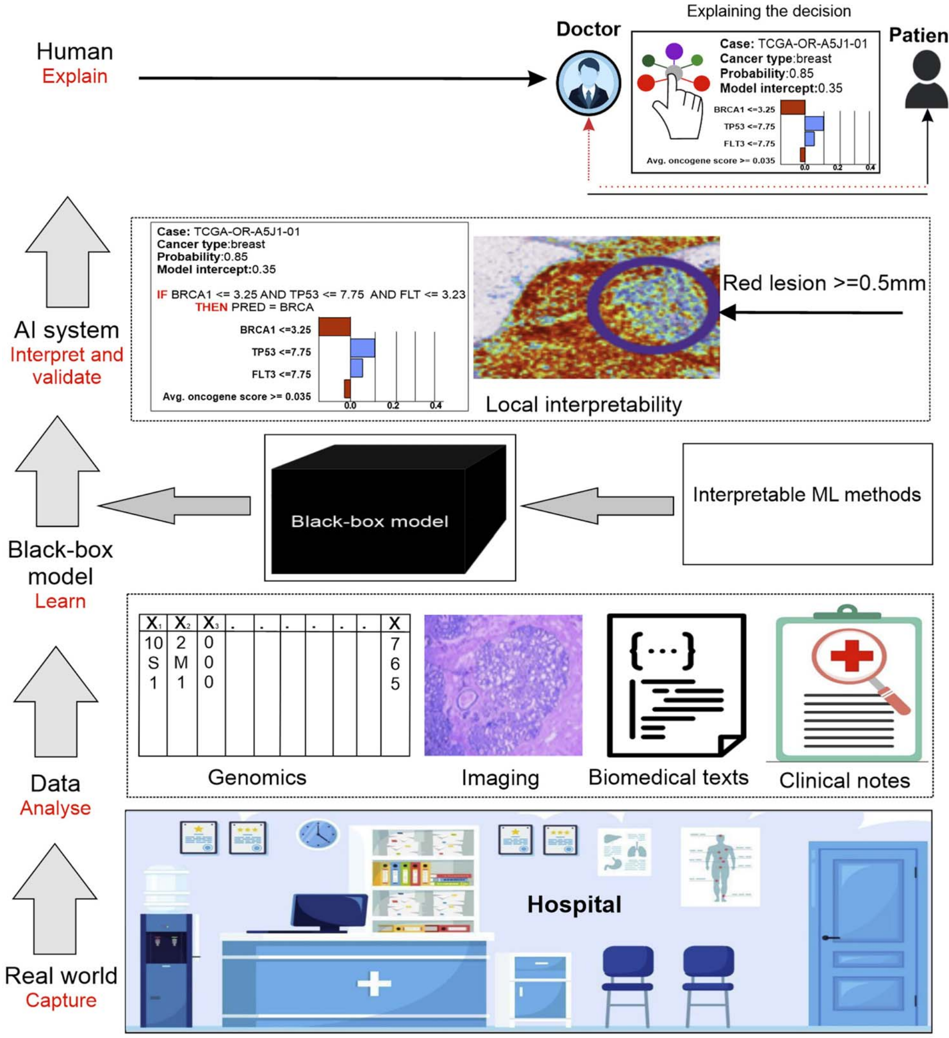


Figure 3. AI for cancer diagnosis in a clinical setting [12].

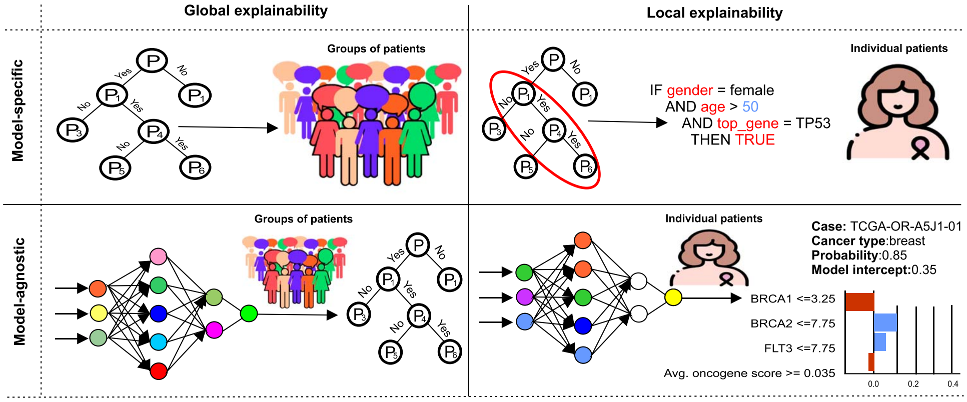


Figure 4. An interpretable model can explain decisions locally, outlining the global behavior (conceptually recreated based on [13]).

feature attributions, they can be ranked based on their relative or absolute importance. These identified genes can serve as cancer-specific marker genes that distinguish specific or multiple tumor classes [17].

Improves decision fairness and trust

Explainability is at the heart of trustworthy AI and must be guaranteed for developing AI systems that empower and engage people across scientific disciplines and industrial settings [7]. With the widespread use of AI, it is essential to address fairness concerns, as AI systems can make significant and impactful decisions in sensitive environments [13]. Bias is a major hindrance to fair decision-making and has been a subject of discussion in philosophy and psychology for a long time [18]. Statistically, bias means a false representation of the truth with respect to the population [18], and it can occur at any stage of the ML pipeline, from data collection, feature selection, model training, hyperparameter setting, to interpretation of results for affected

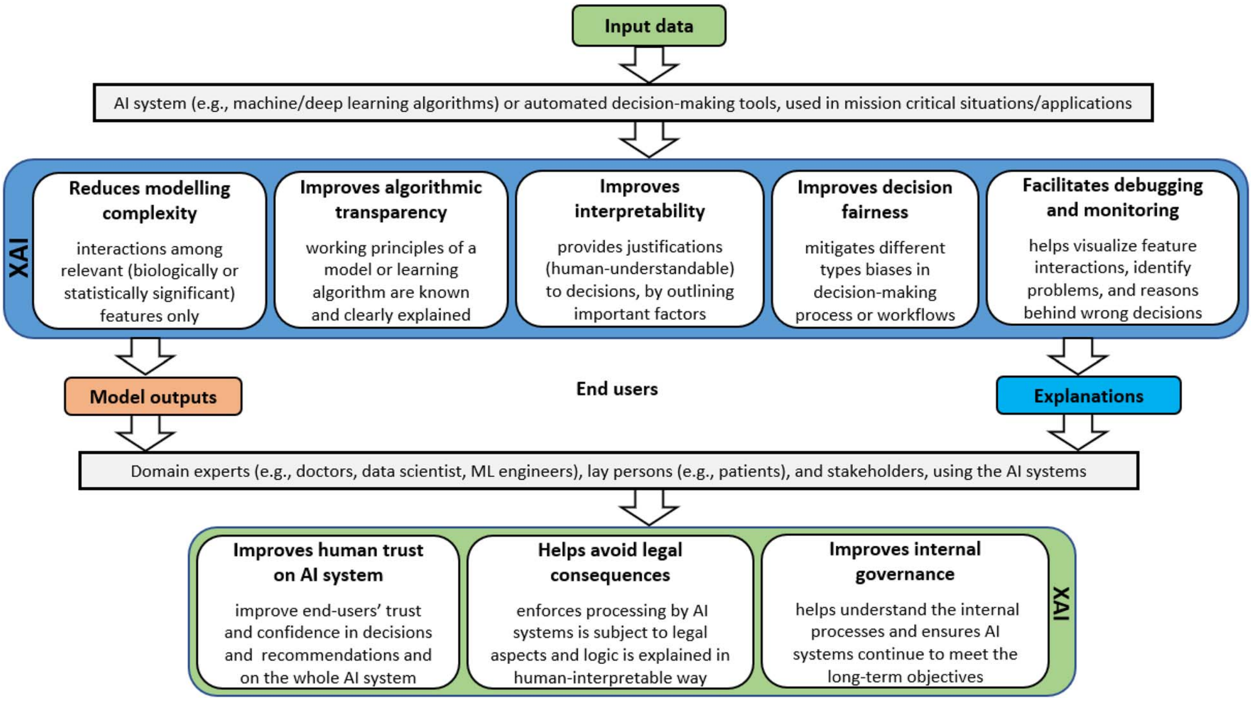

Input data
AI system (e.g., machine/deep learning algorithms) or automated decision-making tools, used in mission critical situations/applications
XAI
Reduces modelling complexity
interactions among relevant (biologically or statistically significant) features only
Improves algorithmic transparency
working principles of a model or learning algorithm are known and clearly explained
Improves interpretability
provides justifications (human-understandable) to decisions, by outlining important factors
Improves decision fairness
mitigates different types biases in decision-making process or workflows
Facilitates debugging and monitoring
helps visualize feature interactions, identify problems, and reasons behind wrong decisions
Model outputs
End users
Explanations
Domain experts (e.g., doctors, data scientist, ML engineers), lay persons (e.g., patients), and stakeholders, using the AI systems
Improves human trust on AI system
improve end-users' trust and confidence in decisions and recommendations and on the whole AI system
Helps avoid legal consequences
enforces processing by AI systems is subject to legal aspects and logic is explained in human-interpretable way
Improves internal governance
helps understand the internal processes and ensures AI systems continue to meet the long-term objectives
XAI

Figure 5. Advantages of XAI in improving interpretability, trust, transparency and fairness in algorithmic decision-making process.

individuals, such as patients [19]. For instance, in biomedicine, representative-, selection- and discriminatory biases can easily be ingested in biophysical data, thereby raising fairness concerns and potentially leading to unfair outcomes in various learning tasks [18]. Therefore, since ML algorithms are a type of statistical discrimination that may become problematic when they give certain privileged groups a systematic advantage and certain disadvantaged groups a systematic disadvantage.

If the data used to train an ML model are biased, the model will also be biased and produce biased decisions. A recent study [21] found that a widely used algorithm exhibits significant racial bias and affects millions of patients. The study showed that at a given risk score, black patients are significantly sicker than white patients and that improving this disparity would increase the number of black patients receiving additional help from 17.7% to 46.5%. Such scenarios can further erode trust in healthcare experts and other stakeholders. In cancer example, predictions based on a biased model can disproportionately impact the diagnosis, potentially leading to incorrect treatments. As humans are inherently biased, data collected or prepared by humans will always contain some level of bias. Therefore, it is important to be aware of common human biases that may appear in the data so that steps can be taken to reduce their impact before training an ML model. Making fair decisions requires a thorough understanding of the context, and interpretability and explainability can help identify factors that may lead to unfair outcomes [18]. By identifying these factors, proactive measures can be taken to prevent discrimination against certain groups or populations [18].

Internal governance and legal compliance

As AI becomes more widespread, the need for transparency of AI decisions grows for ethical, legal and safety reasons [22]. In sensitive domains such as healthcare, where AI may impact human lives, explainability and accountability are not only desirable, but also legally required. The EU General Data Protection Regulation (EU GDPR) recognizes the importance of ethics, accountability and robustness in AI, and requires that automated decision-making processes have suitable safeguards, including the right to understand and challenge the decision, and the right not to be solely subject to an AI-based decision that significantly impacts one's life [23]. Further, it enforces that processing based on automated decision-making tools should be subject to suitable safeguards, including *'right to obtain an explanation of the decision reached after such assessment and to challenge the decision'* and individuals *'have the right not to be subject to a decision based solely on automated processing and whenever human subjects have their lives significantly impacted by an automatic decision-making machine'* [23].

Given the central role AI plays in interactions between organizations and individuals [24], it is important for AI-assisted decisions to be explained in a way understandable to those affected [25]. The GDPR prohibits the use of AI for automated decision-making unless the logic involved is clearly explained. To avoid any legal consequences, the decision-making process should be made as transparent as possible through interpretability [22]. By making explainability a requirement, the organization will have a greater understanding of what the AI system does and why, which will improve oversight and increase precision. This also helps the organization comply with parts of the GDPR and adhere to external policies, practical consequences and processes that regulate business practices [25].

TECHNIQUES AND METHODS FOR INTERPRETABLE ML

A wide range of model-specific and model-agnostic interpretable ML methods have been proposed and developed [68]. Figure 6 shows the brief timeline for XAI methods we discussed in this section and Table 1 categorized interpretable ML techniques based on agnosticism, scopes, underlying methodology and supported data types. All these methods largely fall into three main categories: probing, perturbing, and model surrogation. Further, depending on the level of abstractions, they can be categorized as local interpretability and global interpretability methods. We provide a list of papers, methods and tools as well as supporting Jupyter notebooks, covering bioimaging, cancer genomics, text mining and reasoning examples. The methods discussed in the following subsections largely operate by approximating the outputs of a black-box model via tractable logic or by approximating via a linear model. Besides, tree-, rules- and knowledge-based interpretable methods have been proposed. However, before we discuss different interpretable ML methods, we cover some terminologies and notations related to interpretable ML.

Terminologies and notations

Interpretability and explainability are interchangeably used in literature. However, the former signifies cause and effect of an outcome in a system, whereas the latter is the extent to which the internal working mechanism of an AI system can be explained. According to the Cambridge Dictionary (https://dictionary.cambridge.org/dictionary/english/interpretable), *'if something is interpretable, it is possible to find its meaning or possible to find particular meaning in it'*. Miller *et al.* [26] define *explanation* as the answer to 'why' questions. Das *et al.* [22] define interpretation as a *'simplified representation of a complex domain, such as outputs generated by an ML model, to meaningful, human-understandable and reasonable concepts'*. Interpretability *'is the degree to which a human can understand the cause of a decision'* [26]. Interpretability of the ML model is the extent to which the cause and effect can be observed.

An interpretable ML model refers to methods that make the behavior and predictions of a system understandable to humans [6]. Algorithmic transparency suggests *'factors that influence the decisions made by algorithms should be visible, or transparent, to the people who use, regulate, and are affected by systems that employ those algorithms'* [67]. An interpretable model can outline how input instances are mapped to certain outputs by identifying statistically significant features, while explainability is using the knowledge of what those features represent and their relative importance in explaining the predictions in an understandable way or terms. Further, we refer to following terminologies and notations from our prior work [16] to understand several concepts used in this paper.

