ANNUAL REVIEWS

Annual Review of Pathology: Mechanisms of Disease

Toward Explainable Artificial Intelligence for Precision Pathology

Frederick Klauschen,[1,2,3,4] Jonas Dippel,[3,5] Philipp Keyl,[1] Philipp Jurmeister,[1,4] Michael Bockmayr,[2,6,7] Andreas Mock,[1,4] Oliver Buchstab,[1] Maximilian Alber,[2,8] Lukas Ruff,[8] Grégoire Montavon,[3,5,9] and Klaus-Robert Müller[3,5,10,11]

[1]Institute of Pathology, Ludwig-Maximilians-Universität München, Munich, Germany; email: f.klauschen@lmu.de

[2]Institute of Pathology, Charité Universitätsmedizin Berlin, Berlin, Germany

[3]Berlin Institute for the Foundations of Learning and Data (BIFOLD), Berlin, Germany

[4]German Cancer Consortium, German Cancer Research Center (DKTK/DKFZ), Munich Partner Site, Munich, Germany

[5]Machine Learning Group, Department of Electrical Engineering and Computer Science, Technische Universität Berlin, Berlin, Germany; email: klaus-robert.mueller@tu-berlin.de

[6]Department of Pediatric Hematology and Oncology, University Medical Center Hamburg–Eppendorf, Hamburg, Germany

[7]Research Institute Children's Cancer Center Hamburg, Hamburg, Germany

[8]Aignostics, Berlin, Germany

[9]Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin, Germany

[10]Department of Artificial Intelligence, Korea University, Seoul, Korea

[11]Max Planck Institute for Informatics, Saarbrücken, Germany

ANNUAL REVIEWS CONNECT

www.annualreviews.org

- Download figures
- Navigate cited references
- Keyword search
- Explore related articles
- Share via email or social media

Annu. Rev. Pathol. Mech. Dis. 2024. 19:541–70

First published as a Review in Advance on October 23, 2023

The *Annual Review of Pathology: Mechanisms of Disease* is online at pathol.annualreviews.org

https://doi.org/10.1146/annurev-pathmechdis-051222-113147

Copyright © 2024 by the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See credit lines of images or other third-party material in this article for license information.

Keywords

pathology, deep learning, explainable artificial intelligence, XAI

Abstract

The rapid development of precision medicine in recent years has started to challenge diagnostic pathology with respect to its ability to analyze histological images and increasingly large molecular profiling data in a quantitative, integrative, and standardized way. Artificial intelligence (AI) and, more precisely, deep learning technologies have recently demonstrated the potential to facilitate complex data analysis tasks, including clinical, histological, and molecular data for disease classification; tissue biomarker quantification; and clinical outcome prediction. This review provides a general introduction to

OPEN ACCESS (cc) BY

AI and describes recent developments with a focus on applications in diagnostic pathology and beyond. We explain limitations including the black-box character of conventional AI and describe solutions to make machine learning decisions more transparent with so-called explainable AI. The purpose of the review is to foster a mutual understanding of both the biomedical and the AI side. To that end, in addition to providing an overview of the relevant foundations in pathology and machine learning, we present worked-through examples for a better practical understanding of what AI can achieve and how it should be done.

1. INTRODUCTION

1.1. A Century of Technological Innovation in Pathology

Histopathology was established as the basis of tissue diagnostics more than 100 years ago, and it soon demonstrated the ability to robustly classify diseases and predict outcome. In addition to the technological leap that occurred at that time owing to microscopy, decades of closely correlating clinical observations with histomorphological changes have been complemented by immunohistochemistry and, more recently, molecular profiling techniques that have given pathology its central role in precision medicine. Despite these novel technologies, however, human expertise is still key to integrating clinical imaging and molecular data into a diagnosis, offering a qualitative or sometimes semiquantitative assessment of the pathological findings that guide clinical decisions.

With recent developments in artificial intelligence (AI) and, more precisely, deep learning, hopes are high that another technological leap will transform pathology. Aside from the scientific interest in exploring ways to analyze histopathological data, there is a great medical need because today's precision medicine requires increasingly fine-grained quantitative evaluation of tissue features. While pathologists excel at qualitative assessments of tissue properties for rendering diagnoses, the human brain has limited abilities to quantify observations. Here, AI can assist pathologists in classical tasks such as tumor detection by prescreening tissue for cancer cells and quantification of immunohistochemical stains. While the assistance offered by AI can already help improve diagnostics, AI will reveal its full potential only if novel diagnostic features are identified using end-to-end learning in combination with so-called explainable AI (XAI), which can make the otherwise black-box approach of classical AI transparent.

This review is intended for pathologists and a broader medical audience, as well as computational scientists who would like to better understand the potential as well as the limitations of current machine learning (ML) approaches in the medical domain, especially tissue-based diagnostics. While we provide mathematical descriptions of certain ML concepts for interested readers, these can be skipped without losing the central theme of the review. We focus especially on XAI as a way to extract a link in complex data between concepts of artificial and human intelligence. The review also discusses requirements for implementation of AI in pathology, including technical, organizational, and clinical aspects. Note that this review does not attempt a full treatment of all available literature; instead, we present a somewhat biased point of view illustrating the main ideas by often drawing from the authors' research and providing reference to related work for further reading. The review emphasizes how to insightfully use AI and, in particular, XAI in the quest to bridge the gap between research in disease mechanisms and clinical application. The practical steps of this process are showcased by two examples detailing steps from a pathology challenge to an XAI solution.

2. MACHINE LEARNING

ML, in particular deep learning, is a prominent approach to building predictive models that has successfully solved complex tasks in computer vision (e.g., 1–3), natural language processing (e.g., 4), robotics, and the sciences (e.g., 5, 6), as well as in medical applications (2, 7, 8). ML starts with a data set, for example, a collection of histopathological images with class labels indicating whether the given image contains cancerous tissue. It then learns a complex model (e.g., a neural network) by adjusting its many parameters such that some measure of error is minimized (e.g., the number of misclassified images should be as low as possible).

More formally, we denote $\{(\boldsymbol{x}_1, t_1), \ldots (\boldsymbol{x}_N, t_N)\}$ as the set of data and their respective labels, f_θ as the ML model mapping inputs to predictions, and ℓ as an error or loss function measuring the discrepancy between the ML model's prediction and the labels. Mathematically, the ML problem is to find the parameters θ of the prediction function f_θ that minimize the prediction error averaged over the data set:

$$\mathcal{E}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell(f_\theta(\boldsymbol{x}_i), t_i). \qquad 1.$$

The minimization of $\mathcal{E}(\theta)$ is typically carried out by gradient descent, that is, by repeating iteratively $\theta \leftarrow \theta - \gamma \cdot \nabla\mathcal{E}(\theta)$ until convergence, or variants thereof (9, 10). In the following subsections, we give an overview of the typical practical tasks ML models can solve, models and algorithms to learn these tasks, and methods to verify a model's performance, in particular, verifying that the learned model not only memorizes the current data but also correctly predicts new unseen instances (or in technical terms, does not overfit). We also present several countermeasures to reduce the risk of overfitting. These multiple aspects of ML, along with references, are also presented systematically in **Table 1**.

2.1. Data/Machine Learning Tasks

In the broad ML usage spectrum we distinguish so-called supervised tasks, which aim to learn a specific mapping between input and targets (e.g., predicting for each histological image given as input its associated cancer type), from unsupervised tasks, which aim to reveal the intrinsic structure of data (e.g., the presence of clusters or outliers). Most ML tasks fall into one of the following categories:

1. Classification/regression: predicting a categorical/continuous value from a data point (e.g., the cancer type/survival time from a histology image).
2. Multiple-instance learning (MIL): a specific form of classification (or regression) to learn from a number of observations at once, typically from a number of image regions of a whole-slide image (WSI).
3. Object detection: predicting object positions (and categories) within an image (e.g., cell locations and types).
4. Semantic/instance segmentation: predicting regions/object outlines within an image by assigning a category to each pixel (e.g., stroma regions or cell shapes).
5. Anomaly detection: detecting whether an instance is typical or atypical (the latter outcome may correspond to some rare medical condition to be worthy of special interest).
6. Clustering: identifying whether points naturally aggregate into multiple subgroups, which may give insights into the underlying structure of the data (e.g., multiple cancer subtypes in a population).

In addition to these well-defined tasks, another substantial effort in ML is to build general-purpose data representations. Examples include general feature representations of images, which

Table 1 Representative ML/XAI references[a]

	ML tasks/formulation	Algorithms	Validation and improvement
ML	Classification (1, 15; •30) Regression (31) MIL (32) Segmentation (2) Object detection (3, 33) Representation learning (34, 35) Transfer learning (•7, 11; 36) Self-supervision (12; •22, 37, 38; •23)	Linear models/kernels (15, 39; •40) Decision trees Random forests (•40, 41; •42; •43; ••44) Gradient boosting (•45, •46, •47) Deep learning models (19–21) FFNs/MLPs (•40, •48, •49) CNNs (1, 25; •50; •51) RNNs (26; •52, 53; •54) Transformers (27, 38) Others (•55)	Cross-validation (•40, 41; ••••56) Regularization (•••57) Ensemble models (•40) Data augmentation/prior knowledge (•••58, ••59, ••60, •61)
XAI	Feature types Tabular (•48; 62, 63) Images/text (•56, 64–66) Others (67, 68) Latent variables (64, 69, 70) Higher order (68; 70; •71, 72) Counterfactual explanations (73, 74)	Model-specific (•71, 75, ••76) Model-independent Perturbation (64, 77–79) Backpropagation (•48; •49; •56; •••58; ••60; 64, 65) Others (80, 81) Hybrid approaches Local surrogates (••44; 79, 82; •83–85; •86; •••87) Regularization (88)	Detecting/reducing CH effects (89, 90; •91) Value alignment (92)

[a]Legend: •, histomorphology; •, genomics; •, methylation; •, RNA; •, scRNA; •, proteomics; •, clinical/others.
[b]Abbreviations: CH, Clever Hans; CNN, convolutional neural network; FFN, feed-forward network; MIL, multiple-instance learning; ML, machine learning; MLP, multilayer perceptron; RNN, recurrent neural network; scRNA, single-cell RNA; XAI, explainable artificial intelligence.

are typically obtained by training a large model on some large-scale generic (e.g., nonmedical) image recognition task. Often the learned features are transferable to the specialized (e.g., medical) domain and enable higher predictive accuracy (7, 11). More recently, self-supervised learning (12, 13) has emerged, where a supervised task is built from purely unlabeled data. The resulting so-called pretrained models are then used as a starting point for (supervised) downstream tasks. The promise of this approach is that there are abundant unlabeled data, and models that have seen all these data will generalize better on downstream tasks with restricted data sets. A recent prominent example demonstrating the power of self-supervised learning is the pretrained model GPT (4), which is trained by predicting the next word for a given text and has been refined to ChatGPT by supervised learning in order to respond to user queries. An analogous example in pathology would be training a model that predicts the content of a masked region in a histology image—thereby, for example, learning that immune cells are less likely to appear in dense tumor regions than in sparse ones—and later using this model as starting point for, say, the prediction of clinical endpoints from histology.

2.2. Machine Learning Models and Learning Algorithms

Various ML models and architectures are suitable for different types of data and ML tasks. We can broadly distinguish between linear and nonlinear models. As their name implies, in linear models, data input and model parameters are related in a linear fashion, $f_\theta(\boldsymbol{x}) = \boldsymbol{x}^\top\theta$, as in, for example, linear regression (14) or linear support vector machines (15). Because of the linear relation, the influence of a model parameter on a data input feature can be directly inspected, enabling these models to be deployed with more confidence than pure black boxes.

The large class of nonlinear models facilitates modeling more-complex relations among model parameters and data features. Biological systems, for instance, show highly nonlinear behavior. One category of nonlinear models involves starting with a nonlinear representation $\Phi(\boldsymbol{x})$ of the data, then building a linear model on top of the nonlinear representation rather than on the raw data themselves. This class of methods incorporates so-called kernel methods (16), whose main advantage over hand-designed feature maps is that any operation during learning and prediction can be rewritten in terms of a computationally cheap kernel evaluation, known as the kernel trick. An example of a kernel function is a Gaussian, but kernels can also be engineered with prior knowledge (16, 17). This property helps establish nonlinear versions of linear methods by "kernelizing" (18), leading to kernel support vector machines, kernel regression, kernel principal component analysis (PCA), and so forth. Note, however, that the higher predictive power obtained through nonlinearity comes at the cost of increasing the model's black-box nature. In Section 3 we describe various techniques to bring some level of transparency back into these more-complex models.

In contrast to shallow methods, which build on top of kernels or preextracted features, deep learning methods (19–21) follow the idea of learning the features relevant for a task from the data themselves via multilayered (deep) neural networks. This approach of learning the data representation via deep neural networks in combination with enormous quantities of data currently sets the state of the art in many domains (1, 2, 5) for different data types; recent breakthroughs have relied especially on self-supervised learning (4, 12, 22, 23). There exist different architectures that incorporate different assumptions, or "inductive biases" (24), about data representation; these include convolutional neural networks (CNNs) (1, 25), recurrent neural networks (RNNs) (26), and transformers (27). The parameters of these networks, which can number in the billions for modern deep networks, are trained mostly with variants of stochastic gradient descent (SGD) (9, 10), which iteratively update the network parameters following the gradient of the error function [computed via backpropagation (e.g., 28, 29)]. While deep neural networks were often referred to initially as black boxes, a lot of research has gone into making deep networks explainable, as we show in Section 3.

Lastly, powerful solutions for digital pathology often incorporate multiple ML models (shallow and deep) that are ultimately combined into composite models for patient characterization. For example, comprehensive characterizations of the tumor microenvironment (TME), which may serve as digital biomarkers in immuno-oncology, often involve deep models for cell detection and classification, tissue region segmentation, and so forth. Combining these models allows to derive informative and interpretable features such as the ratio, density, and infiltration of different cell types within different tissue regions. These features can then be fed into more classical models (e.g., linear regression models) to predict patient outcomes such as response to therapy or overall survival.

2.3. Validating and Improving a Machine Learning Model

Minimization of functions similar to that in Equation 1 enables the model to minimize the prediction error on the training data. Therefore, as a result of training, the ML model will tend to achieve accurate predictions on these data. However, there is no guarantee that the model will work accurately outside the training data, for instance, on future observations, which the ML model has not yet seen. The resulting gap in prediction accuracy is called overfitting. Overfitting is particularly severe in the context of high-dimensional molecular data, where a model can easily identify correlations in the available data (e.g., between the predicted cancer type and one of the 20,000-plus gene mutations available as input). These correlations on the training data are often spurious and do not generalize to new data. The problem only gets worse when nonlinear features

are added to the model. Overfitting results in models that work poorly on new data and do not offer many scientific insights because of their reliance on spurious correlations. Methods based on cross-validation (**Figure 1**), where the data are split into several parts, trained on some, and tested on others, have become the gold standard for detecting overfitting.

Another popular approach to avoid overfitting is to artificially augment the input data with small random perturbations to favor correlations that are more robust (and more likely to hold on new data). An approach that is particularly effective in practice is Monte Carlo Dropout (or Dropout) (93). Dropout trains the model while simultaneously applying noiselike perturbations in the input layer as well as in the multiple intermediate layers of the neural network. Monte Carlo Dropout has been applied broadly in biomedical applications ranging from models of molecular/omics data (e.g., 61, 94) to CNNs for histopathology (e.g., 95). Dropout layers are available in common neural network frameworks such as PyTorch (see **https://pytorch.org**) or TensorFlow (see **https://www.tensorflow.org**). Other sources of overfitting are mislabelings (e.g., a misdiagnosis, a measurement error, or a fault in the data preparation), which may artificially generate incorrect correlations in the data. Some mislabeling may be avoidable, but in other cases it is intrinsic to the data acquisition [ranging from falsely unexpressed genes in the context of single-cell RNA sequencing (96) to potential ambiguities in disease taxonomies]. A more direct approach to address this type of overfitting is through robustness-inducing loss or error functions. These robust loss functions are designed to tolerate more incorrect labels while yielding stable classification results. Common ML frameworks such as PyTorch and TensorFlow include a number of losses with or without robustness properties.

Another issue that can harm model performance and is often observed in the context of biological data arises from spurious correlations that occur systematically on the whole available data set, including the data reserved for testing (97, 98). Consider a scenario in which data from multiple hospitals are aggregated (e.g., different hospitals are equipped with different slide scanners). To solve the prediction task, the ML model might find it easier to recognize the distinct color signature of each slide scanner rather than the truly predictive biology. Techniques to reduce overfitting such as Dropout do not help in this case. Classifiers leveraging these systematic spurious correlations are commonly referred to as Clever Hans classifiers (89), in memory of the horse Hans, because they predict correctly but for the wrong reason. Such classifiers are, again, at high risk of becoming inaccurate on future observations. Continuing our biological example, the ML strategy of detecting the scanner's color signature may start to fail dramatically when a hospital renews its equipment or reorients its practice. Furthermore, a model relying on spurious features is of little interest if we want to extract scientific knowledge from it.