Definition 1. (**Dataset**) $D = (\tilde{X}, \tilde{Y})$ is a dataset, where $\tilde{X}$ be an N-tuple of M-instances, X be the set of all instances in $\tilde{X}$, and $\tilde{Y}$ be an N-tuple of labels $l \in L$.

Definition 2. (**Model and prediction**) Let the pair $(\boldsymbol{name}, \boldsymbol{value})$ be a parameter and Θ be a set of parameters. A *model* f is a parametric function $f : X \times \Theta \to \mathbb{R}$ that maps an input instance x from its feature space X to a decision $y \in L$ and returns a real-valued output called *prediction* $\hat{y}$. A prediction $\hat{y} = f(x_i, \theta)$ is accurate for model f and parameter $\theta \in \Theta$ if and only if $\hat{y} = \tilde{Y}[i]$ (classification) or $\hat{y} \approx \tilde{Y}[i]$ (regression) for $x = \tilde{X}[i]$, where $1 \leq i \leq M$.

Definition 3. (**Black-box and interpretable models**) Let f be a model and Θ be a set of parameters. Model, f is a *black box* if its internal working principle and θ are hidden or uninterpretable by humans owing to lack of traceability of how f makes predictions. Model f is *interpretable* if the parameters θ are known and there exists a mathematical interpretation λ_i showing how and why a certain prediction $\hat{y}$ is generated by f.

Definition 4. (**Algorithmic transparency**) A model f is transparent if there exists a mathematical interpretation λ_t by which a learning algorithm learns model f by mapping relations between X and Y to make predictions.

Definition 5. (**Feature importance**) Let a_i be a feature in instance x and A be the set of all features. An importance function $h : A \to [0, 1]$ assigns to each element of A a non-negative number between 0 and 1: the larger the number, the higher the importance of a_i. Local feature importance for x is a set of feature and importance pairs $I_x = \left\{(a_1, h(a_1)), (a_2, h(a_2)), \ldots, (a_M, h(a_M))\right\}$ for all $a_i \in x$. Global feature importance for X is a set of feature and importance pairs $\bar{I}_X = \left\{(a_1, \bar{p}_1), (a_2, \bar{p}_2), \ldots, (a_k, \bar{p}_M)\right\}$, where $\bar{p}_i$ is the mean local feature importance of a_i.

Definition 6. (**Feature impacts**) Let a_i be a feature of instance x, and A the set of all feature names. An impact function $g : A \to [-1, 1]$ takes an element of A as input and results in a real number between -1 and 1. Local feature impact for x is a set of feature and impact pairs $T_x = \left\{(a_1, g(a_1)), (a_2, g(a_2)), \ldots, (a_M, g(a_M))\right\}$ for all $a_i \in x$. Global feature impact for X is a set of feature and impact pairs $\bar{T}_X = \left\{(a_1, \bar{q}_1), (a_2, \bar{q}_2), \ldots, (a_k, \bar{q}_M)\right\}$, where $\bar{q}_i$ is the mean of all local feature impacts of feature a_i.

Definition 7. (**Top-k and bottom-k features**) Let f be a model, I be the global feature importance for D, and k be an integer. The top-k features is a k-tuple such that for all $i \leq k \leq m$, $I[k[i]] \geq I[k[m+1]]$, and $I[k[i]] \geq I[k[i+1]]$, where k is the number of top features used to explain the decision. A bottom-k features is a k-tuple such that for all $i \geq k \geq m$, $I[k[i]] \leq I[k[m+1]]$, and $I[k[i]] \leq I[k[i+1]]$. (Since the learning algorithm for a model involves stochasticity and the way they are computed could be different, feature importance scores (p) could be different if I is sorted in ascending or descending orders of p.)

Definition 8. (**Top-k and bottom-k impactful features**) Let f be a model, T be global feature impacts for D, and k be an integer. The top-k features is a k-tuple such that for all $i \leq k \leq m$, $T[k[i]] \geq T[k[m+1]]$, and $T[k[i]] \geq T[k[i+1]]$, where k is the number of top features used to explain a decision. A bottom-k impactful features is a k-tuple such that for all $i \geq k \geq m$, $T[k[i]] \leq T[k[m+1]]$, and $T[k[i]] \leq T[k[i+1]]$. (Since the learning algorithm for a model involves stochasticity and the way they are computed could be different, feature impact scores (q) could be different if T is sorted in ascending or descending orders of q.)

Definition 9. (**Interpretability**) *Interpretability* is the degree to which humans can understand the cause of a decision [26]. Global interpretability refers to an explanation E explaining why model f has predicted $\hat{Y}$ for all instances in X, outlining conditional interactions between dependent- and independent variables using some functions $\sigma(\cdot,\cdot)$. Local interpretability e reasons why $\hat{y}$ has been predicted for an instance x, showing conditional interactions between dependent variables and $\hat{y}$, focusing on individual predictions.

Definition 10. (**Algorithmic fairness**) An algorithm is fair if the predictions do not favor or discriminate certain individuals or groups based on sensitive attributes.

Definition 11. (**Explanations**) An explanation e for a prediction $\hat{y} = f(x)$ for an instance $x \in X$ is an object derived from model f using some function $\sigma(\cdot,\cdot)$ that reasons over f and x for y such that $e = \sigma(f, x)$ and $e \in E$, where E is the human-interpretable domain and σ is an explanation function.

Definition 12. (**Decision rules**) A decision rule r is a formula $p_1 \wedge p_2 \wedge \cdots \wedge p_k \rightarrow y$, where p_i are boolean conditions on feature values in an instance x and $\hat{y}$ is the decision. Each decision rule r is evaluated w.r.t x by replacing the feature with a feature value from x, evaluating the boolean condition and reaching a conclusion y if the condition is evaluated to be true.

Definition 13. (**W-perturbations**) Let $D = (\tilde{X}, \tilde{Y})$ be a dataset, let X be the set of all instances in $\tilde{X}$, and let x be an instance of an m-tuple in X. Let x' for x be a resulting vector by applying minimum change Δx to some feature values v_i using an optimization method (e.g. ADAM [9]) such that $y' = f(x')$ against original prediction $y = f(x)$, where $x' = x + \Delta x$. Then, x' is called a W-perturbation of x and y'.

Definition 14. (**Counterfactual rules**) Let $r : p \rightarrow y$ be a decision rule for an instance x and x' is the perturbed vector of x (e.g. w-perturbation. (Making changes to certain features may lead to a different outcome.) A *counterfactual rule* $r^{\ddagger}$ for the boolean conditions p is a rule of the form $r^{\ddagger} : p[\delta] \rightarrow y'$ for $y' = f(x')$ s.t. $y' \neq y$, where δ is a *counterfactual* for original decision $y = f(x)$ for model f.

Definition 15. (**Surrogate model**) Let $x \in X$ be an instance of an m-tuple, $\tilde{Y}$ be N-tuple of labels $l \in L$, and f_b be a black-box model. Model f is a *surrogate model* of f_b if the variance R^2 of f_b captured by f is ≈ 1.

Definition 16. (**Causally interpretable model**) Let f be an interpretable model and q be a causal or counterfactual question. Then, the model f is casually interpretable if there exists at least one answer $\hat{y}_a$ to question q.

Definition 17. (**Local explanations**) Let f be an interpretable model, $\hat{y} = f(x)$ be the prediction for instance x, r be a *decision rule* for $\hat{y}$, Φ be the set of *counterfactual rules* for r, and I be the set of *local feature importance* for $\hat{y}$. A local explanation e explaining decision $\hat{y}$ for x is a triple (I, r, Φ). The domain E_ℓ of e is a N-tuple of triples $(e_1, e_2, \cdots, e_N)$, where N is the number of instances in X.

Definition 18. (**Global explanations**) Let f be a model, $\hat{y} = f(x)$ be the prediction for instance x, $\bar{I}$ and $\bar{T}$ be the sets of global feature *importances* and *impacts* for the set of predictions $\hat{Y}$ for instances X, $\bar{I}_{k_t} = \{(a_1, \bar{q}_{1_t}), (a_2, \bar{q}_{2_t}), \ldots, (a_k, \bar{q}_{k_t})\}$ be the set of top-k features, and $\bar{I}_{k_b} = \{(a_1, \bar{q}_{1_b}), (a_2, \bar{q}_{2_b}), \ldots, (a_k, \bar{q}_{k_b})\}$ be the set of bottom-k features. A global explanation e_g is a pair $\langle \bar{I}_{k_t}, \bar{I}_{k_b} \rangle$. The domain E_g of e_g is a pair $(\bar{I}, \bar{T})$.

Probing black-box models

Several probing techniques have been developed to understand their logic with the view to learn the inner working principles of *black-box* models. Examples of probing techniques include gradient-based methods like gradient-weighted class activation mapping (Grad-CAM++) [30] and layer-wise relevance propagation (LRP) [58], which use first-order gradient information from the black-box model to generate heatmaps indicating the relative importance of input features. These techniques are useful in bio-imaging (e.g. CT/MRT/X-Ray), where a convolutional neural network (CNN) learns features, e.g. class-discriminating pixels using filters and edge detectors across convolutional layers, and generate attention maps to highlight the most important pixels in an image.

Saliency map and gradient-based methods

Saliency maps (SMs) and gradient-based techniques are applied to locate crucial areas and then assign significance to each feature, for instance, a pixel in an image. These techniques include guided backpropagation [36], class activation maps (CAM) [69], Grad-CAM[29], Grad-CAM++ [30] and LRP. In gradient-based methods, heatmaps (HMs) are displayed using absolute output gradients and input nodes' negative gradients are zeroed out at the rectified linear layers of the network during the backward pass. This rectification of gradients results in more precise HMs [70]. The Class-discriminatory Attention Map is utilized to show the weighted combination of feature maps (FMs). To highlight where a CNN focuses more, CAM computes weights for each FM-based on the final convolutional layer. However, if the classifier is replaced with linear layers, the network must be retrained and the classifier's non-linearity disappears.

An improved version of CAM, called Grad-CAM [29], has been proposed that uses globally average gradients of FMs as weights for target class c instead of pooling. The guided backpropagation in Grad-CAM produces more interpretable but less class-sensitive visualizations compared to SMs. Since SMs use true gradients, network weights tend to impose a stronger bias towards specific input pixels. Grad-CAM highlights class-relevant pixels instead of generating random noise [71] by using HMs to focus attention on and locate class-discriminating regions of an image. The class-specific weights for each FM are gathered from the final convolutional layer through globally averaged gradients (GAG) instead of pooling [30].

Grad-CAM has a limitation in visualizing multiple occurrences of the same class in an image with slightly different orientations, causing some objects to disappear from the SMs. This is due to its inability to recognize significance disparities among pixels, leading to parts of objects being rarely localized. To address this, Grad-CAM++ was introduced, which replaces the GAG with a weighted average of pixel-wise gradients. A typical example of using Grad-CAM++, as shown in Figure 7, involves taking a radiograph as input, passing it through convolutional layers, rectifying the convolutional feature maps using guided backpropagation and Grad-CAM++, and then feeding it into a fully-connected

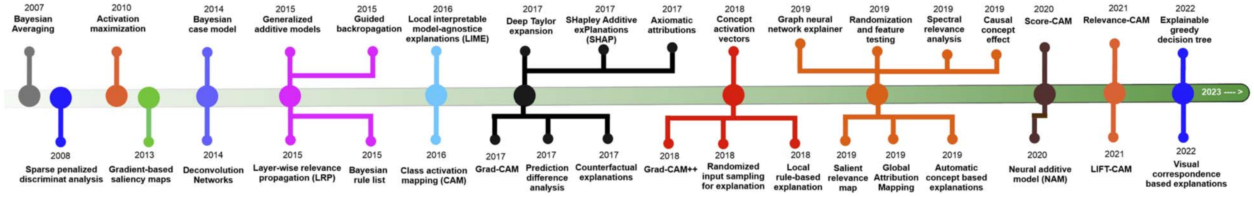

2007
Bayesian Averaging
2008
Sparse penalized discriminat analysis
2010
Activation maximization
2013
Gradient-based saliency maps
2014
Bayesian case model
2014
Deconvolution Networks
2015
Generalized additive models
2015
Layer-wise relevance propagation (LRP)
2015
Guided backropagation
2015
Bayesian rule list
2016
Local interpretable model-agnostice explanations (LIME)
2016
Class activation mapping (CAM)
2017
Grad-CAM
2017
Deep Taylor expansion
2017
Prediction difference analysis
2017
SHapley Additive exPlanations (SHAP)
2017
Counterfactual explanations
2017
Axiomatic attributions
2018
Grad-CAM++
2018
Concept activation vectors
2018
Randomized input sampling for explanation
2018
Local rule-based explanation
2019
Graph neural network explainer
2019
Salient relevance map
2019
Randomization and feature testing
2019
Global Attribution Mapping
2019
Spectral relevance analysis
2019
Automatic concept based explanations
2019
Causal concept effect
2020
Score-CAM
2020
Neural additive model (NAM)
2021
Relevance-CAM
2021
LIFT-CAM
2022
Explainable greedy decision tree
2022
Visual correspondence based explanations
2023 ---- >

Figure 6. Timeline and evolution of interpretable ML methods, covering scopes, methodology and usage level [12].