Unlike overfitting, Clever Hans effects cannot be properly detected using cross-validation (we present an alternative approach to detect such flawed classifiers in Section 3.3). However, the issue described above can often be avoided proactively by, for instance, an appropriate data set design or by introduction of prior knowledge. As an example, assuming metadata with regard to acquisition device, age of the subject, and so on are available, one can use a proper sampling strategy to ensure that the samples are stratified. This process can considerably reduce the reliance on spurious correlations. If the data set has already been collected and cannot be changed, one can instead try to induce the proper prediction behavior by various means, such as normalizing the input data (e.g., contrast/brightness normalization) or augmenting the training data with artificial color variations.

3. EXPLAINABLE ARTIFICIAL INTELLIGENCE

XAI (e.g., 99–102) is a major development in ML, driven by the need to make ML models more transparent and understandable to their users. The practical need for transparency in ML models

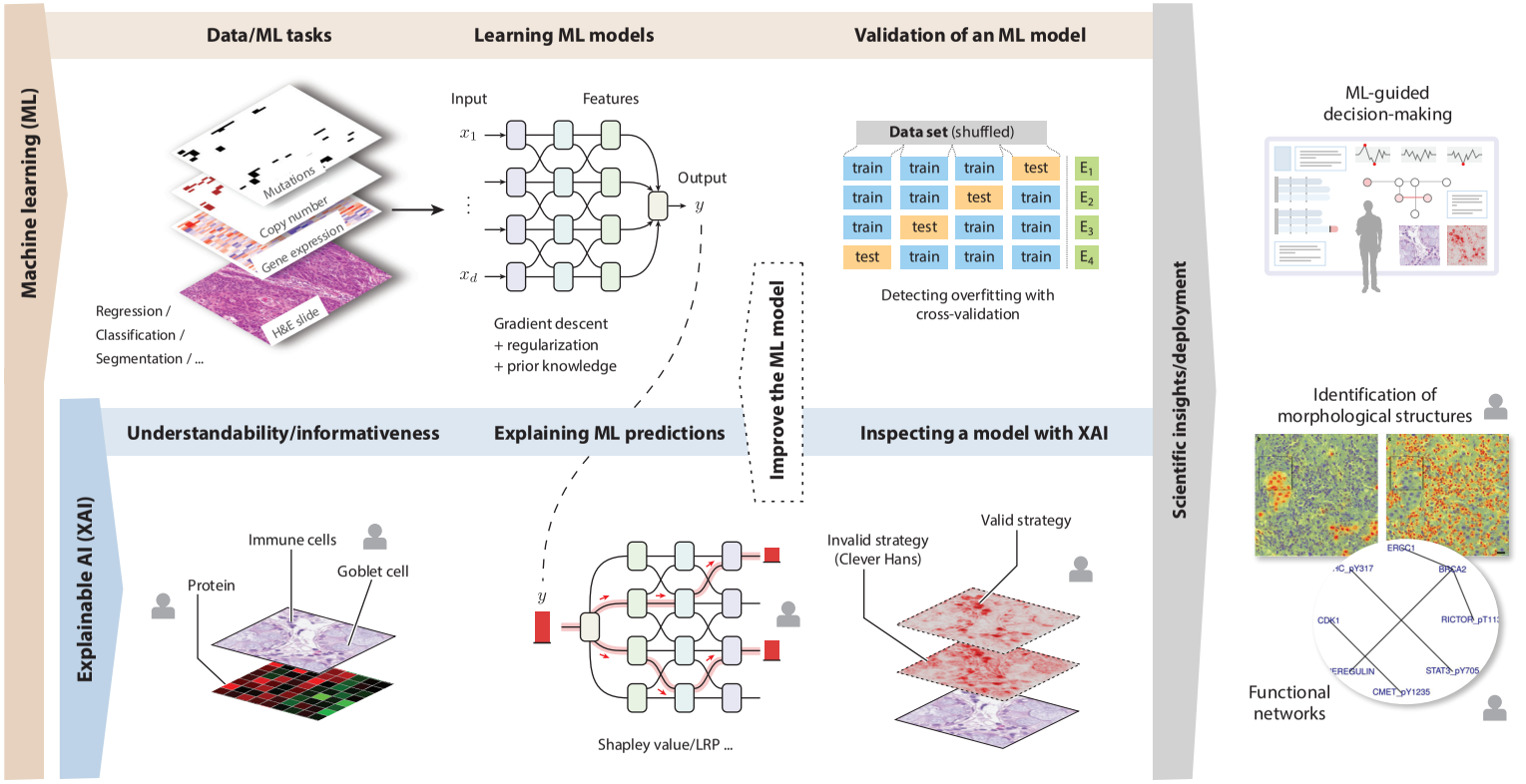


Figure 1

Overview of the typical ML/XAI pipeline. In ML, a prediction task is formulated from the available data. An ML model (e.g., a neural network) is trained to predict the task. After the model is trained and validated, it can be deployed to make decisions, assist the decision-making process, or extract insights. XAI enhances the transparency of the ML pipeline. Data are prepared so that features (e.g., images) are interpretable. The ML model is equipped with explainability so that its predictions can be explained in terms of input features. Explanations can be used to further verify the correct behavior of the model as well as to generate deeper insights or scientific hypotheses about the input–output relation learned by the ML model. Gray icons of people indicate that a user looks at the data or visualization for explainability reasons. Abbreviations: LRP, layer-wise relevance propagation; ML, machine learning; XAI, explainable artificial intelligence.

in medicine has been acknowledged many times (7, 103). Motivations include verifying that the ML model uses a valid/well-generalizing decision strategy (see Section 3.3) and gaining more insight into the data being predicted and thereby increasing scientific knowledge. While classical ML models can provide basic insights, such as testing whether one data modality (e.g., cancer type) is predictable from another (e.g., histological images), XAI can generate much deeper insights, uncovering the exact input features (e.g., pixels or their nonlinear correlations) that are used by the ML model for prediction. In this section we focus on the technical aspects of XAI (for a systematic overview and references, see also **Table 1**); we return to medical aspects in Sections 4 and 5.

3.1. Understandability and Informativeness

A common formulation of the problem of explanation is to map a data point $\boldsymbol{x} \in \mathbb{R}^d$ and its prediction by the function f into a collection of scores $(R_1, \ldots, R_d)$, highlighting the relevance of each input feature to the prediction (62, 65, 77):

$$(\boldsymbol{x}, f) \mapsto (R_1, \ldots, R_d). \qquad 2.$$

Before entering into the technicalities of how to produce such explanations, it is essential to ensure that the explanation being generated is understandable and informative for the human—that is, it enables the user to identify and correct a flawed ML model or arrive at new scientific knowledge in a data-driven manner.

First, for an explanation to be understandable, the meaning of individual input features on which the explanation is based should be clear to the human. When the data have a tabular structure (e.g., showing for each subject an array of protein expressions, gene mutations, or clinical features), ensuring that the user has the expertise to interpret these features is a prerequisite. An explanation can then take the form of a bar plot, where each bar represents a particular input feature and the bar length its contribution to the prediction. Another common type of data is image data (e.g., in histopathology). In this case, input features received by the model are pixels, which, unlike tabular data, rarely carry meaning on their own. For this type of data, explanations are better rendered as a highlighting of the input image (or heatmap), enabling the human to identify which visual pattern the model has used for its prediction. **Figure 1** depicts some heatmap-based explanations. At the interface between visual data and tabular data are efforts to generate dictionaries of human-interpretable concepts that the network uses for predicting (e.g., midlevel visual features) and that can be used to support an explanation (69, 70). Lastly, identifying the contribution of individual features to the prediction may be of limited use, and one may instead want to determine how features interact with other features to arrive at the prediction (68, 70, 72).

Second, it is important to pay attention to the precise question the explanation answers. For example, asking what makes the output neuron associated with class 1 activate differs from asking what makes this image predicted to be of class 1 rather than another class, and so should the associated explanations (104). Another subtle difference is to ask what makes an image $\boldsymbol{x}$ of class 1 whereas another image $\tilde{\boldsymbol{x}}$ is predicted to be of another class. The latter question is addressed by counterfactual explanations (73, 74). In the context of molecular data, a counterfactual example may be a vector of mutations similar to the original vector, without the few mutations that cause the model to predict cancer, or a histopathological image without the region containing the cells relevant for the prediction.

3.2. Techniques of Explanation

We now focus on the question of how to explain models technically, knowing that most ML models used in practical applications are complex and highly nonlinear (e.g., deep neural networks). We

distinguish between two general families of explanation techniques: model-specific and model-independent (for discussion, see 102, 105, 106).

The model-specific (i.e., self-interpretable) approach imposes a predefined structure to the ML model, so that an explanation can be easily extracted. For example, when the prediction function is of the type $f(\boldsymbol{x}) = \sum_{i=1}^{d} g_i(x_i)$, where $g_1, \ldots, g_d$ can be any nonlinear functions, an attribution to the input features can be easily extracted by taking the summands (71, 75):

$$R_i = g_i(x_i). \tag{3.}$$

Other structures that enable a limited form of interpretability include decision trees, until a depth of three or four layers (72, 107). Beyond a certain depth, they are no longer interpretable by humans.

When such restrictions on the structure of the model are not practical (e.g., because high prediction accuracy or low run time is of primary concern), one usually resorts to a model-independent (i.e., post hoc) approach, which makes no assumptions about the function f. An intuitive way of extracting model-independent explanations is to measure the effect on the prediction of adding/removing input features. For example, one can compute

$$R_i = f(\boldsymbol{x}) - f(\boldsymbol{x}_{-i}). \tag{4.}$$

Here, $\boldsymbol{x}_{-i}$ denotes the data point $\boldsymbol{x}$ where feature i has been removed (e.g., set to zero or replaced by the corresponding feature of the counterfactual $\tilde{\boldsymbol{x}}$). Many variants of feature removal techniques have been proposed (for a review, see 78). A popular formulation with good theoretical properties is the Shapley value (77, 79), which considers multiple joint feature perturbations and weighs them appropriately. Note that perturbation approaches in general require the function to be evaluated many times (typically at least once per input feature), which can significantly slow down the generation of explanations and make it impractical for large models.

Layer-wise relevance propagation (LRP) (65) is another post hoc explanation approach that addresses the scaling issue by leveraging the underlying sequential (deep-layered) structure of the prediction function. LRP starts at the output of the model and backpropagates the prediction layer by layer until the input features are reached. A simple propagation rule between two layers is

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_j a_j w_{jk}} R_k, \tag{5.}$$

where j and k are indices of neurons in the two consecutive layers, a_j denotes the activation of neuron j, and w_{jk} is the weight connecting neuron j to neuron k. Different propagation rules have been designed to address layers encountered in specific architectures such as CNNs (104), long short-term memory models (66), and transformers (108).

At the frontier between model-specific and model-independent approaches, we find other types of XAI methods. These include methods such as LIME (local interpretable model-agnostic explanations) (82) and SHAP (Shapley additive explanation) (79), which build a readily interpretable local surrogate of the original ML model, and methods that regularize the ML model to ease the process of post hoc explanation (88).

3.3. Inspecting a Model with Explainable Artificial Intelligence

In Section 2.3, we note that for a model to perform well on new data, it is critical to verify that the model neither overfits nor uses spuriously correlated features (e.g., predicting on the basis of the scanner's color profile or other types of artifacts). Use of spurious features (so-called Clever Hans effects) is hard to detect on the basis of classical cross-validation techniques; however, XAI can systematically unmask Clever Hans effects (89, 90). The reliance of the model on a spuriously

correlated feature is indeed readily highlighted in the explanation. Several XAI-based approaches have been proposed for systematically uncovering and removing Clever Hans effects from large neural network models (e.g., 89, 90). A worked-through example illustrating the process of detecting and removing Clever Hans effects in a histopathology context is provided in Section 6.1.

XAI also has a role in imparting further scientific insight. Given an ML model trained to predict some complex, nonlinear biological system of interest, XAI can extract from the model biologically meaningful network structures (e.g., related to cells' signaling pathways) that go beyond what can be achieved with common statistical/bioinformatic approaches in terms of resolution and overall accuracy (48). A use case of XAI for such scientific purposes is provided in Section 6.2. Further applications of ML/XAI in pathology are presented systematically in Section 7.

4. COMPUTATIONAL PATHOLOGY

Computational or digital pathology aims to create and implement tools that provide assistance in the conventional clinical pathology workflow. This workflow consists of three phases: preanalytical (sample collection, accessioning, specimen preparation, grossing, tissue and slide preparation), analytical (interpretation), and postanalytical (report preparation, transmission). At all levels, there are specific tasks, including associated error sources, that could benefit from computational assistance (109). The availability of slide scanners that produce WSIs is the precondition for pathology to enter the digital age as well as to benefit from image analysis techniques in computer vision (110), notably deep learning. Open source software like QuPath (111) facilitates preprocessing (annotation) of the digitized slides. Libraries like TIAToolbox (112) allow easy module integration of common subtasks, namely reading WSI data, patch extraction, stain normalization and augmentation, model inference, and visualization.

4.1. Artificial Intelligence for Histological Data Analysis

With its ability to classify complex unstructured data, AI is highly suitable for histological image analysis. The following subsections review strongly and weakly supervised as well as unsupervised learning approaches.

4.1.1. Strongly supervised learning in histopathology. AI is being used successfully in various histological image analysis steps (113). The classical application is to replicate tasks performed by humans, with a side effect of higher reproducibility, such as in cell detection (114) and tissue segmentation (2). For tumor diagnostics, features of different cells from the TME with tumor content, the cytological features of the tumor cells, and the cells' spatial relationships can be combined. Such morphological profiles can then be either further analyzed by conventional (spatial) statistics or again used as input for training ML models to predict (clinical) endpoints. A detour via established morphological features can reduce dimensionality and include prior knowledge and/or hypotheses on the relevance of certain tissue properties. A disadvantage of this fully supervised approach on strongly annotated data, however, is the need to generate annotations for training down to the single-cell level—an expensive and cumbersome task.

4.1.2. Weakly supervised learning and explainable artificial intelligence. An alternative approach is the direct prediction of clinical endpoints or molecular properties from histological images (i.e., end-to-end learning). Here, labels are required only at the slide level, so-called weak annotations (e.g., survival), in contrast to pixel-wise strong annotations (which require more cases for training). Another limitation of conventional ML that applies to both approaches, but particularly to end-to-end learning, is the inability to understand the AI decision process. This so-called black-box characteristic of ML limits the user's ability to understand the AI-based

decision process. Recently, XAI has been used to highlight the image regions that contribute the most to the AI's decision in form of heatmaps (56, 91). Unlike conventional heatmaps, which simply visualize the resulting classifier scores for an image patch, XAI-generated heatmaps offer more fine-grained information in the form of pixel-wise scores even when (only weak) annotations exist only at the slide level (56, 65, 91).

While cell detection and tissue segmentation approaches can achieve high accuracy on par with that of pathologists, the accuracy of predictions of molecular markers from histology still falls behind that of sequencing techniques. Whereas the detection of, say, oncogenic mutations can be achieved with state-of-the art sequencing at an accuracy of nearly 100%, ML-based prediction reaches an area under the curve (AUC) of no more than 70–80% for the best cases (56, 115). This discrepancy may not be surprising given that the complex molecular states governed by genes, genetic regulatory networks, and ultimately proteins are integrated at the histological level. To be predictable by AI, molecular features need to have a strong impact on cellular morphology. An interesting example of a relatively simple-to-predict molecular feature is microsatellite instability (MSI), which is predictable from histology with a good AUC of more than 80% (115). MSI is related to a deficiency in mismatch repair, resulting in a high mutational load and high neoepitope generation that, in turn, cause an inflammatory response in the TME. This response provides relatively clear morphological clues in the form of massive lymphocyte infiltration. Most current models used for the prediction of molecular properties rely on MIL, and approaches using support vector machines show similar accuracy (56). A challenge with these models is that their generalizability is still limited, as the models are trained mostly with single-institution data sets or single-clinical-trial data.

In the future, novel ML approaches may offer improvement by allowing users to incorporate knowledge from more data, including decentralized learning or foundational models. While by definition they are never as effective as centralized training, federated or swarm learning approaches (116, 117) can be used to extend the database by working around regulatory or data protection limitations by leaving data at their respective locations and training the AI models in a distributed fashion (i.e., sharing only the trained weights θ).

4.1.3. Unsupervised learning. In contrast to supervised learning, unsupervised learning can help extend the database by allowing the model to learn from all data, not just annotated data (i.e., no label needed). This can expand the accessible data by orders of magnitude and can lead to breakthrough results in current natural language processing applications like ChatGPT (118). The approach works by imposing a task, such as masking an image patch that needs to be reconstructed, and requires the trained model to understand complex concepts while not requiring any human input, for example, certain tumor types showing a prominent desmoplastic stroma reaction. The resulting so-called foundation or pretrained model can then be refined to solve a downstream task like predicting survival. If what is learned from general domains like text understanding can be translated to medicine and pathology, then, given enough data, such techniques could enable solutions to a complex task like performing a continuous text-based diagnosis as a pathologist would perform it.

Although unsupervised expansion of the training data allows for a complex understanding of the model, a final, supervised step is typically needed to unlock the model's power for applications such as classification, regression, or segmentation; this step thus poses some limitations. One such limitation is the prediction of rare disease, for which it is hard to gather enough annotated examples to train any ML model. This long tail of disease can be approached by anomaly detection (119, 120), an ML concept that revolves around detecting data that are not similar to most of the data at hand and, therefore, allows the model to train on naturally distributed data and promises

to flag data that are uncommon to it. This orthogonal approach is fundamental to AI in real-world scenarios in order to catch potential fail cases that should be referred to human experts.