Table 1. Interpretable ML techniques, categorized based on agnosticism, scopes, underlying methodology and supported data types

Methodology	Approach	Scope	Agnosticism	Supported data types
Back-prop. and integrated gradients	CAM [27, 28]	Local	Model-agnostic	Image
	Grad-CAM [29]	Local	Model-specific	Image
	Grad-CAM++ [30]	Local	Model-specific	Image
	Respond CAM [31]	Local	Model-specific	Image
	DeepLIFT [32]	Local, global	Model-specific	Image, biological sequences
	Slot activation vectors [33]	Global	Model-agnostic	Text
	Peak response mapping (PRM) [34]	Local	Model-agnostic	Image
	DeConvolutional nets [35]	Local	Model-agnostic	Image
	Guided back-propagation [28, 36]	Local	Model-agnostic	Image
	Activation maximization [37]	Local	Model-agnostic	Image
	Gradient-based saliency maps [38]	Local	Model-specific	Image
	Deep attribution maps [39]	Local	Model-agnostic	Image, text
	Axiomatic attributions for deep networks [40]	Local	Model-agnostic	Image, text
	PatternNet and pattern attribution [41]	Local	Model-agnostic	Image
	Spectral relevance analysis (SpRAy) [42]	Global	Model-agnostic	Image
	Salient relevance (SR) map [43, 44]	Local, global	Model-agnostic	Image
Perturbation-based	LIME [45, 46]	Local, global	Model-agnostic	Image, text, tabular
	MUSE [47]	Global	Model-agnostic	Text
	Prediction difference analysis (PDA) [48]	Local	Model-agnostic	Image
	SHapley Additive exPlanations (SHAP) [49]	Local, global	Model-agnostic	Image, text, tabular
	Global attribution mapping (GAM) [50]	Global	Model-agnostic	Image
	Randomized sampling for explanation (RISE) [51]	Local	Model-agnostic	Image
Bayesian	Bayesian averaging over decision trees [52]	Global	Model-specific	Tabular
	Bayesian case model (BCM) [53]	Global	Model-specific	Image, text, tabular
Discriminative	Generative discriminative models (GDM) [54]	Global	Model-specific	Tabular
Decomposition	Deep Taylor decomposition [57]	Local	Model-agnostic	Image
	Layer-wise relevance propagation (LRP) [58]	Local	Model-agnostic	Image, Text
Shape approximation	Neural additive models (NAMs) [59]	Global	Model-specific	Image
Attention-based	ProtoAttend [60]	Global	Model-agnostic	Image
	Self-attention networks (SANs) [61]	Local, global	Model agnostic	Tabular
Graph-based	GNNExplainer [62]	Global	Model-agnostic	Graph
	GCFExplainer [63]	Global	Model-specific	Graph
Concept-based	Concept activation vectors (CAVs) [64]	Global	Model-agnostic	Image
	Automatic concept-based explanations (ACE) [65]	Global	Model-agnostic	Image
	Causal concept effect (CaCE) [66]	Global	Model-agnostic	Image

softmax layer for classification. The critical areas of the image are localized using HMs.

Although CAM variants back-propagate gradients upto the inputs, gradients are essentially propagated only up to the final convolutional layer, and they are specific to the architecture. LRP is a technique suitable for imaging tasks, as it is based on the idea that the likelihood of a class can be traced backwards through a network to the individual layer-wise nodes of the input [73]. For image recognition, LRP produces HMs that highlights the important pixels of an image for the model's prediction. This is achieved by running a backward pass in a CNN, which is a conservative redistribution of relevance, where nodes that contribute the most to the higher layers receive the most relevance. First, an image is classified through a forward pass, and then the relevance is back-propagated to generate a relevance map. The relevance of each node in a layer is calculated recursively, and if the node-level relevance is negative, it is calculated using the ReLU activation function [73].

While regions highlighted by Grad-CAM++ for class discrimination are less precisely localized and scattered, LRP highlights them more accurately [72]. This is because Grad-CAM++ replaces the GAG with a weighted mean, which highlights conjoined features more precisely. As a result, if a diagnosis has to be based on microscopic histopathology images, an AI-assisted image analysis tool could help confirm the presence of breast cancer.

The class-discriminative regions can be localized with respect to pixel relevance, making Grad-CAM++ and LRP useful in improving the reliability of diagnoses. For instance, Figure 10

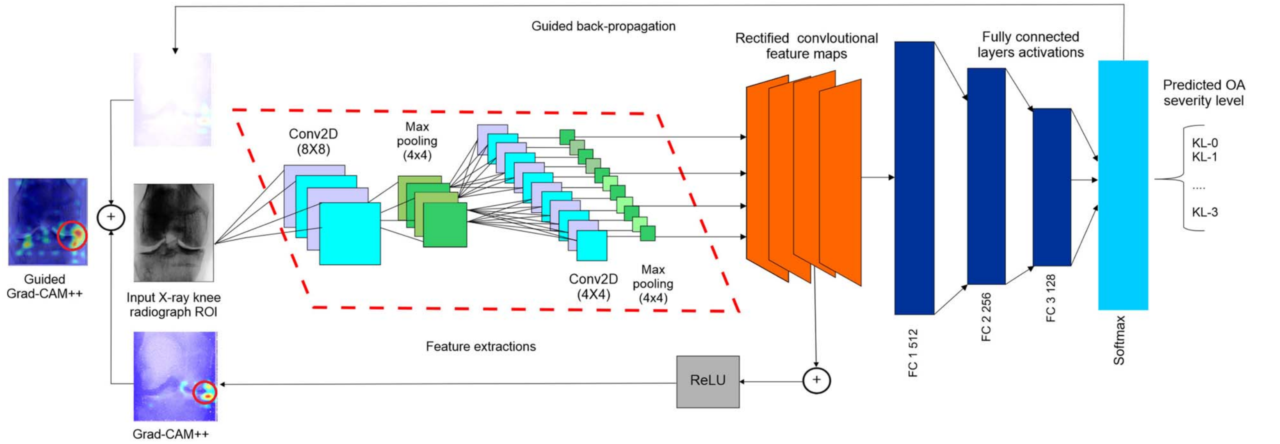


Figure 7. Example of Grad-CAM++ showing pixel relevances to localize critical regions in knee radiography (based on [72]).

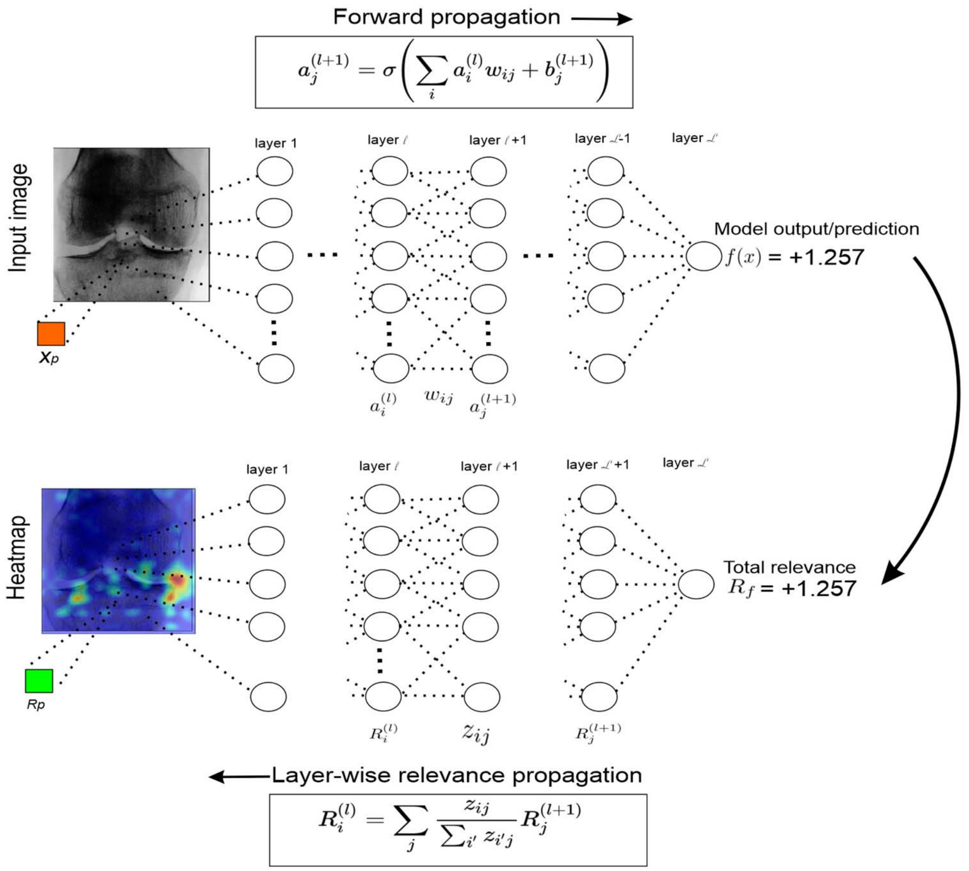


Figure 8. Example of LRP showing pixel relevances to localize critical regions in knee radiography (based on [72]).

shows an example of an explainable diagnosis of osteoarthritis using MRT and X-ray images. As seen, both Grad-CAM++ and LRP generate reliable heat maps that highlight critical image regions and provide precise localization while marking similar regions as important.

Further, saliency map and gradient-based methods can be applied to genomic sequencing data as well. An example (Based on: https://shap.readthedocs.io/en/latest/genomic_examples.html) of computing importance scores on simulated genomic data is shown in Figure 9. It employs three different variants of *DeepLIFT*: integrated gradients, guided backprop and rescale on the conv layers. The model assigns contributions based on difference-from-reference highlighting relevant nucleotide. For example, assuming an one-hot-encoded position $[1, 0, 0, 0]$ for 'A' in the actual sequence and $[0.3, 0.2, 0.2, 0.3]$ in the reference, importance is assigned to the difference (1–0.3), (0–0.2), (0–0.2) and (0-0.3) in the A, C, G and T channels, respectively.

Attention-based probing techniques

Attention mechanisms are designed to identify significant portions of features, leading to enhanced precision in various language modeling assignments. Propositional self-attention networks (PSAN) [74] is an early approach relying on attention, where attention heads symbolize different connections between input features. Transformer language models (TLMs) such as bidirectional encoder representations from transformers (BERT)

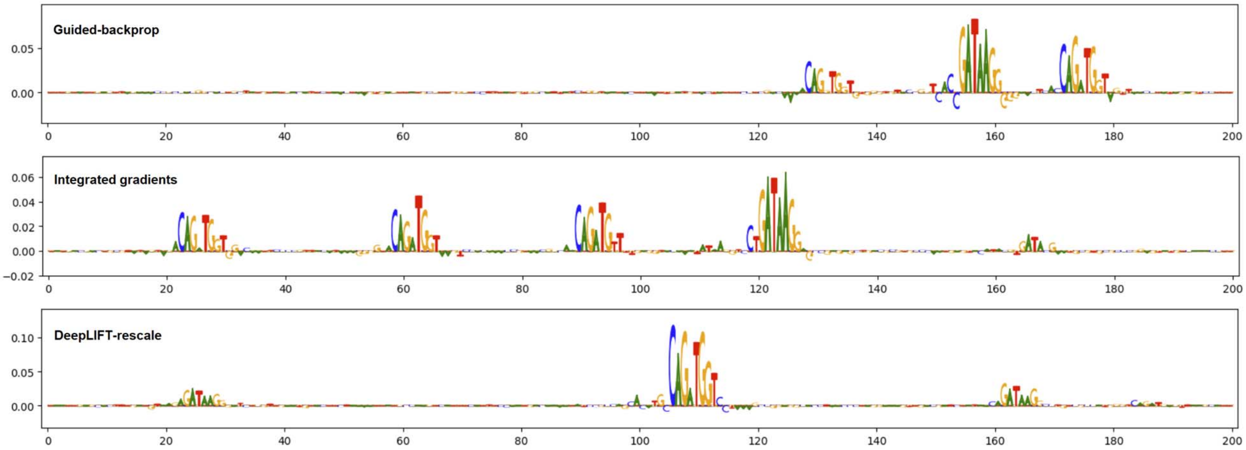


Figure 9. Computing importance scores on simulated genomic data using integrated gradients, guided backprop, and rescale on convolutional layers.

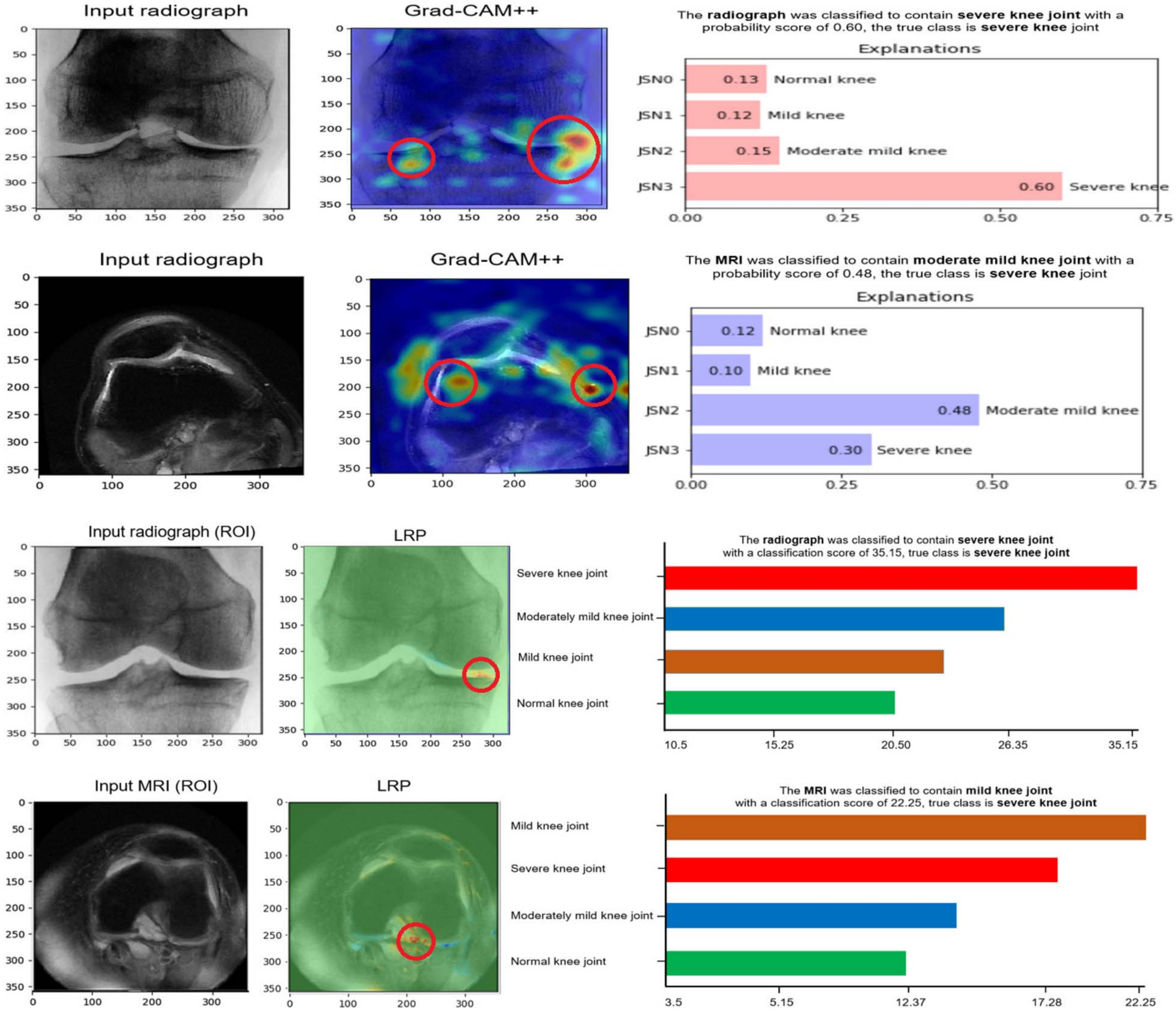

Input radiograph
Grad-CAM++
The **radiograph** was classified to contain **severe knee joint** with a probability score of 0.60, the true class is **severe knee** joint
Explanations
JSN0
0.13
Normal knee
JSN1
0.12
Mild knee
JSN2
0.15
Moderate mild knee
JSN3
0.60
Severe knee
0.00
0.25
0.50
0.75
Input radiograph
Grad-CAM++
The **MRI** was classified to contain **moderate mild knee joint** with a probability score of 0.48, the true class is **severe knee** joint
Explanations
JSN0
0.12
Normal knee
JSN1
0.10
Mild knee
JSN2
0.48
Moderate mild knee
JSN3
0.30
Severe knee
Input radiograph (ROI)
LRP
Severe knee joint
Moderately mild knee joint
Mild knee joint
Normal knee joint
The **radiograph** was classified to contain **severe knee joint** with a classification score of 35.15, true class is **severe knee joint**
10.5
15.25
20.50
26.35
35.15
Input MRI (ROI)
LRP
Mild knee joint
Severe knee joint
Moderately mild knee joint
Normal knee joint
The **MRI** was classified to contain **mild knee joint** with a classification score of 22.25, true class is **severe knee joint**
3.5
5.15
12.37
17.28
22.25
0
50
100
150
200
250
300
350