Concepts like decentralized learning, foundation models, and anomaly detection are introduced here in the context of histopathology. In principle, however, they can translate to any other biomedical domain and beyond.

In light of the limited clinical impact of computational pathology, and despite some studies' claims of "clinical-grade" computational pathology (121, 122), there is still a lack of solid clinical validation (115, 123, 124). Also, aspects of the pre- and postanalytical workflow, including the varying depth of digitization and implementation of structured reporting, slow down clinical adoption.

4.2. Artificial Intelligence for Molecular Data Analysis

While AI applications in pathology have so far focused mostly on histomorphological images, as described in the preceding section, the increasing use of molecular profiling approaches (ranging from single-gene mutational analysis to omics technologies in diagnostics and research down to the single-cell level) in combination with increasingly structured clinical data requires novel computational techniques capable of dealing with complex heterogeneous data. In the future, XAI methods may be used not only to analyze the different data modalities but also to help integrate and interpret those data through multimodal learning concepts. In this section, we review current AI approaches for the analysis of different omics data modalities, for both bulk tissue and single-cell data, and provide an outline of what exists and what is expected regarding multimodal learning approaches.

4.2.1. Molecular data and artificial intelligence for treatment guidance. While the use of AI in clinical practice is still under development, it has already demonstrated the potential to support the use of molecular data for precision pathology. Significant milestones have been achieved, including the improved identification of single-nucleotide polymorphisms (SNPs) through the use of neural networks (50, 125). Additionally, ML models like neural networks and random forests can predict cancer prognosis or response to therapy on the basis of mutations, copy number variations, and RNA-sequencing data (e.g., 76, 126). For a more detailed review of AI approaches in these areas, we refer readers elsewhere (127).

4.2.2. Artificial intelligence in DNA methylation–based tumor diagnostics. DNA methylation is a so-called epigenetic modification. When located in a gene promoter, DNA methylation usually leads to transcriptional repression (128). It is well known that global DNA methylation signatures are tissue-specific, defining the expression profile of different cell types according to their individual function (129). Although DNA methylation is altered in virtually all types of cancer, the global DNA methylation profile of malignant cells still contains substantial information about their cells of origin (130). Currently, DNA methylation analyses are performed mostly using an array-based method that measures approximately 850,000 CpG sites (131). Such analyses can be performed on fresh-frozen or paraffin-embedded formalin-fixed tissue without major batch effects, enabling the simultaneous analysis of different data sets (41).

Due to their high complexity, analyses of DNA methylation data require ML-based approaches. Unsupervised dimensionality reduction and clustering methods can be used to define epigenetic classes, which were the basis for the definition of several new tumor entities and subtypes (132–135). Furthermore, ML-based classification algorithms developed for tumor classification outperformed conventional approaches as well as other molecular methods (40, 41, 94, 136). Most importantly, the DNA methylation–based Heidelberg Brain Tumor Classifier has become the gold standard for the diagnostic classification of several brain tumor entities and is

one of a few ML-based methods that are used directly in the clinic (41) and are included in the current World Health Organization classification (137).

Beyond their diagnostic utility, DNA methylation data can be used for prognosis prediction or the estimation of tumor-infiltrating lymphocytes (138–141). Due to their application in the two most popular algorithms (41, 136), random forests are the most commonly used techniques. However, several studies have shown that other techniques, such as support vector machines or neural networks, can clearly outperform random forest classifiers in specific applications (40, 94, 142).

Currently, there are no well-established XAI methods for unsupervised or supervised ML analyses of DNA methylation data. For classification tasks, the relevance of CpGs can in principle be computed; however, their biological interpretation is difficult. Specific XAI methods are therefore needed for validation and for a better understanding of DNA methylation–based predictions, which will be key for wider clinical deployment of these approaches.

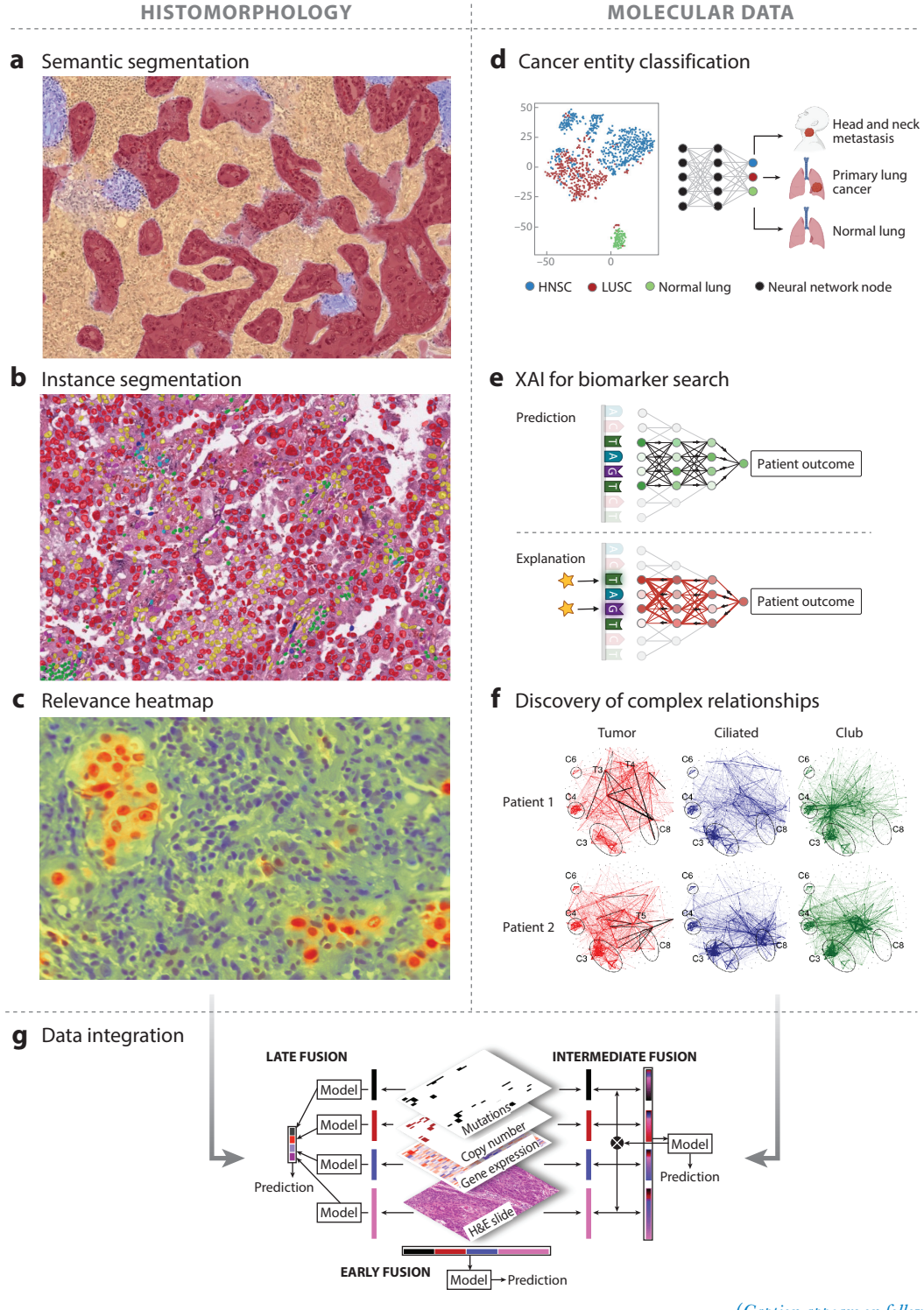
4.2.3. Molecular marker discovery in precision pathology. Precision pathology aims to find treatment-guiding patient-specific characteristics. In cancer diagnostics, great progress has been made in characterizing patient subgroups that are predictive of the efficacy of targeted therapies. Prominent examples are the identification of HER2/neu expression as a biomarker for trastuzumab treatment response in breast cancer (143) and $BRAF^{V600E}$ mutations predictive of vemurafenib efficacy in melanoma (144). A recent breakthrough was the identification of a patient group suffering from advanced colorectal cancer with mismatch repair defect (145) that went into full remission after targeted treatment with a PD-L1 inhibitor.

The identification of molecular biomarkers holds great potential for precision pathology, but experimental investigations are limited given the complexity and heterogeneity of disease mechanisms in different patients. ML approaches can help by modeling the complex associations between disease and biomarker candidates for large sample numbers. Below, we outline how XAI can leverage these models to extract underlying biological mechanisms and identify key molecular features.

Based on an ML model that predicts a certain clinical endpoint from molecular data, XAI determines the most important markers for this prediction (**Figure 2*e***). Chen et al. (76) predicted the risk of cancer death on the basis of imaging and molecular profile data and explained these predictions by an integrated gradient approach. This approach allowed them to find prognostic biomarkers such as *IDH1* mutations in low-grade glioma. Model-agnostic XAI methods like SHAP (79) have been applied to explain survival predictions of different models, such as random survival forests, survival support vector machines, and gradient boosting (44, 47, 146). These studies identified prognostic biomarkers for colorectal, pancreatic, and breast cancers. Kim et al. (54) used an RNN to identify factors that determine progression from atrophic gastritis to gastric cancer.

Another common application of XAI is searching for molecular markers that identify cancer subtypes or distinguish metastases from primary tumors. Here, established knowledge about cancer subtype behavior (e.g., with respect to treatment) is linked to newly identified markers (e.g., epigenetic and transcriptomic data) (83–86). While these approaches have found markers for a known outcome, even unsupervised methods have been combined with XAI to find multiomics markers that cluster cancer patients into distinct molecular subtypes (87).

4.2.4. Using higher-order explainable artificial intelligence to include prior knowledge for marker identification. Molecular profiles screened by XAI for informative markers can encompass tens to hundreds of thousands of molecular features. In most cases, this search space is too large for the reliable identification of biomarkers, and prior knowledge may be used to restrict the model. Several existing approaches integrate such functional information in the form of biological



(*Caption appears on following page*)

Figure 2 (*Figure appears on preceding page*)

AI applications in precision pathology based on histology and molecular data. (*a*) Semantic segmentation of histological slides distinguishes cancer (*red*) from healthy tissue (*yellow*) and necrosis (*blue*). (*b*) Instance segmentation identifies cancer cells (*red*), fibroblasts (*yellow*), lymphocytes (*green*), and others. (*c*) XAI heatmap visualizes relevance of each pixel for the prediction "cancer." Low (negative) relevance is shown in blue and high relevance in red. (*d*) Neural networks distinguish primary lung cancer, normal lung tissue, and head and neck metastases on the basis of DNA methylation profiles. (*e*) Biomarker identification with XAI. An outcome of interest is predicted on the basis of genomic data. XAI then determines the most relevant single-nucleotide polymorphisms for this outcome. Red color intensity and line width indicate the relevance of neural network nodes for the prediction. (*f*) XAI for a single-cell gene regulatory network prediction based on single-cell tumor RNA-sequencing data identifies tumor-specific (T; *black lines*) and normal (C; *ellipses*) network modules. (*g*) Architectures of multimodal modeling. Abbreviations: AI, artificial intelligence; HNSC, head and neck squamous cell carcinoma; LUSC, squamous cell lung carcinoma; XAI, explainable AI.

networks. Chereda et al. (60) used known protein–protein interactions and gene expression data to model the metastasis of breast cancer with a graph-CNN. They then applied an LRP variant, GLRP, to identify relevant subnetworks for the prediction of breast cancer metastasis, and they even reported actionable genes. Pfeifer et al. (59) integrated a protein–protein interaction network to identify relevant disease network modules. Schulte-Sasse et al. (58) applied a graph-CNN to multimodal data, including single-nucleotide variants, copy number alterations, gene expression, and DNA methylation, as well as protein–protein interaction network information, for the prediction of cancer genes. Using LRP, they determined how different data modalities contribute to such predictions. Bourgeais et al. (61) integrated gene ontology information with gene expression data to leverage cancer detection and used a self-explainable network to find cancer-specific gene ontology functions.

4.2.5. Discovery of functional interactions by explainable artificial intelligence. Network inference methods for the prediction of molecular networks can help uncover functional relationships in omics data (reviewed in 147–150). Here we focus on ML/XAI for network predictions for single-cell tumor RNA-sequencing data. Even small patient cohorts can provide sufficient training data; thousands of sequenced cells per tumor offer a unique view of intratumoral heterogeneity, which is important for understanding clonal evolution and resistance to therapy (151, 152).

Based on an ML model trained to predict the expression of a gene by other genes, XAI can determine the most relevant genes for that prediction (49, 153), not only for specific cell types but also for individual cells (**Figure 2*f***). The regulatory relationships have been modeled using, for instance, random forests (153) and gradient boosting (154). The single-cell gene regulatory network prediction identifies networks specific to cancer cells, information that may be used to understand dysregulation and identify potential drug targets.

5. MULTIMODAL DATA INTEGRATION

The diagnosis and risk stratification of a tumor disease, as well as the evaluation of treatment success, are based on a multitude of clinical and diagnostic assessments. These include patient characteristics, clinical examination, histopathological and molecular characterization of the tumor, blood work, imaging (computed tomography, magnetic resonance imaging, ultrasound), and patient-reported outcomes. Methodologies for multimodal data integration can use these digital biobanks (155) from cancer patients to develop data-driven biomarkers. In cases where the different data layers represent orthogonal information, they promise to yield better predictive performance than the unimodal model. The prospect of combining molecular data with histopathology predates AI-based image analysis. In this section, we focus on important aspects of processing data from histopathology and molecular analyses and present an overview of architectures for multimodal data integration.

5.1. Feature Selection in Molecular Data

The combined molecular data of a tumor (methylome, genome, transcriptome, proteome) can exceed a million data points, making feature selection essential for successful multimodal integration in cohorts of usually a few hundred patients. To account for the long tail of rare mutations in cancer, cutoffs for the recurrence of a mutation or copy number alteration within the cohort enable a straightforward approach to feature selection [e.g., 5–10% (76, 156, 157)]. In contrast, feature engineering in transcriptomic data is more challenging. Curated gene sets [e.g., from the Molecular Signatures Database (158)] have been used to select gene expression features (76, 156) but, without further filters, also include genes with low expression or low variation between samples and likely have no meaningful representation of the transcriptome. Ultimately, a large fraction of the variance observed in the transcriptome should be retained after feature selection. The expression of a gene is dependent on other genes in the gene regulatory network; therefore, one promising approach is to represent gene expression by interactions [e.g., by using a sparse graph-CNN (159)] rather than as a vector of single-gene expression levels.

5.2. Architecture: Early and Intermediate Versus Late Fusion

Multimodal models differ by time of integration (**Figure 2*g***). In an early-fusion architecture, the selected features from the individual data layer are concatenated and serve as joint model input. Late-fusion architecture models every data layer individually and then fuses the learned parameters at the end. In both early- and late-fusion architectures, the unimodal embeddings are not affected by the embeddings from other data layers. In contrast, in an intermediate-fusion architecture, feature representations of the unimodal data are iteratively improved by backpropagation from the multivariate model. For a more extensive overview of model architectures, we refer readers to excellent reviews in this area (160, 161). We envision that XAI will be particularly powerful when used in an intermediate-fusion architecture, enabling the assessment of links between morphological and molecular cancer properties (56).

6. WORKED-THROUGH EXAMPLES

In this section, we work through two specific, real-world examples to demonstrate the modeling and evaluation process and provide some best practices.

6.1. Example 1: Cancer Classification From Images

Deep learning models are in principle able to capture morphological features in histopathological images in order to, for instance, differentiate between cancer and noncancer. However, developing models that perform robustly in a routine clinical setting remains a difficult task. In this example, we highlight common pitfalls and best practices for training a deep learning model to classify colorectal cancer tissue. Our example also outlines some of the specific issues encountered in pathology AI and how best to address them.

We use the NCT-CRC-HE-100K patch data set (162) to differentiate among nine different colon tissue types. The data set consists of 100,000 annotated patches from two different source sites without any color normalization. We split the data set into 50% training, 25% validation, and 25% testing.

Let us start by training a CNN classifier on the training split. We use the standard ResNet18 (25) architecture with 11 million parameters to balance computational requirements and model capacity. We use a collection of functions built on top of the PyTorch framework to train and test

our model (the code can be found at **https://github.com/jonasd4/pathology-worked-through**). As an objective, we use the categorical cross-entropy loss and optimize our model with SGD (**Figure 3*a***, **subpanel *i***). To prevent overfitting (see Section 2.3), we make the usual recommendation to use regularization, which aims to limit model complexity and therefore supports generalizability to new data.