Figure 10. Example of using Grad-CAM++ and LRP for osteoarthritis diagnosis from knee MRI and X-ray images, highlighting critical knee regions to emphasize plus textual explanations (based on [72]).

[75], employ attention to identify significant tokens for next word prediction by representing diverse connections between input features through bidirectional attention. BERT uses both bidirectional attention and vast amounts of unsupervised data to produce context-sensitive representations [76]. The attention technique is widely used in NLP, computer vision, and speech recognition. Research by Xu *et al.* [77] has demonstrated the potential of TLMs in achieving high accuracy in biomedical text

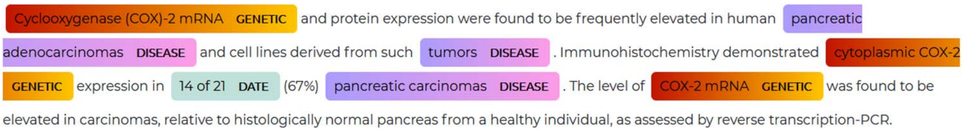

Cyclooxygenase (COX)-2 mRNA GENETIC and protein expression were found to be frequently elevated in human pancreatic adenocarcinomas DISEASE and cell lines derived from such tumors DISEASE . Immunohistochemistry demonstrated cytoplasmic COX-2 GENETIC expression in 14 of 21 DATE (67%) pancreatic carcinomas DISEASE . The level of COX-2 mRNA GENETIC was found to be elevated in carcinomas, relative to histologically normal pancreas from a healthy individual, as assessed by reverse transcription-PCR.

Figure 11. Example of an explainable biomedical NER, where the model classifies and highlights relevant entities.

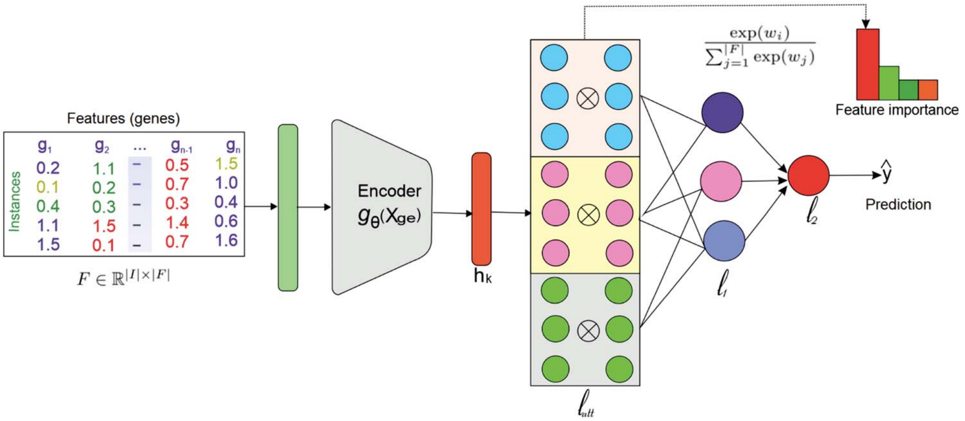


Figure 12. Example of using SAN for cancer type prediction from gene expression data (based on [12]).

mining tasks like named entity recognition from unstructured text [78]. TMLs are effective for domain-specific fine-tuning in a transfer learning setting and thus become the de-facto standard for representation learning for text mining and information extraction in NLP.

Besides, the attention mechanism is also increasingly applied. Self-attention network (SAN) [61] is proposed to identify important features from datasets having a large number of features, indicating that having not enough data can distil the relevant parts of the feature space [61]. TabNet [79] is another approach, which uses a sequential attention mechanism to choose a subset of semantically meaningful features to process at each decision step. It visualizes feature importance and how they are combined to quantify individual feature contributions leveraging local and global interpretability. Approaches such as SAN and TabNet, place attention layers (ALs) as hidden layers that map real values to parts of the human-understandable input space [61]: an element-wise product with X is computed in the forward pass to predict labels $\hat{y}$ in which two consecutive dense layers l_1 and l_2 contribute to predictions, ⊗ and ⊕ are Hadamard product- and summation across k heads. Once the training is finished, weights of ALs are activated using softmax [74]. Top-k features then can be extracted as diagonal of $W^k_{l_{att}}$ and ranked w.r.t. their respective weights.

Perturbing black-box model

Feature-based attributions, game theoretic approach and sensitivity analysis fall in this category that perturb the *black-box* models.

Feature-based attribution methods

Knowing what features are statistically most important to a model help achieve and generate human-level interpretability. Some features have a higher impact than others. This notion of feature importance (ref. definition 5) can be computed as permutation feature importance (PFI). PFI works by randomly permuting a single column in the validation dataset leaving all the other columns intact, where a feature is considered *important* if and only if the model's accuracy drops significantly, thereby increasing the prediction error. A feature is considered *unimportant* if switching its position does not significantly affect the accuracy or performance of the model.

Feature importance can be conceptualized both locally and globally: effects of a feature for a single prediction, over a large number of samples, or for the overall predictions. Let x_i be a feature in instance x. Methods supporting feature importance define an explanation function $g : f \times \mathbb{R}^d \mapsto \mathbb{R}^d$ and return the importance scores $g(f, x) \in \mathbb{R}^d$ for all features for f and a point of interest in x. Besides, feature importance can be extracted from tree-based models, e.g. DTs or tree ensembles.

Game theoretic approach

For a non-linear function, the order in which features are observed by the model matters. Scott *et al.* [49] showed that tree-based methods are similar to using a single ordering defined by a tree's decision path, yielding inconsistency in PFI [49]. A widely used perturbation-based method is SHapley Additive exPlanations (SHAP). SHAP is a game theory-based approach inspired by Shapley values (SVs). SVs are based on coalitional game theory, i.e. the average marginal contribution of a feature and a way to distribute the gains to its players. [80]. Therefore, SVs give an idea of how to fairly distribute the payout [6]. Let f be an interpretable model, x_i be a feature in instance x, and $\hat{y}$ be the prediction. SHAP explains the prediction $\hat{y}$ by computing the contribution of each feature x_i w.r.t SVs and used as measure of feature attributions, where each feature x_i acts as a player in a coalition. SHAP models an explanation as [6]:

$$f(z') = \phi_0 + \sum_{i=1}^{C} \phi_i z'_i, \tag{1}$$

where $z' \in \{0,1\}^C$ is the coalition vector (Coalition vector is a simplified feature representation for tabular data.) such that the effect of observing or not observing x_i is calculated by setting $z'_i = 1$ or $z'_i = 0$), C is the maximum coalition size which is equal to the number of input features M and $\phi_i \in \mathbb{R}$ is the feature attribution for x_i or SVs. The importance ϕ of a feature x_i is computed by comparing what model f predicts with and without

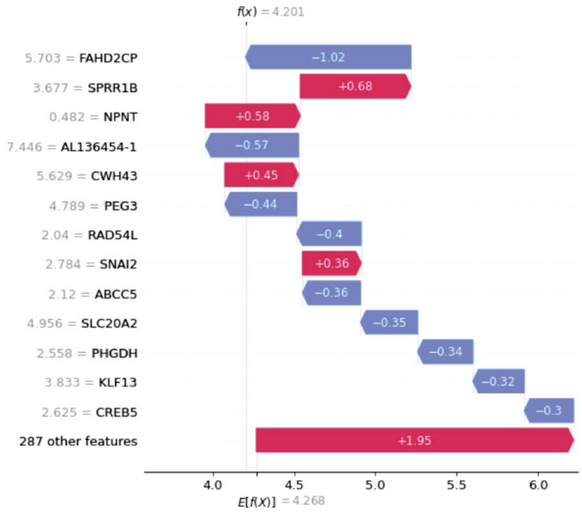


Figure 13. Explaining a single cancer prediction using SHAP waterfall plot: the bottom starts as the expected output value, each row shows how the positive or negative contribution of individual features pushes the value from the expected output over the training set to model output. Assuming the true class label is bladder urothelial carcinoma (BLCA), the positive values imply probabilities above 50% that the patient is diagnosed correctly.

x_i for all possible combinations of $M - 1$ features (i.e. except for feature x_i) in x [81].

SHAP values explain the output of a function as a sum of the effects ϕ_i of each feature being introduced into a conditional expectation. SHAP averages over all possible orderings for computing the mean SVs. If x_i has no or almost zero effect on the predicted value, it is expected to produce an SV of 0. If two features x_i and x_{i+1} contribute equally to the prediction, SVs should be the same [49]. To compute global importance, absolute SVs per feature across instances are averaged as [81].

Sensitivity analysis

For a *black-box* or interpretable model f, sensitivity analysis (SA) is used to explain a prediction $\hat{y}$ based on the model's locally evaluated gradient or partial derivatives. Sensitivity $R_{x_i} = \| \frac{\partial}{\partial x_i} f(x) \|$ quantifies the importance of an input feature x_i at a low level (e.g. image pixel). This measure assumes that the most relevant features are those for which the output is sensitive. Often HMs are plotted to visualize which pixels need to be changed to make the image look similar to the predicted class. However, such an HM does not indicate which pixels are pivotal for a specific prediction, making them not suitable for quantitative evaluation and to validate globally important features. Therefore, SA is more suitable for tabular data to inspect which features a model is more sensitive to.

To perform SA for a tabular dataset, the value for a feature x_i is changed by keeping other features unchanged. If any change in its value significantly impacts the prediction, the feature is considered to have a high impact and is statistically significant. For this, a new test set $\hat{X}^*$ is often created by applying w-perturbation over feature x_i, followed by measuring its sensitivity at the global level. To measure the change in the predictions, the mean square error (MSE) between actual and predicted labels are compared. However, SA requires a large number of calculations (i.e. $N \times M$; N and M are the number of instances and features) across predictions. (Therefore, some approaches [16] recommends making minimal changes to top-k features only, thereby reducing the computational complexity.)

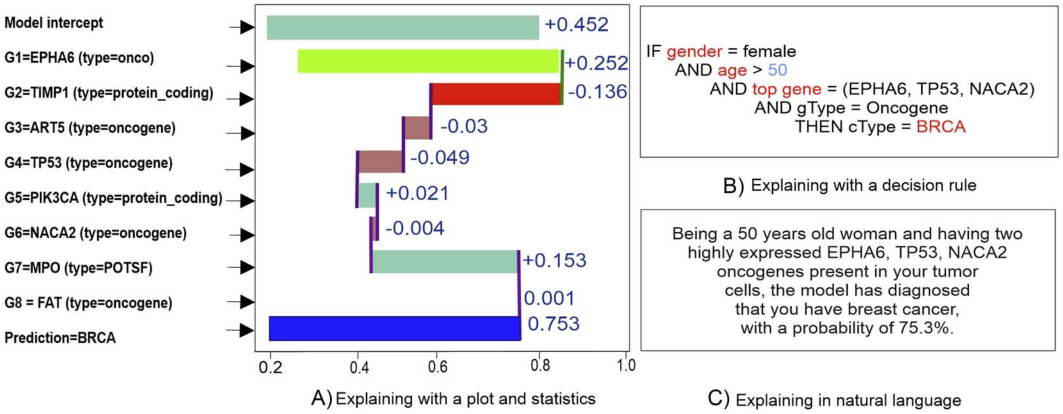
Tree-, text- and rules-based approaches

For the cancer diagnosis scenario, suppose a DT classifier is used for cancer type prediction. As depicted in Figue 14, consider a test instance where the model predicts that the patient has breast cancer with a probability of 75.3% based on their GE profile. This prediction shows an average response (i.e. model intercept) of +0.452. The patient's gender, age and some marker genes are found to be the most influential features based on their feature impact score. However, combining these concepts to explain the decision may not be easily understandable for all users, such as patients. This is because using plots and charts to explain a decision can be helpful for exploration and discovery, but may be challenging for patients to interpret. Rule-based explanations are more easily understood as they relate the feature values of a sample to its prediction [82].

Using a *decision rule* (DR), it is easier to explain the decision in a way that is intuitive to humans. DRs have several features: (i) a *general structure* – if particular conditions are met, then make a certain prediction, (ii) the *number of conditions* – there must be at least one *feature=value* statement in the condition, with no upper limit and additional statements can be added using the 'AND' operator, (iii) *single or multiple DRs* – while multiple rules can be used to make predictions, sometimes a single DR is sufficient to explain the outcome. The same decision can be translated into a DR: '*IF gender = female AND age > 50 AND top gene= (EPHA6, NACA2) AND gType = Oncogene, THEN type = BRCA.*' However, interpreting this for a patient may still be difficult unless it is explained in a human-interpretable way. A simple explanation, in a natural language, could be '*increased breast cancer is associated with risk for developing lymphedema, musculoskeletal symptoms and osteoporosis. Being a 50 years old woman and having two highly expressed oncogenes EPHA6 and NACA2 mutated in your tumor cells, you are diagnosed with breast cancer positive, with a probability of 75.3%.*'. The doctor could further explain: '*the model learns attention on the patches from the image and localizes the malignant and normal regions, based on which the decision has been made*'.

Local interpretable model-agnostic explanations (LIME) approximates a *black-box* model via an interpretable model to explain the decisions locally. Anchor [83] is another rule-based method that extends LIME. It computes DRs by incrementally adding equality conditions on antecedents w.r.t precision threshold [56]. DRs are arguably the most interpretable predictive models as long as they are derived from intelligible features and the length of the condition is short. However, a critical drawback of rule-based explanations could arise due to overlapping and contradictory rules. Sequential covering and Bayesian rule lists are proposed to deal with these issues. Sequential covering iteratively learns a single rule covering the entire training data rule-by-rule, by removing data points already covered by new rules [6], while Bayesian rule lists combine pre-mined frequent patterns into a decision list using Bayesian statistics [6, 55].

Tree-based methods such as DTs utilize tree structures, where internal nodes signify feature values relative to Boolean conditions and leaf nodes signify predicted class labels. DTs split the training set into multiple subsets based on the threshold values of features at each node, until each subset contains instances belonging to only one class. Each branch in a DT represents a potential outcome, while the paths from the root to the leaf represent the classification rules. The outcome of a decision tree is the predicted label of a leaf, while the conjunctions of conditions in the *IF* clause match up with different conditions in the path [84]. The total of importance values is normalized to 1, with a mean importance of 0 indicating that a feature x_i is



IF gender = female
AND age > 50
AND top gene = (EPHA6, TP53, NACA2)
AND gType = Oncogene
THEN cType = BRCA

Being a 50 years old woman and having two highly expressed EPHA6, TP53, NACA2 oncogenes present in your tumor cells, the model has diagnosed that you have breast cancer, with a probability of 75.3%.

Figure 14. Explaining a decision in a human-interpretable way: (A) explaining with plot, (B) explaining with rule, (C) explaining in natural language [12].

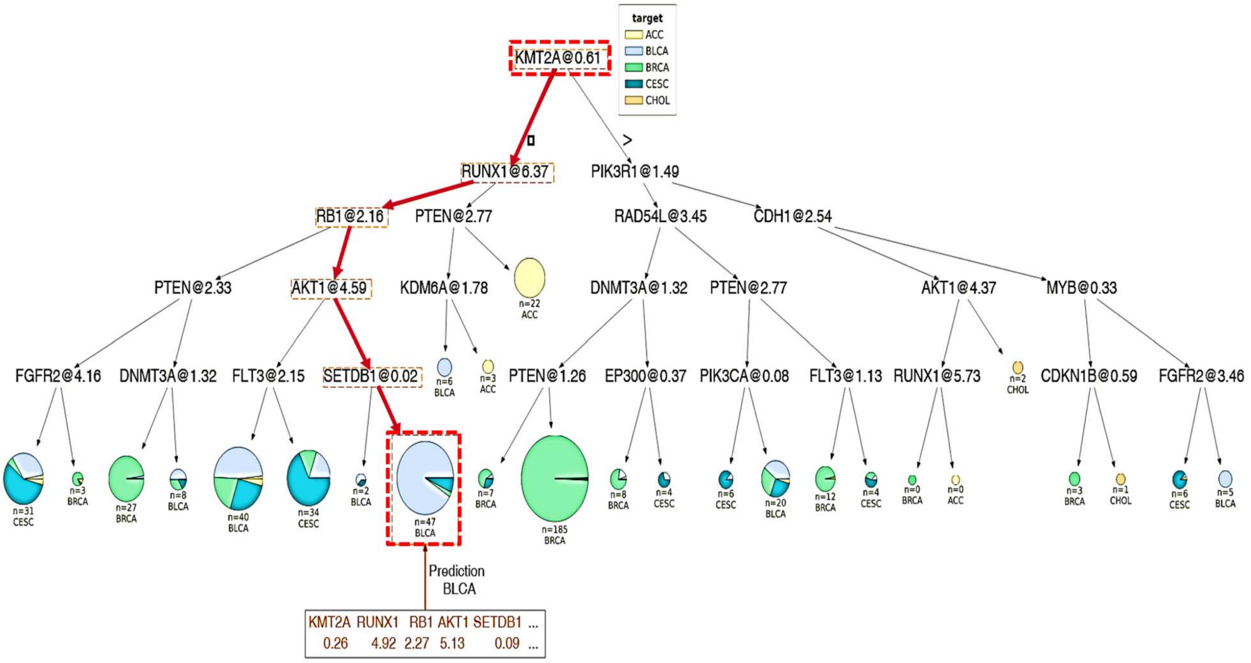


Figure 15. Fig/images/An example showing how a tree-based model has arrived at decision (bladder urothelial carcinoma (BLCA)).

highly important, and a mean importance of 0 indicating that the feature is least important.

As for tree ensembles, e.g. boosted trees (GBTs) or RF, the prediction function $f(\mathbf{x})$ is defined as the sum of individual feature contributions and the average contributions for initial nodes in a DT, and the K possible class labels that change along the prediction paths after each split w.r.t. information gain [85]. DRs can be derived from root-leaf paths in a DT by starting at the root node of the DT and satisfying the split condition of each decision node, until reaching a leaf node containing the decision.

Model surrogation strategies

Since interpretability comes at the cost of accuracy vs. complexity trade-off, research has suggested learning a simple interpretable model to imitate a complex model [6]. A *surrogate* or a simple proxy model is often developed to learn a locally faithful approximation of the *black-box* [13]. Model surrogation is a model interpretation strategy, which involves training an inherently interpretable model by approximating the predictions of a *black-box* [6]. Depending on the complexity of the problem, the surrogate model is trained on the same data the *black-box* was trained on or on sampled data, as shown in Figure 16. Since most important features can be identified by the *black-box* with higher confidence, training an interpretable model on top-k feature space (ref. definition 7) is reasonable. Let X^* be a sampled (e.g. a simplified version of the data containing top-k features only) data for the original dataset X and Y be the ground truths. A surrogate model f can be trained on X^* and for Y. The advantage of model surrogation is that any interpretable model can be used [6], e.g. LR, DTs or GBT or RF classifiers. (Although tree ensembles are complex and known to be *black-boxes*, decision rules can be extracted and FI can be computed from them.)

Casual inference and contrastive explanations

ML models that produce statistical outputs are based on correlation, not causality (i.e. focus on association instead of causality). These models map relevant features (x) to a target variable (y) based on association, not causality. It is important to note that just because there is a correlation between x and y, it does not mean that x causes y, i.e. '*correlation does not imply causation*'. Recent

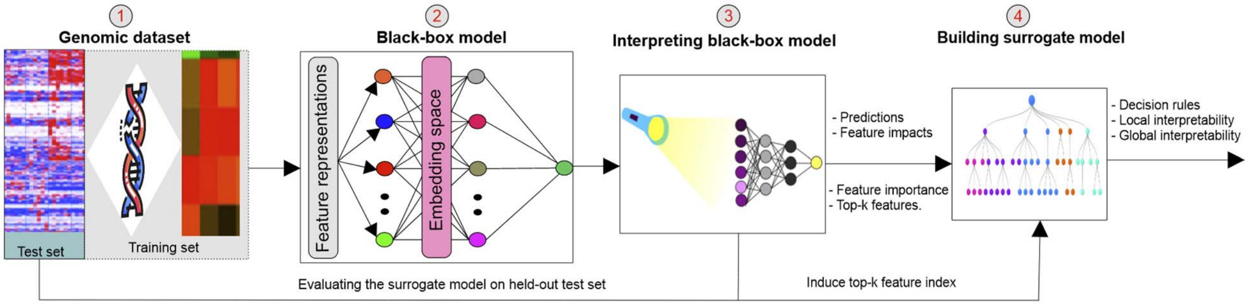


Figure 16. Creation of interpretable surrogate model based on the multimodal black-box model and using the surrogate model to explain the decision (gray circles with red numbers represent different steps of the process) [12].

interpretable ML methods attempt to address causality, such as determining which feature caused a specific diagnosis decision made by the model *'Was it a specific feature that caused the diagnosis decision made by the model?.'* Statistical interpretability reveals associations, while causal interpretability answers 'what-if' and "why" questions, offering a higher level of interpretability [86]. Kim *et al.* [87] suggests learning an oracle model to estimate causal effects for all observed instances and using an interpretable model to approximate the oracle. Another approach is to first train a black-box to learn causal effects, then build an interpretable model using model surrogation strategies. Lipton *et al.* [88] argues that a causally interpretable model is often necessary to ensure fairness.

By using a set of DRs users can focus on learned knowledge instead of underlying data representations [82]. Further, humans tend to think in a counterfactual way by asking a question such as *'How would the prediction have been if input x had been different?'*. *Local rule-based explanations* (LORE) [56] is an approach that learns an interpretable model by computing a neighborhood using a genetic algorithm. LORE derives explanations from the interpretable model in the form of DRs and *counterfactuals* (ref. definition 14). Partial dependence plot (PDP) is another way to depict the marginal effect of features on predicted outcomes. PDP allows measuring the change in predictions after making an intervention (w-perturbations), which can help to discover the features' causal relationship.

Human-in-the-loop and knowledge-based approaches

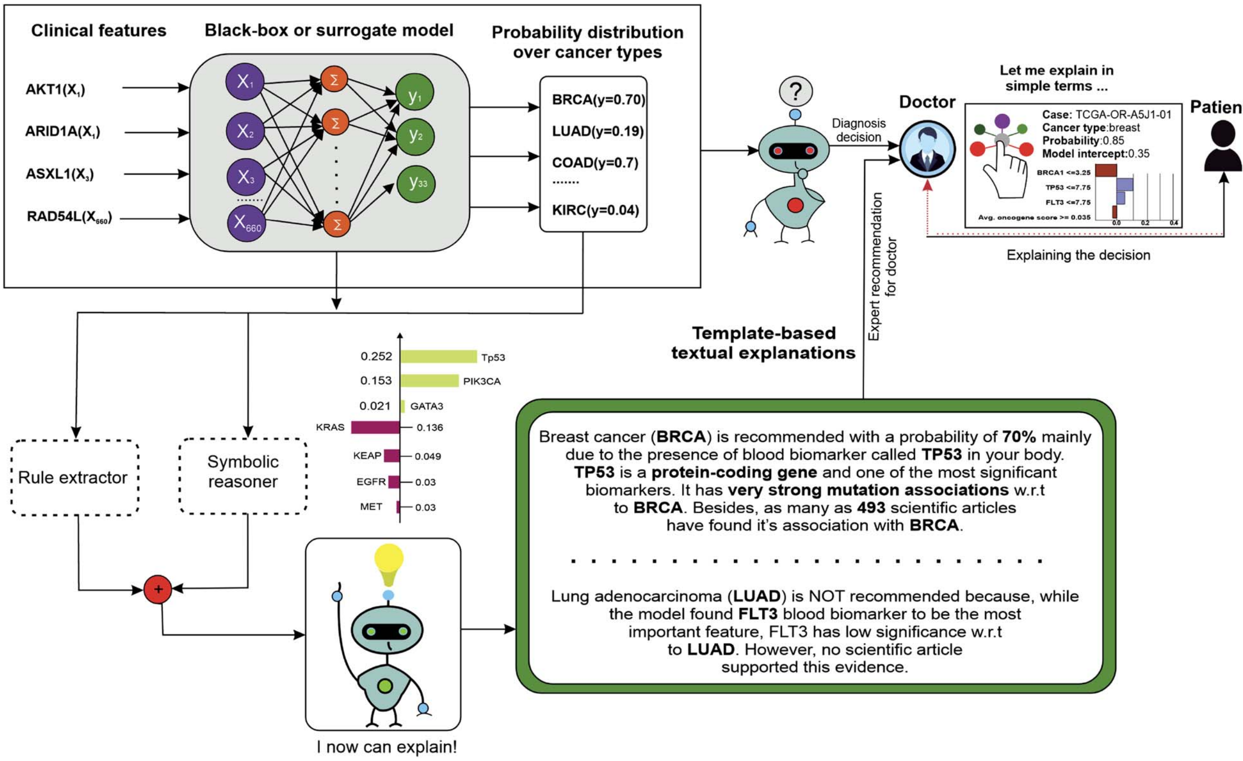
Explanations serve as a bridge between humans and AI systems [84]. By allowing for re-enactment and retracing of AI/ML results, interactive ML systems can incorporate human expertise into AI processes, making them more user-friendly [89]. The human–computer interaction (HCI) community has a long history of advocating for algorithmic transparency in AI systems [90]. As shown in Figure 17, an interactive human–AI interface can be created to evaluate the quality of AI explanations. A more concrete example could be developing an explainable chatbot to enable humans to interact with AI systems and receive interactive answers, such as explanations about cancer diagnosis made by ML models.

ML models are typically trained using data and improved through optimization process, without incorporating structured domain knowledge. In many scientific research, where a deeper understanding is the goal, relying solely on sophisticated ML methods and high accuracy is not enough. While an ML model may outperform humans in prediction and pattern recognition, it lacks the ability to understand the reasoning behind its decisions. This means that it cannot think like a human and make inferences, abstractions and connections as a human would [91]. If the model is trained with data that includes expert knowledge or metadata, a well-fitted *black-box* model would have a deeper understanding of the statistically significant features from a domain perspective.