6.1.1. Regularize the machine learning model. We use a combination of regularization techniques, including weight decay. We augment the images (random horizontal and vertical flipping, random crops) and choose the regularization parameters that lead to the lowest validation loss. We then measure the test error on a separate test set, which we have used neither for training the model nor for selecting the regularization parameters (that were chosen with the validation data set):

```
model = resnet18 ()
model . fit (training_data)
model . predict (training_data)
# accuracy: 99.54% (good)
model . predict (validation_data)
# accuracy: 99.24% (good)
model . predict (test_data)
# accuracy: 99.19% (good)
```

6.1.2. Validate the findings on an external cohort. The classifier achieves an accuracy of 99.2% on the test set. As a further verification experiment, we now test our model on an external test set (CRC-VAL-HE-7K) consisting of tissues collected at different source sites and on different patients. The external test set contains 7,180 image patches from 50 patients:

```
model . predict (external_test_data)
# accuracy: 47.89% (bad)
```

Unlike the original test set accuracy, we observe a poor performance on this external test. The classifier achieves an accuracy of 47.9%, showing that it does not generalize well to new data. Not performing this additional experiment would have resulted in an overoptimistic assessment of the model performance and would have made the practical use of this model quite hazardous, hence our third recommendation, below.

6.1.3. Inspect the machine learning model with explainable artificial intelligence methods. To understand the poor performance of the classifier on the external test set, we apply XAI by computing LRP heatmaps with the Zennit framework (163). The resulting heatmaps highlight the contribution of individual pixels to the model's prediction (**Figure 3*a***, **subpanel *iii***). We notice that our model does not seem to focus on the relevant image features. Instead, larger regions of the image are marked as equally relevant. This suggests that the model relies on a Clever Hans–type strategy, that is, a strategy that exploits spurious correlations in the available data but fails to generalize on instances outside those data.

6.1.4. Use stain normalization and data augmentation. Reliance on color distribution rather than the more predictive geometrical shapes is a common Clever Hans strategy that is usually much easier to learn by a model. To understand the emergence of the Clever Hans strategy, we conduct a simple additional experiment wherein we represent each image by its average color (a vector containing the pixel-wise mean of the three color channels). **Figure 3*a***, **subpanel *ii***

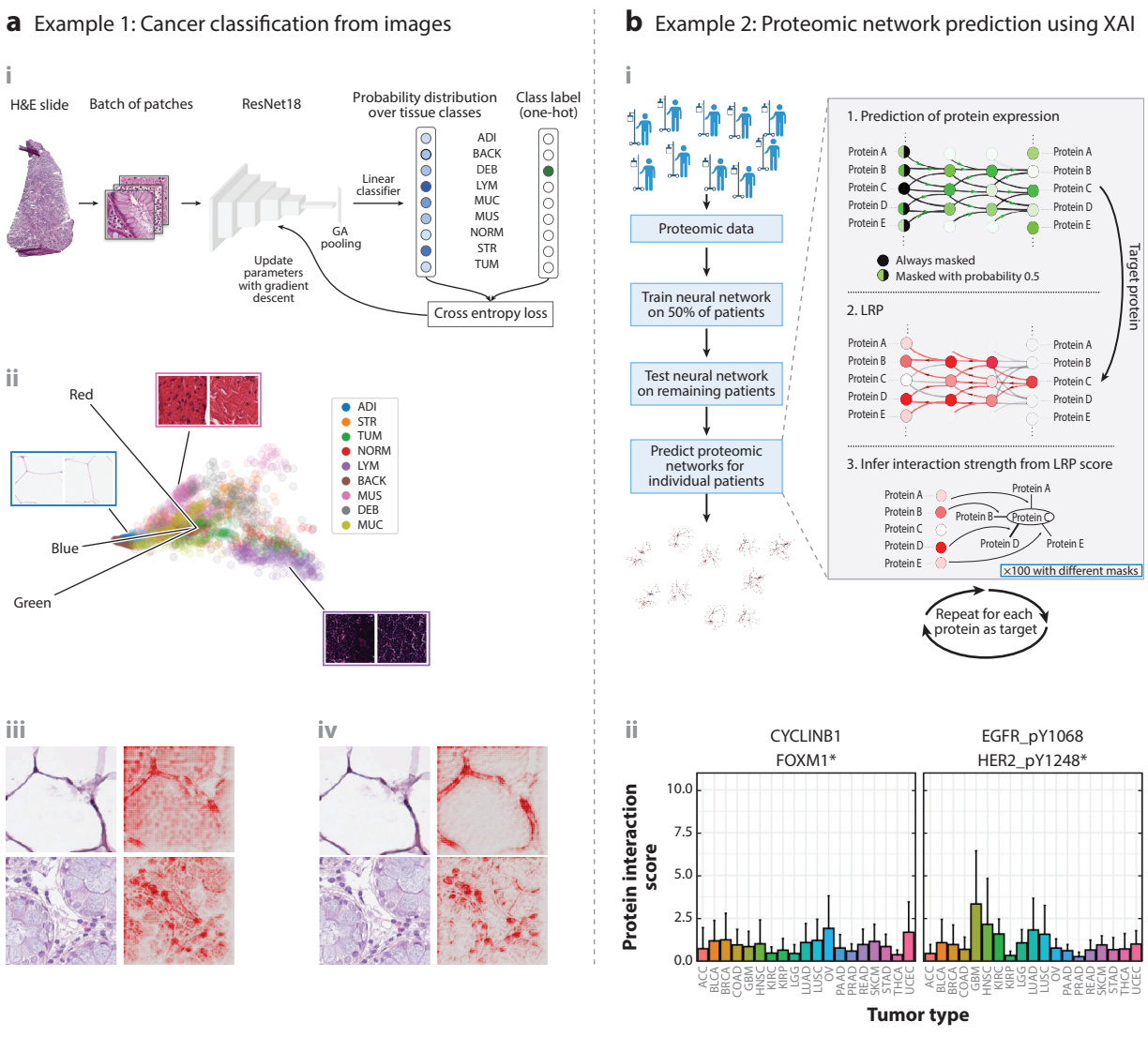


Figure 3

Depiction of our two worked-through examples. (*a*) (*i*) Training pipeline for the image analysis example. (*ii*) A principal component analysis plot of the training image mean colors shows that color information is correlated with class labels. (*iii*) LRP explanation heatmaps of the simple model. Large areas of the image are highlighted in the relevance map, which indicates a bias toward the image color. (*iv*) LRP heatmaps of the improved model. By applying stain normalization and color augmentations during the training, we reduce the color bias of the model. (*b*) (*i*) Enhanced ML/XAI approach to reconstruct protein–protein interaction networks for individual tumors (49). (*ii*) Examples of protein interaction scores resolved by cancer type for pairs of functionally related proteins. Abbreviations: ADI, adipose; BACK, background; DEB, debris; GA, global average; H&E, hematoxylin and eosin; LRP, layer-wise relevance propagation; LYM, lymphocytes; ML, machine learning; MUC, mucus; MUS, smooth muscle; NORM, normal colon mucosa; STR, cancer-associated stroma; TUM, colorectal adenocarcinoma epithelium; XAI, explainable artificial intelligence. Data from Reference 48.

Table 2 Model accuracy on the first cohort's test set (NCT-CRC-HE-100K) and on the external cohort (CRC-VAL-HE-7K)

Model	Color augmentation	Stain normalization	First cohort	External cohort
Baseline			99.19	47.89
+ color augmentation	×		98.61	87.91
+ stain normalization		×	98.80	93.12
Improved	×	×	98.14	94.09

shows a PCA biplot visualization of the data set in this simple 3D space. For some of the classes, the color-based representation is highly predictive of the tissue class, and a *k*-nearest neighbor classifier built on this simple representation achieves 78.2% accuracy on the test set.

Now that we have identified the problem with XAI and our color analysis, we want to improve our training procedure by preventing it from relying solely on color information. Therefore, we use stain normalization (164) as a preprocessing step and introduce color augmentations. We use color jitter, randomly equalize the color histogram of the image, and randomly make the image grayscale. The improved model achieves an accuracy of 94.1% on the external test set. **Table 2** shows the individual contributions of the color augmentations and stain normalization on model performance.

We observe that both stain normalization and color augmentation have a strong positive effect on the external test set performance. By inspecting the LRP heatmaps of the improved model, we can see that the model focuses less on the overall background of the image and more on class-discriminating features (**Figure 3*a*, subpanel *iv***). We stop our analysis here but note that applying XAI is an iterative process: Further inspection of the model might yield further improvements. Notably, simple pixel-wise attribution techniques such as standard LRP might not always be the best way to identify flawed decision strategies, and one may benefit from enriching the explanation with counterfactuals or latent representations (see Section 3).

6.2. Example 2: Proteomic Network Prediction Using Explainable Artificial Intelligence

In this application example, we highlight how XAI can predict functional protein interactions from large proteomics data sets that are difficult to interpret functionally with conventional bioinformatics approaches. We follow the main steps from Reference 48, apply them to data from 5,144 patients across 19 cancer types with 147 proteins, and validate our predictions with data from the Reactome Pathway Database (165), a curated resource for protein interactions.