Interpreting a complex *black-box* model can reveal the most important features in a dataset, e.g. in our cancer diagnosis example, the model can outline, which oncogenes or protein-coding genes are most significant. However, it cannot determine if these features are biologically significance (e.g. are all features in the antecedents biologically relevant?). Thus, generated explanations would be based on statistical learning theory, making it difficult to trust the decisions made by the model. Without validation from domain knowledge, these unreliable decisions could have serious consequences in a medical diagnosis, leading to incorrect treatments. To build a reliable and trustworthy AI system, it is essential to understand the biological mechanisms behind carcinogenesis. An interpretable model is not enough and requires rigorous clinical validation with domain expertise before being used in a clinical setting.

The integration of an ML model with a knowledge-based system would provide human operators with reasoning and question-answering abilities. (This paradigm, which is being emerged as *neuro-symbolic AI* combines both connectionist AI and symbolic AI paradigms together.) Domain experts may need to rely on latest findings from a vast array of sources. For instance, knowledge and facts about drugs, genes, protein, and their mechanism are spread across a huge number of structured (knowledge bases [KBs]) and unstructured (e.g. scientific articles) [12, 92]. These sources can play a crucial role in understanding biological processes, such as diseases, to inform the development of prevention and treatment strategies [92]. The extraction and integration of facts, such as simple statements like *'TP53 is an oncogene'* or quantified statements like *'oncogenes are responsible for cancer,'* into a KG involves three steps: *named entity recognition* (NER) to identify named entities in biomedical texts, *entity linking* to link the extracted entities with the concepts in a domain-speciifc KB, and *relation extraction*.

NER is recognizing domain-specific proper nouns in a biomedical corpus. It can be performed by fine-tuning a domain-specific BERT variant such as BioBERT [93] and SciBERT [94] on relevant articles. For example, for the abstract: *'Cyclooxygenase (COX)-2 mRNA and protein expression were found to be frequently elevated in human pancreatic adenocarcinomas and cell lines derived from such tumors. Immunohistochemistry demonstrated cytoplasmic COX-2 expression in 14 of 21 (67%) pancreatic carcinomas. The level of COX-2 mRNA was found to be elevated in carcinomas, relative to the histologically normal pancreas from a healthy individual, as*



Breast cancer (**BRCA**) is recommended with a probability of **70%** mainly due to the presence of blood biomarker called **TP53** in your body. **TP53** is a **protein-coding gene** and one of the most significant biomarkers. It has **very strong mutation associations** w.r.t to **BRCA**. Besides, as many as **493** scientific articles have found it's association with **BRCA**.

. . . .

Lung adenocarcinoma (**LUAD**) is NOT recommended because, while the model found **FLT3** blood biomarker to be the most important feature, FLT3 has low significance w.r.t to **LUAD**. However, no scientific article supported this evidence.

Figure 17. Decision reasoning with rules in textual and natural human language based facts from a domain-knowledge graph [12].

assessed by reverse transcription-PCR.', an interpretable BioBERT [95]-based NER model would be able to recognize named entities classified as diseases, chemical or genetic, as shown in, where named entities are highlighted with different HMs. These facts can be extracted as a triplet fact in the form of $(u, e, v) = (\textit{subject}, \textit{predicate}, \textit{object})$, where each triple forms a connected component of a sentence for the KG, e.g. for the sample text: *'TP53 is responsible for a disease called BRCA (Breast cancer). TP53 has POTSF (Proto-oncogenes with tumor-suppressor function.) functionality, which is mentioned in numerous PubMed articles,'* *TP53*, *disease*, *BRCA*, *POTSF* and *PubMed* are the extracted named entities. Then, by linking entities with the concept in a domain-specific ontology, the following relation triples can be generated and integrated into a KG: `(TP53, causes, BRCA)`, `(TP53, hasType, POTSF)`, `(BRCA, a, Disease)`, `(POTSF, hasEvidence, PubMed)`.

AI could benefit from external knowledge to support domain experts in understanding why the algorithms came up with certain results [96]. Tiddi *et al.* [96] outlined that decisions in the form of counterintuitive predictions could be more understandable if a model provides evidence in the form of explanations based on external, machine-readable knowledge sources, e.g. ontologies can be used to model a component of the AI system to automatically compose explanations exposing different forms of knowledge in order to address a variety of tasks performed by the agent. Therefore, the full potential of an AI system can only be exploited by integrating both domains and human knowledge [97–99].

A KG can be viewed as the discrete symbolic representations of knowledge [100]. Inference rules (IRs) are a straightforward way to provide automated access to deductive knowledge [100]. An IR encodes IF-THEN-style consequences: $p \rightarrow y$, where both body and head follow graph patterns in KG. Therefore, reasoning over a KG using a reasoner would naturally leverage the symbolic technique and would allowing question answering (in combinations of graph-based explanations techniques such as graph neural network explainer (GNNExplainer) [62] and global counterfactual explainer for GNNs (GCFExplainer) [63]) andjb reasoning. (See supporting Jupyter notebook in the GitHub repo.) It helps deduce implicit knowledge from existing facts [101], e.g. assuming two facts *'TP53 is an oncogene; oncogenes are responsible for cancer'* are already present, reasoning over the KG would entail an extended knowledge *'TP53 is responsible for cancer'* [100]. Then, a doctor with their expertise and by combining the facts from KG, could explain the decision with additional interpretation (e.g. biomarkers and their relevance w.r.t. specific cancer types) [12]. In such a setting, a neuro-symbolic system could be an effective means to generate casual reasoning if the KG contains sufficient knowledge [102].

Measure of interpretability and explainability

When it comes to providing qualitative or quantitative measures of explainability, there is little consensus on what interpretability in ML is and how to evaluate it for benchmarking [89]. Current interpretability evaluation falls into two categories. The first category evaluates explainability in a quantifiable way: a domain user/expert first claims that some model family, e.g. linear models, rule lists are interpretable and then presents algorithms to optimize within that family [26]. The second category evaluates the explainability in the context of its applications from a qualitative point of view. These approaches rely on the notion of *'you will know it when you see it.'* This is too naïve, leaving many questions unanswerable, e.g. *'are all models in all defined-to-be-interpretable'?* or *'are all model classes equally interpretable?.'* However, answers to these questions can only be realized w.r.t. some metrics that could qualitative or quantitatively measure the quality of explanations by an AI system.

Kusner *et al.* [103] proposed a metric for measuring how fair decisions w.r.t. counterfactuals. A decision $\hat{y}$ is fair for instance *x* if the prediction is the same for both the actual world and the counterfactual world in which the instance belonged to a different demographic group. Since surrogate models are often used to explain the decisions, the quality of the explanations depends on the surrogate's predictive power [16]. To measure how well a surrogate f model has replicated a *black-box* f_b, R-squared measure (R^2) is calculated as an indicator of goodness-of-fit. It is conceptualized as the percentage of variance of f_b captured by the surrogate [6]:

- (i) If R^2 is close to 1 (low error), the surrogate model approximates the behavior of the *black-box* model very well. Hence, the surrogate model f can be used instead of black-box model f_b.
- (ii) If R^2 is close to 0 (high error), the surrogate fails to approximate the *black-box*, hence cannot replace f_b (should not use).

Literature [104, 105] outlined rational explanation as a key to understanding an AI system since translating the rationales results into usable and understandable formats. Deyoung *et al.* [105] proposed comprehensiveness and sufficiency to measure the quality of explanations in NLP. Comprehensiveness signifies whether all features needed to make a prediction, while sufficiency signifies whether extracted rationales contain sufficient signal to support the prediction. These metrics are based on the concept of rationales introduced by Zaidan *et al.* [106] in NLP, where a human annotator would highlight a part of the texts that could support the labeling decision.

Let $f(x)_c$ be the original prediction probability for a model (black-box or interpretable) f for the class c. To measure the sufficiency, Deyoung et al. [105] proposed to create contrasting example $\tilde{x}$ for each sample x, by removing predicted rationales r:

$$s = f(x)_c f(r)_c\,, \tag{2}$$

where s measures the degree to which extracted rationales are adequate for the model f for making the prediction [105]. Let $f(x\backslash r)_c$ be the predicted probability of $\tilde{x}$ $(= x\backslash r)$. Thus, it is expected the prediction to be lower on removing the rationales, where comprehensiveness e is calculated as [105]:

$$e = f(x)_c - f(x\backslash r)_c\,. \tag{3}$$

The same idea can be conceptualized for our cancer example w.r.t. leave-one-feature-out: the rationale can be computed based on the number of extracted features (e.g. top-k genes) divided by the number of features in x. A prediction $\hat{y}$ is a match if the overlap with any of the ground truth rationales $r_i \geq \sigma$, where σ is a predefined threshold set by a domain experts. A high value of comprehensiveness implies that the rationales were influential w.r.t prediction $\hat{y}$. If an AI system is useful, then it must somehow be interpretable and able to provide human-understandable explanations [107].

System casuability scale (SCS) [89] is another measure based on the notion of causability [108]. It is proposed to quickly determine whether and to what extent an explainable user interface, an explanation, or an explanation process itself is suitable for the intended purpose [89]. SCS combines with concepts adapted from a widely accepted usability scale and SCS is computed based on responses from 10 usability questionnaires listed in Table 2. SCS system uses a five point scale: rating 1 = strongly disagree; 2 = disagree; 3 = neutral; 4 = agree; 5 = strongly agree. SCS is measured by dividing the acquired ratings by the total ratings, i.e. $\text{SCS} = \sum_{i=1}^{10} \text{Rating}_i/50$.

INTERPRETABLE ML TOOLS AND LIBRARIES

Following the timeline of interpretable and XAI methods (ref. Figure 6), numerous libraries and tools have been developed. Majority of the tools are suitable for general purpose problems. Besides, some existing libraries were customized to solve domain-specific problems.

General purpose XAI tools

Most of these tools and libraries are developed with a view to improving the interpretability and explainability of *black-box* ML models, covering general-purpose problems in computer vision, text mining or structured data, and are based on well-known interpretable ML methods such as LIME [45], model understanding through subspace explanations (MUSE) [47], SHAP [49] (and its variants such as kernel SHAP and tree SHAP), partial dependence plot (PDP), individual conditional expectation (ICE), permutation feature importance (PFI) and counterfactual explanations (CE) [9]. Following are some widely used and general purpose interpretable tools and libraries, and links and their GitHub repositories:

- (i) **DeepLIFT**: https://github.com/kundajelab/deeplift
- (ii) **ExplainerDashboard**: https://github.com/oegedijk/explainerdashboard
- (iii) **Xplique**: https://github.com/deel-ai/xplique/
- (iv) **GNNExplainer**: https://github.com/RexYing/gnn-model-explainer
- (v) **DALEX**: https://dalex.drwhy.ai/python/
- (vi) **Alibi**: https://github.com/SeldonIO/alibi
- (vii) **SHAP**: https://github.com/slundberg/shap
- (viii) **LIME**: https://github.com/marcotcr/lime
- (ix) **PyTorch-Grad-CAM**: https://github.com/jacobgil/pytorch-grad-cam
- (x) **ELI5**: https://github.com/TeamHG-Memex/eli5
- (xi) **InterpretML**: https://github.com/interpretml/interpret
- (xii) **Interpret-Text**: https://github.com/interpretml/interpret-text
- (xiii) **ExplainerDash**: https://github.com/oegedijk/explainerdashboard
- (xiv) **CNNViz**: https://github.com/utkuozbulak/pytorch-cnn-visualizations
- (xv) **iNNvestigate**: https://github.com/albermax/innvestigate
- (xvi) **DeepExplain**: https://github.com/marcoancona/DeepExplain
- (xvii) **Lucid**: https://github.com/tensorflow/lucid
- (xviii) **TorchRay**: https://facebookresearch.github.io/TorchRay/
- (xix) **Captum**: https://captum.ai/
- (xx) **AIX360**: https://github.com/Trusted-AI/AIX360
- (xxi) **BERTViz**: https://github.com/jessevig/bertviz

Customizing XAI methods for bioinformatics

General purpose XAI tools are not specialized to tackle bioinformatics problems by default. This makes the direct application to bioinformatics problems challenging without customization and domain-specific adaptation. For instance, LIME is mostly suitable for tabular data, even though it supports image, text and tabular data. Now in order to perform time series classification, LIME needs to be extended so that it is able to deal with time series data. One approach could be perturbing (e.g. parts of its

Table 2. SCT and it is usability questionnaires with interpretation (adopted from [89])

Usability question	Interpretation
Factors in data	Data included all relevant known causal factors with sufficient precision and granularity
Understood	Explanations are within the context of my work
Change detail level	Level of detail can be changed on demand
Understanding causality	No support was necessary to understand the explanations
Use with knowledge	Explanations helped me understand the causality
No inconsistencies	Explanations could be used with my knowledge base
Learn to understand	No inconsistencies between explanations
Needs references	Additional references were not necessary in the explanations
Efficient	Explanations were generated in a timely and efficient manner.

features are *switched off*, pixels grayed out) input instance several times before feeding into the *black-box*. The approximating model then learns which features to have the most influence on the final prediction.

Methods like SHAP were developed to determine how each input alters the model prediction. However, interpretation of models trained from biological sequences remains more challenging because model interpretation often ignores ordering of inputs. 'Positional SHAP' (PoSHAP) [109] is proposed to interpret models trained from biological sequences by exploiting SHAP to generate positional model interpretations. Study has found that PoSHAP helps improve interpretability for DNN models trained on biological sequences in a variety of tasks such as peptide binding motifs, reflected known properties of peptide CCS, and provided new insights into interpositional dependencies of amino acid interactions. A ML model can be trained on biological sequence data as an input to predict peptide collisional cross section and to predict peptide binding affinity to major histocompatibility complex (MHC) isoforms. (See supporting Jupyter notebook in the GitHub repo.) However, in order to enable positional interpretation for the predictions, input indexes need to be added to the inputs to calculate the SVs from the models. Then, for every sequence, top-five matches can be identified to a given position weight matrix (PWM) and also investigate the total importance assigned to the positions underlying those matches.

Studies have found that domain shift in imaging can result in two major differences in image quality and appearance, with sharpening, changes in contrast, brightness and intensity [110]. Therefore, in case of bioimaging (e.g. MRIs, CT and X-rays), modality-specific preprocessing steps (e.g. rescaling and horizontal flipping in case of radiographs, while MRIs may require contrast enhancement, intensity regulation, noise elimination) are necessary. Further, the network weights should not be initialized with ImageNet like pretrained models, as they often contain photos of general objects, otherwise it would activate the internal representation of the network's hidden layers with geometrical forms, colorful patterns or irrelevant shapes that are usually not present in radiography images [72]. These are necessary to improve model generalization capability as well as not to influence imaging interpretability methods such as Grad-CAM++ and LRP. The histocartography [111] is library designed to facilitate the development of graph-based computational pathology pipelines, where Grad-CAM++ were extended as GraphGradCAM and GraphGradCAM++ for cell graph explainer to generate an explanation to highlight salient nodes.

TLMs such as BERT were originally pre-trained on English corpus such as Wiki dump, newspapers and books. Thus, without domain-specific finetuning they cannot be directly applied to biomedical texts containing a considerable number of domain-specific proper nouns [12]. BioBERT [93] or SciBERT [94] can be initialized with a case-sensitive version of BERT, followed by finetuning their weights on PubMed abstracts to perform the NER and enity linking tasks. Local explanations for individual predictions then can be provided by highlighting important features in an input biomedical text sample in a post-hoc fashion. For instance, the relevance score as a measure of importance can be computed with relevance conservation LRP [112]. The output value for each bio-entity predicted then can be back-propagated layer-wise onto the token level, where token relevances can be visualized with HMs (e.g. [113]) or feature attributions (BERT-LRP [114]).

Sometimes trained models may need to be converted to support the above general purpose XAI tools. For instance, in order to embed the capability for computing importance scores using DeepLIFT, a Keras model need to convert as DeepLIFT model. Further, not each form of explanations can be generated using a single tool, which implies that a single tool or even multiple may need to be customized or be combined, to develop XAI applications for bioinformatics. Following are examples of customized XAI tools that can be used in the contexts of biomedical/bioinformatics use cases:

(i) **Bio-NER**: https://github.com/librairy/bio-ner
(ii) **LIME for time**: https://github.com/emanuel-metzenthin/Lime-For-Time
(iii) **Positional-SHAP**: https://github.com/jessegmeyerlab/positional-SHAP
(iv) **BERT-LRP**: https://github.com/frankaging/BERT-LRP
(v) **Histocartography**: https://github.com/BiomedSciAI/histocartography

CONCLUSION

Interpretability is a key to generating insights on why and how certain predictions are made by a model. In this paper, we discussed the importance of interpretability in bioinformatics and provided a comprehensive overview of interpretable methods. Via several examples of bioimaging, genomics and biomedical texts, we demonstrated how bioinformatics research could benefit from interpretability and different means of explanation types (e.g. rules, plots, heatmaps, textual and knowledge-based). Besides, we analyzed existing interpretable ML tools and libraries that can be employed to improve interpretability for complex bioinformatics research problems.

Although interpretability could contribute to transparent AI [87], interpretability alone cannot guarantee the trustworthiness of an AI system. The benefits of explainability still need to be proven in practical settings. The EU's guidelines on AI robustness and explainability [115] emphasize three key elements for the proper utilization of AI: transparency, reliability and safeguarding of individual data. We argue that there are other important considerations too in the development and deployment phases of an AI system – especially for mission-critical applications like clinical use [25]:

(i) Before building an interpretable model, decision-makers and experts should consider factors such as: (i) what kind of data to use – imaging, text, tabular, graphs?, (ii) what types of explanations to provide, e.g. visual-, tree-, rule- or textual?, (iii) how could the full potential of global interpretability be achieved to tailor or generalize the model better for unseen data?, (iv) how could local interpretability be used if the model fails, so it can be diagnosed before re-training? and (v) what potential impacts the model could have on the targeted users? [25].

(ii) One of the first steps to improving an AI system is to understand its weaknesses: the better we understand which factors cause it to make right predictions or fail, the easier it becomes to improve it [26]. However, such weakness analysis on *black-box* or interpretable models is not straightforward [107], but requires close monitoring and debugging by zooming individual data points.

(iii) Explanations generated by AI-assisted decisions may not only reveal commercially sensitive information, but also the inner workings of an ML model [25]. There are potential dangers associated with the rationale, fairness and types of data explanation, such as information on how similar individuals were treated and details on the input data used to make a decision. To mitigate these risks, it is crucial to restrict the amount of detail provided, such as feature weightings or importance, and to thoroughly evaluate the risk as part of a data protection plan.

(iv) AI systems may be vulnerable to adversarial attacks, bias against underrepresented groups and inadequate protection of individual data, which not only negatively affects user experience but also undermines societal trust. To address these issues, an AI system must be robust to potential adversaries by taking both reactive and proactive measures. It is crucial to ensure that the AI system's predictions remain consistent and reliable even in the presence of minor variations in input data, e.g. adding small amounts of noise to the input should not drastically change the predictions.

(v) An AI system should not only provide *meaningful information* (e.g. clinical outcomes) and clarify the reasoning behind its decisions through supporting explanations, but it should also offer *insights* (e.g. treatment recommendations). Further, biological relevance of important factors needs to be validated both clinically and based on domain knowledge, e.g. oncologists can combine their expertise with evidence from a domain KG.

We believe that the techniques and methods discussed in the paper offer valuable insights into interpretable machine learning and explainable AI. We hope that it will benefit domain experts such as doctors and data scientists, lay people including patients, and stakeholders, ultimately accelerating bioinformatics research.

Key Points

- **Ante-hoc or post-hoc interpretable methods?** Interpretable models should be the first choice due to their simplicity and ease of comprehension. However, in complex scenarios, a surrogate model may not sufficiently approximate the behavior of a *black-box* model. This could result in incorrect decisions if the surrogation process is performed without proper evaluation [16]. It is therefore recommended to build a *black-box* model first, followed by incorporating interpretable ML logic and using the interpretable model to explain decisions. The latter approach can combine both ante-hoc and post-hoc approaches in a single pipeline.
- **Local or global explainability?** Global explainability is important for monitoring and having a holistic view of a model's performance and interpretability (e.g. what features across training instances are more important to the model?). Local explainability is important for individual instances without providing a general understanding of the model. Further, local explainability helps in diagnosing which factors contributed to wrong predictions.
- **What ML models and explanation types?** The choice depends on requirements and data types, as well as target users (decision recipients), e.g. decision rules and counterfactuals are more effective in providing intuitive explanations compared to visualization explanations, an AI system needs to build such that relevant information can be extracted for a range of explanation types [16]. Via a user-study, Tiddi *et al.* [96] identified that different types of explanations required at the different steps of the automated reasoning for scenarios like cancer diagnosis and treatment in clinical settings: 'everyday explanations' for diagnosis, 'trace-based explanations' for planning the treatment, 'evidence-based explanations' to provide scientific evidence from existing studies, and 'counterfactual explanations' to allow clinicians to view/add/edit the diagnosis or treatment recommendations.
- **Accuracy or explainability?** In critical scenarios, both accuracy and reliability are key considerations, as mediocre performance is not acceptable. However, balancing accuracy and explainability can be challenging, given the varying level of complexity of the problem. If the priority is solving a complex problem by building an efficient model, attaining higher accuracy may not be the highest priority, as long as the solution is still interpretable.
- **Can explainability help mitigate unfairness in the decision-making process?** Interpretability and explainability are crucial factors in identifying sensitive features that may lead to discriminatory outcomes in decision-making. They allow for proactive measures to be taken to prevent unfair treatment of certain groups or populations based on such attributes.
- **Is explainability always necessary?** Increased interpretability leads to better understanding of the task and supporting explanations for the end users [13]. However, not all predictions need to be explainable, e.g. in disease diagnosis in translational bioinformatics, addressing AI learning security or fixing learning flaws may be more important than creating an interpretable model [2].

- **Human-in-the-loop and domain knowledge?** HCI researchers should work closely with domain experts, e.g. data scientists and doctors, throughout the development phase, utilizing cutting-edge ML algorithms and interpretable methods. This collaboration will enable the AI developers to design user interfaces that allow for human operators from various domains to ask 'why,' 'how' and 'what-if' questions and receive clear and contrasting explanations in diverse formats.

ACKNOWLEDGMENT

This article is based on a chapter of the Ph.D. dissertation titled "Interpreting black-box machine learning models with decision rules and knowledge graph reasoning" [12] by the first author. Besides, we would like to thanks anonymous reviewers too for their valuable comments and improvement suggestions that further lead us to improve this paper.

FUNDING

Michael Cochez is in part funded by the Graph-Massivizer project, which receives funding from the Horizon Europe research and innovation program of the European Union under grant agreement 101093202.

ABBREVIATIONS

ALs attention layers
AI artificial intelligence
AEs autoencoders
BRL Bayesian rule lists
BCM Bayesian case model
CNN convolutional neural network
CAVs concept activation vectors
CE counterfactual explanations
CLRP contrastive layer-wise relevance propagation
CAE convolutional autoencoder
CAM class activation mapping
CaCE causal concept effect
CI casual inferencing
DL deep learning
DNNs deep neural networks
DRs decision rules
DT decision tree
XAI explainable artificial intelligence
FMs feature maps
GDPR general data protection regulation
GDM generative discriminative models
GAM global attribution mapping
GE gene expression
GB guided back-propagation
Grad-CAM++ gradient-weighted class activation mapping
GAP global average pooling
GNNExplainer graph neural network explainer
GCFExplainer global counterfactual explainer for GNNs
GAG globally averaged gradients
GBT gradient boosted trees
GNNs graph neural networks
HMs heatmaps
Isomap isometric mapping
ICE individual conditional expectation
KNN K-nearest neighbour
KG knowledge graphs
LRP layer-wise relevance propagation
LIME local interpretable model-agnostic explanations
LORE local rule-based explanations
ML machine learning
MACE model-agnostic counterfactual explanation
MUSE model understanding through subspace explanations
NLP natural language processing
NAMs neural additive models
PCA principal component analysis
PDP partial dependence plots
PRM peak response mapping
PSAN propositional self-attention networks
PDA prediction difference analysis
PFI permutation feature importance
RF random forest
ReLU rectified linear unit
SM saliency maps
SAN self-attention network
SCS system causability scale
SC sequential covering
SHAP SHapley additive exPlanations
SA sensitivity analysis
SR salient relevance
SpRAy spectral relevance analysis
SVs Shapely values
SVD singular value decomposition

DATA AVAILABILITY AND IMPLEMENTATION

Codes and interactive Jupyter notebooks are available on our GitHub repository at https://github.com/rezacsedu/XAI-for-bioinformatics.

REFERENCES

1. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthcare J* 2019;**6**(2):94.
2. Han H, Liu X. *The Challenges of Explainable AI in Biomedical Data Science*, Vol. **22**. Springer, 2022.
3. Karim MR, Beyan O, Zappa A, *et al.* Deep learning-based clustering approaches for bioinformatics. *Brief Bioinform* 2020;**22**: 393–415.
4. Fournier Q, Aloise D. Empirical comparison between autoencoders and traditional dimensionality reduction methods. In: *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*. IEEE; 2019, 211–4.
5. Aggarwal CC, Reddy CK. *Data Clustering: Algorithms and Applications*. CRC Press, 2014.
6. Molnar C. *Interpretable machine learning*. Lean Publishing, 2020.
7. Giannotti F, Naretto F, Bodria F. *Explainable for Trustworthy AI. In: Human-Centered Artificial Intelligence: Advanced Lectures*. Springer, 2023, 175–95.
8. Holzinger A, Goebel R, Fong R, Moon T, Müller KR, Samek W. xxAI-Beyond explainable artificial intelligence. In: *xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020*, July 18, 2020, Vienna, Austria, Revised and Extended Papers. Springer; 2022, 3–10.

9. Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harv JL & Tech* 2017;**31**:841.
10. Tjoa E, Guan C. A survey on explainable artificial intelligence (xai): toward medical xai. *IEEE Trans Neural Netw Learn Syst* 2020;**32**(11):4793–813.
11. Weber L, Lapuschkin S, Binder A, Samek W. Beyond explaining: opportunities and challenges of XAI-based model improvement. *Inform Fusion* 2022;**92**:154–76.
12. Karim MR, Rebholz-Schuhmann D, Decker S. *Interpreting Black-box Machine Learning Models with Decision Rules and Knowledge Graph Reasoning*. Aachen, Germany; 2022. Available from:https://publications.rwth-aachen.de/record/850613.
13. Stiglic G, Kocbek P, Fijacko N, *et al.* Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdiscip Rev: Data Mining Knowl Discov* 2020;**10**(5):e1379.
14. Kourou K, Exarchos TP, Exarchos KP, *et al.* Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J* 2015;**13**:8–17.
15. Zednik C. *Solving the black box problem: a normative framework for explainable artificial intelligence*. Philos Technol, 2019, 24, 1.
16. Karim M, Shajalal M, Graß A, *et al.* Interpreting Black-box Machine Learning Models for High Dimensional Datasets . arXiv preprint. arXiv:220813405. 2022.
17. Karim MR, Cochez M, Beyan O, Decker S, Lange C. OncoNetExplainer: explainable predictions of cancer types based on gene expression data. In: *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE; 2019, 415–22.
18. Mehrabi N, Morstatter F, Saxena N, *et al.* A survey on bias and fairness in machine learning. *ACM Comput Surv. (CSUR)* 2021;**54**(6):1–35.
19. Xu C, Doshi T. Fairness Indicators: Scalable Infrastructure for Fair ML Systems. Google Research, 2019.
20. Mitchell M. Fairness. *Crush Course on Introduction to Machine Learning*. Google Research; 2019. Available from:https://developers.google.com/machine-learning/crash-course/fairness/video-lecture.
21. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;**366**(6464):447–53.
22. Das A, Rad P. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. arXiv preprint. arXiv:200611371. 2020.
23. Kaminski ME. The Right to Explanation, Explained. *Berkeley Tech LJ* 2019;**34**:189.
24. Meske C, Bunde E, Schneider J, Gersch M. Explainable artificial intelligence: objectives, stakeholders, and future research opportunities. *Inform Syst Manage* 2022;**39**(1):53–63.
25. Kazim E, Koshiyama A. Explaining decisions made with AI: a review of the co-badged guidance by the ICO and the Turing Institute. Available at SSRN 3656269. 2020.
26. Miller T. Explanation in artificial intelligence: insights from the social sciences. *Artif Intell* 2018;**267**:1–38.
27. He K, Zhang X, Ren S, Sun J. *Computer Vision and Pattern Recognition*. CVPR, 2016.
28. Izadyyazdanabadi M, Belykh E, *et al.* Weakly-supervised learning-based feature localization for confocal laser endomicroscopy glioma images. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2018, 300–8.
29. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision*; 2017, 618–26.
30. Chattopadhay A, Sarkar A. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In: *Conference on Applications of Computer Vision (WACV). IEEE*; 2018, 839–47.
31. Zhao G, Zhou B, Wang K, Jiang R, Xu M. *Respond-cam: analyzing deep models for 3d imaging data by visualizations*. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2018, 485–92.
32. Shrikumar A, Greenside P, Kundaje A. *Learning important features through propagating activation differences*. In: International Conference on Machine Learning. PMLR, 2017, 3145–53.
33. Jacovi A, Sar Shalom O, Goldberg Y. Understanding Convolutional Neural Networks for Text Classification. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Brussels, Belgium: Association for Computational Linguistics; 2018, 56–65.
34. Zhou Y, Zhu Y, Ye Q, *et al.* Weakly supervised instance segmentation using class peak response. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 3791–800.
35. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: *European Conference on Computer Vision*. Springer, 2014, 818–33.
36. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller MA. Striving for Simplicity: The All Convolutional Net. In: *3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA, USA, May 7–9, 2015, Workshop Track Proceedings; 2015.
37. Erhan D, Courville A, Bengio Y. *Understanding representations learned in deep architectures* . Technical Report, 2010.
38. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. In: *2nd International Conference on Learning Representations, ICLR 2014*, Banff, AB, Canada, April 14–16, 2014, Workshop Track Proceedings; 2014.
39. Ancona M, Ceolini E, Öztireli C, Gross M. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In: *6th International Conference on Learning Representations, ICLR 2018*, Vancouver, BC, Canada, April 30–May 3, 2018, Conference Track Proceedings; 2018.
40. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: *International Conference on Machine Learning*. PMLR; 2017, 3319–28.
41. Kindermans PJ, Schütt KT, Alber M, *et al.* Learning how to explain neural networks: Patternnet and patternattribution. In: *6th International Conference on Learning Representations*. ICLR 2018, Vancouver, BC, Canada, April 30–May 3, 2018, Conference Track Proceedings; 2018.
42. Lapuschkin S, Wäldchen S, Binder A, *et al.* Unmasking clever Hans predictors and assessing what machines really learn. *Nat Commun* 2019;**10**(1):1–8.
43. Li H, Tian Y, Mueller K, Chen X. Beyond saliency: understanding convolutional neural networks from saliency prediction on layer-wise relevance propagation. *Image Vision Comput* 2019;**83**: 70–86.
44. Wang Z, Liu Z, Wei W, Duan H. SalED: saliency prediction with a pithy encoder-decoder architecture sensing local and global information. *Image Vision Comput* 2021;**109**:104149.
45. Ribeiro M, Singh S, Guestrin C. *Local Interpretable Model-Agnostic Explanations (ref45): An Introduction*, 2019.
46. Ribeiro MT, Singh S, Guestrin C. Why should i trust you? Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2016, 1135–1144.

47. Lakkaraju H, Kamar E, Caruana R, Leskovec J. Faithful and customizable explanations of black box models. In: *Proceedings of AAAI/ACM Conference on AI, Ethics, and Society*; 2019, 131–138.
48. Zintgraf LM, Cohen TS, Adel T, Welling M. Visualizing deep neural network decisions: Prediction difference analysis. arXiv preprint. arXiv:170204595. 2017.
49. Lundberg SM, Lee SI. *A unified approach to interpreting model predictions*. In: Advances in Neural Information Processing Systems, 2017, 4765–74.
50. Ibrahim M, Louie M, Modarres C, Paisley J. Global explanations of neural networks: mapping the landscape of predictions. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*; 2019, 279–287.
51. Petsiuk V, Das A, Saenko K. Rise: randomized input sampling for explanation of black-box models. arXiv preprint. arXiv:180607421. 2018.
52. Schetinin V, Fieldsend JE, Partridge D, *et al.* Confident interpretation of Bayesian decision tree ensembles for clinical applications. *IEEE Trans Inf Technol Biomed* 2007;**11**(3):312–9.
53. Kim B, Rudin C, Shah JA. The bayesian case model: a generative approach for case-based reasoning and prototype classification. *Adv Neural Inform Process Syst.* 2014;**27**:1–9.
54. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. I ntelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2015, 1721–30.
55. Letham B, Rudin C, McCormick TH, Madigan D. Interpretable classifiers using rules and bayesian analysis: building a better stroke prediction model. *Ann Appl Stat* 2015;**9**(3):1350–71.
56. Guidotti R, Monreale A, Ruggieri S, *et al.* Local rule-based explanations of black box decision systems. arXiv preprint arXiv:180510820. 2018.
57. Montavon G, Lapuschkin S, Binder A, *et al.* Explaining nonlinear classification decisions with deep Taylor decomposition. *Patt Recogn* 2017;**65**:211–22.
58. Bach S, Binder A, Montavon G, *et al.* On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS One* 2015;**10**(7):e0130140.
59. Agarwal R, Melnick L, Frosst N, *et al.* Neural additive models: interpretable machine learning with neural nets. *Adv Neural Inform Process Syst* 2021;**34**:4699–711.
60. Arik SÖ, Pfister T. Protoattend: attention-based prototypical learning. *J Mach Learn Res* 2020;**21**(1):8691–725.
61. Škrlj B, Džeroski S, Lavrač N, Petkovič M. Feature importance estimation with self-attention networks. In: *Proceedings of the 24th European Conference on Artificial Intelligence*; 2020, 1491–98.
62. Ying Z, Bourgeois D, You J, *et al.* Gnnexplainer: generating explanations for graph neural networks. *Adv Neural Inform Process Syst* 2019;9240–51.
63. Huang Z, Kosan M, Medya S, Ranu S, Singh A. Global counterfactual explainer for graph neural networks. In: *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*; 2023, 141–9.
64. Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F, *et al.* Interpretability beyond feature attribution: quantitative testing with concept activation vectors (tcav). In: *International Conference on Mmachine Learning*. PMLR; 2018, 2668–77.
65. Ghorbani A, Wexler J, Zou JY, Kim B. Towards automatic concept-based explanations. *Adv Neural Inform Process Syst* 2019;**32**:9277–86.
66. Goyal Y, Feder A, Shalit U, Kim B. Explaining classifiers with causal concept effect (cace). arXiv preprint. arXiv:190707165. 2019.
67. Diakopoulos N, Koliska M. Algorithmic transparency in the news media. *Digital Journalism* 5 2017;**7**:809–28 2017.
68. Azodi CB, Tang J, Shiu SH. Opening the black box: interpretable machine learning for geneticists. *Trends Genet* 2020;442–55.
69. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2016, 2921–9.
70. Böhle M, Eitel F, Weygandt M, Ritter K. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. *Front Aging Neurosci* 2019;**11**:194.
71. Nie W, Zhang Y, Patel A. A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. In: *International Conference on Machine Learning*. PMLR; 2018, 3809–18.
72. Karim MR, Jiao J, Döhmen T, *et al.* DeepKneeExplainer: explainable knee osteoarthritis diagnosis from radiographs and magnetic resonance imaging. *IEEE Access* 2021;**9**:39757–80.
73. Iwana BK, Kuroki R, Uchida S. Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation. In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. *IEEE*; 2019, 4176–85.
74. Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. *Adv Neural Inform Process Syst* 2017;**30**:5998–6008.
75. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, Minneapolis, MN, USA, June 2–7, 2019, Vol. **1** (Long and Short Papers). Association for Computational Linguistics; 2019, 4171–86.
76. Xue K, Zhou Y, Ma Z, Ruan T, Zhang H, He P. Fine-tuning BERT for joint entity and relation extraction in chinese medical text. In: *Int. Conf. Bioinformatics and Biomedicine (BIBM)*. *IEEE*; 2019, 892–7.
77. Xu J, Kim S, Song M, *et al.* Building a PubMed knowledge graph. *Scientific Data* 2020;**7**(1):1–15.
78. Anantharangachar R, Ramani S, Rajagopalan S. Ontology guided information extraction from unstructured text. *International Journal of Web & Semantic Technology (IJWesT)* 2013;**4**:19–36.
79. Arik SO, Pfister T. TabNet: Attentive Interpretable Tabular Learning. In: AAAI. Vol. **35**; 2021, p. 6679–6687.
80. Branzei R, Dimitrov D, Tijs S. *Models in Cooperative Game Theory*, Vol. **556**. Springer Science & Business Media, 2008.
81. Lundberg SM, Lee SI. Consistent feature attribution for tree ensembles. In: *Proceedings of the 34 th International Conference on Machine Learning*, Sydney, Australia; 2017.
82. Ming Y, Qu H, Bertini E. RuleMatrix: visualizing and understanding classifiers with rules. *IEEE Trans Vis Comput Graph* 2018;**25**(1):342–52.
83. Ribeiro MT, Singh S, Guestrin C. Anchors: high-precision model-agnostic explanations. In: *Thirty-Second AAAI Conference on Artificial Intelligence*; 2018.
84. Guidotti R, Monreale A, Ruggieri S, *et al.* A survey of methods for explaining black box models. *ACM Comput Surv (CSUR)* 2018;**51**(5):1–42.
85. Al-Obeidat F, Rocha Á, Akram M, *et al.* (CDRGI)-cancer detection through relevant genes identification. *Neural Comput Appl* 2021; 1–8.

86. Moraffah R, Karami M, Guo R, *et al.* Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explor Newsletter* 2020;**22**(1):18–33.
87. Kim C, Bastani O. Learning interpretable models with causal guarantees. arXiv preprint. arXiv:190108576. 2019.
88. Lipton ZC. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. *Queue* 2018;**16**(3):31–57.
89. Holzinger A, Carrington A, Müller H. Measuring the quality of explanations: the system causability scale (SCS). *KI-Künstliche Intelligenz* 2020;1 6.
90. Liao QV, Pribić M, Han J, *et al.* Question-driven design process for explainable ai user experiences. arXiv preprint arXiv:210403483. 2021.
91. Kapanipathi P, Abdelaziz I, Ravishankar S, *et al.* Question answering over knowledge bases by leveraging semantic parsing and neuro-symbolic reasoning. arXiv preprint arXiv:201201707. 2020.
92. Karim M, Ali H, Das P, *et al.* Question answering over biological knowledge graph via Amazon Alexa . arXiv preprint arXiv:221006040. 2022.
93. Lee J, Yoon W, Kim S, *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2019;**36**:1234–40.
94. Beltagy I, Lo K, Cohan A. *SciBERT: Pretrained Language Model for Scientific Text*. EMNLP, 2019.
95. Lee J, Yoon W, Kim S, *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;**36**(4):1234–40.
96. Tiddi I, Schlobach S. Knowledge graphs as tools for explainable machine learning: a survey. *Artif Intell* 2022;**302**:103627.
97. Tocchetti A, Brambilla M. The role of human knowledge in explainable AI. *Data* 2022;**7**(7):93.
98. Zhu M, Weng Y, He S, *et al.* ReasonChainQA: text-based complex question answering with explainable evidence chains. In: *2022 China Automation Congress (CAC)*, Xiamen, China, 2022, pp. 5431–36.
99. Rajabi E, Etminani K. Knowledge-graph-based explainable AI: a systematic review. *J Inform Sci* 2022;01655515221112844.
100. Hogan A, Blomqvist E, Cochez M, *et al.* Knowledge graphs. *ACM Comput Surv (CSUR)* 2021;**54**(4):1–37.
101. Futia G, Vetrò A. On the integration of knowledge graphs into deep learning models for a more comprehensible AI—three challenges for future research. *Information* 2020;**11**(2): 122.
102. Xian Y, Fu Z, Muthukrishnan S, De Melo G, Zhang Y. Reinforcement knowledge graph reasoning for explainable recommendation. In: *Proceedings of the 42nd international ACM SIGIR Conference on Research and Development in Information Retrieval*; 2019, 285–94.
103. Kusner MJ, Loftus J, Russell C, Silva R. Counterfactual fairness. *Adv Neural Inform Process Syst* 2017;**30**:4069–79.
104. Yu M, Chang S, Zhang Y, Jaakkola T. Rethinking Cooperative Rationalization: Introspective Extraction and Complement Control. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics; 2019, 4094–103.
105. DeYoung J, Jain S, Rajani NF, Lehman E, Xiong C, Socher R, *et al.* ERASER: A Benchmark to Evaluate Rationalized NLP Models. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics; 2020, 4443–58.
106. Zaidan O, Eisner J, Piatko C. Using "annotator rationales" to improve machine learning for text categorization. *The Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies* 2007; 260–7.
107. Bhatt U, Xiang A, Sharma S, Weller A, Taly A, Jia Y, *et al.* Explainable machine learning in deployment. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*; 2020, 648–657.
108. Holzinger A, Langs G, Denk H, *et al.* Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscipl Rev: Data Mining Knowl Discov* 2019;**9**(4):e1312.
109. Dickinson Q, Meyer JG. Positional ref49 (ref111) for interpretation of machine learning models trained from biological sequences. *PLoS Comput Biol* 2022;**18**(1):e1009736.
110. Zhang L, Wang X, Yang D, *et al.* Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE Trans Med Imaging* 2020;**39**(7):2531–40.
111. Jaume G, Pati P, Anklin V, Foncubierta A, Gabrani M. Histo-Cartography: A toolkit for graph analytics in digital pathology. In: *MICCAI Workshop on Computational Pathology*. PMLR; 2021, 117–28.
112. Arras L, Montavon G, Müller KR, Samek W. Explaining recurrent neural network predictions in sentiment analysis. In: *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Copenhagen, Denmark: Association for Computational Linguistics; 2017 159–68.
113. Karim MR, Dey SK, Islam T, Sarker S, Menon MH, Hossain K, *et al.* Deephateexplainer: explainable hate speech detection in under-resourced bengali language. In: *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*. *IEEE*; 2021, 1–10.
114. Wu Z, Ong DC. On explaining your explanations of bert: an empirical study with sequence classification. ACL'2011. arXiv preprint arXiv:210100196. 2021.
115. Hamon R, Junklewitz H, Sanchez I. Robustness and explainability of artificial intelligence. *Publ Office Eur Union* 2020.