6.2.1. Prediction of protein interactions with explainable artificial intelligence. Here, we demonstrate how to predict protein interactions using XAI (48) (**Figure 3*b*, subpanel *i***). The data are first split into a training set and a test set. We train a fully connected neural network on the training set to impute proteins that are masked in the input with a probability between zero and one. This masking task procedure can also be regarded as a variant of the dropout method; consequently, it has a beneficial regularization effect on the neural network.

Once the neural network has been trained, we compute predictions over 100 randomly generated masks (where proteins are masked with probability 0.5), which are then attributed to the proteins at the input of the neural network. Averaging over the 100 samples yields a matrix of protein–protein scores for each patient. We can interpret these matrices as individualized measures of protein interactions. The direction of interaction can be ignored by applying the absolute value and adding the transpose of the matrix.

Note that our ML/XAI approach differs from a classical correlation approach in the following ways:

1. Instead of one global (average) interaction matrix, we compute one matrix for each individual patient.
2. The type of modeled interactions is no longer restricted to linear or monotonous but can be any form of interaction learned by the ML model.

Let us now test whether the interactions predicted by ML/XAI match our ground truth. As with the correlation-based approach, we generate a global protein interaction matrix by averaging patient-specific interaction matrices. We find that the ML/XAI approach, where 56 of the strongest 100 predicted interactions are confirmed by Reactome, outperforms the standard correlation approach, which identifies only 40 correct interactions.

6.2.2. Predictions for individual patients. An advantage of the XAI approach used here is its ability to make predictions for individual patients, instead of the population average made with standard statistical approaches. This can help reveal differences among patients with one cancer type and/or across different cancers (**Figure 3*b***, **subpanel *ii***). The analysis indicates that the functional relation between proteins depends strongly on the cancer type and may help formulate hypotheses about proteomic network differences among cancers.

7. A SYSTEMATIC VIEW OF MACHINE LEARNING AND EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR PATHOLOGY

In this section, we aim to provide a systematic view of possible uses of ML/XAI in pathology. We identify three major prototypical use cases: (*a*) assisting a human decision-maker, (*b*) building autonomous decision systems, and (*c*) extracting scientific insights (**Figure 4**).

7.1. Machine Learning and Explainable Artificial Intelligence as a Personal Assistant

Human decision-making is a resource-intensive process. Especially in pathology, manual evaluation of multimodal diagnostic data [(immuno)histology, mutational profiles, clinical data] is

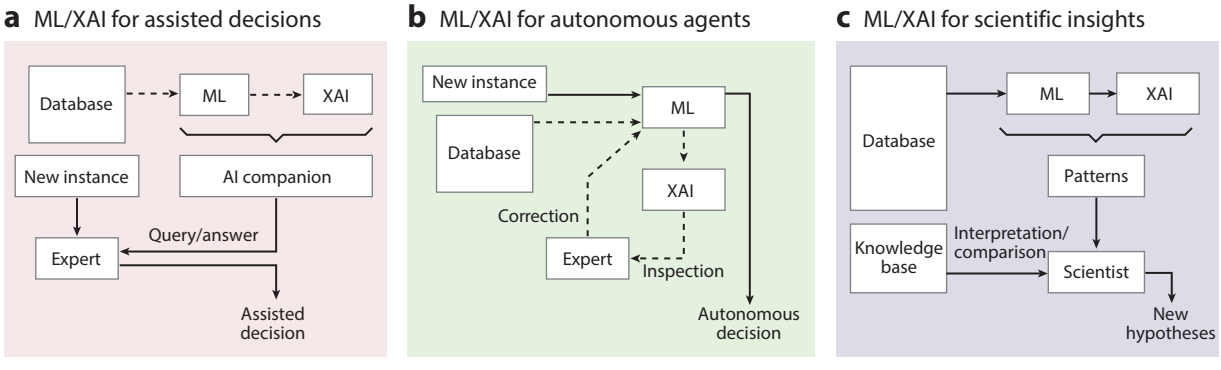


Figure 4

Overview of practical uses of XAI that are relevant in biomedicine. XAI can be used to (*a*) assist the medical expert in making decisions, (*b*) validate and improve a decision system, and (*c*) reveal scientifically interesting data features that could lead to new hypotheses. Dashed arrows indicate steps such as training and validation that are taken before the ML or ML/XAI model is deployed to perform autonomously. Abbreviations: AI, artificial intelligence; ML, machine learning; XAI, explainable AI.

complex. ML/XAI can be used as a preprocessing tool for such data by, for instance, highlighting suspicious regions. In this context, XAI has been instrumental in generating heatmaps for histopathological images (56) that draw the human eye to specific tissue features. Highlighting tissue regions is already part of medical products such as Paige Prostate (see **https://info.paige.ai/prostate**); however, Paige Prostate uses so-called attention maps that are not true XAI. An ML model can also be used as a second opinion for the pathologist. In case of disagreement, further inspection is recommended (e.g., using XAI), a functionality that is already present in medical products such as the Ibex Second Read system (see **https://ibex-ai.com**). When using ML/XAI as a personal assistant, where the model is queried in real time, fast solutions based on neural networks combined with attention maps or propagation methods such as LRP are particularly suitable.

7.2. Machine Learning and Explainable Artificial Intelligence for Autonomous Decisions

While diagnostic decisions can be critical and ultimately require board-certified pathologists, time-consuming quantitative evaluations of tissue features [e.g., counting mitoses (166)] and a growing global shortage of pathologists pose a challenge. Fully autonomous diagnostic ML/XAI tools may be a solution, but obtaining regulatory approval is hard and systems need to prove robustness and highest accuracy. Best ML practices, as discussed above, such as choosing a proper model and regularization scheme and verifying the model performance not only via cross-validation but also on independent data sets, are all applicable here. In addition, XAI can be used to reveal flaws that are undetectable with cross-validation, such as reliance on artifactual features (89, 90). XAI may also be used to identify flaws after deployment. Several regulatory frameworks have been developed for such decision processes [e.g., General Data Protection Regulation (167), In Vitro Diagnostic Regulation], and discussions on how to standardize explanations are ongoing.

7.3. Machine Learning and Explainable Artificial Intelligence for New Scientific Insights

Biomedical research is producing increasingly large quantities of complex data. Prominent examples include The Cancer Genome Atlas (see **https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga**), the Gene Expression Omnibus database (168) for genomics data, and BigPicture.eu (see **https://bigpicture.eu**) for histological imaging data. These data sets have many potential scientific uses, ranging from descriptive statistics to research into disease mechanisms. Here, XAI provides a powerful extension to classical bioinformatics. ML models can learn complex, nonlinear, multivariate relations among different variables that are further evaluated by XAI and subsequently exploited in data-driven hypothesis generation (48, 49, 56).

8. DISCUSSION AND CONCLUSION

In addition to improving quantitative image evaluation, XAI enables physicians to verify results and obtain novel insights. Furthermore, XAI may support regulatory approval. In research, it may help solve a long-standing conundrum wherein data-driven (omics) approaches applied to large patient cohorts result only in high-dimensional static descriptions of disease states without the ability to functionally explore diseases, while insights gained from cell culture or animal models often cannot be transferred to the human situation. Here, XAI can help infer functional

properties and even higher-order associations in static high-dimensional data. In this manner, XAI can bridge the gap between data-driven hypothesis generation and conventional hypothesis–driven approaches. While our review has showcased successful pathology applications of ML in general and XAI specifically, there remain challenges that require further research between the disciplines.

A fundamental property of medical data is their uneven distribution. Relatively few diseases occur very frequently, while the majority of diseases are relatively rare. This is in contrast to most typical ML applications and makes the collection of training data covering the full spectrum of diseases very difficult. In addition, there is often a misfit between the number of cases and data dimensionality; for example, for omics data with tens of thousands of parameters, even the largest currently available data sets may not be sufficient for training robust models. Here, data-rich single-cell sequencing approaches may hold promise. Furthermore, very rare diseases yield an imbalance among classes that so far has been nearly impossible to model. While novel loss functions and data augmentation have been tested to improve the classification of rare classes, fundamental progress is still urgently needed. Long-tail distribution effects can be alleviated through anomaly detection. Cases that do not belong to any of the classes known to the classifier would typically be assigned to the most similar class, which could yield highly nonsensical predictions. More useful would be a rejection of classification for such rare cases, reflecting their anomalous characteristics. ML models should also be able to reflect their decision uncertainty.

An important aspect of the impressive recent success of foundation models (e.g., GPT, Lambda) in natural language processing is the size of the training data, which essentially consist of all text data in the Internet. In contrast, most cohort sizes in biomedicine are comparatively small. Therefore, extracting knowledge from the sheer quantity of data is currently infeasible in the medical domain. Thus, models have to rely on including prior medical knowledge, such as results from gene or protein interaction studies, that effectively increases data efficiency.

A further challenge is to better harness multimodal heterogeneous data. The issue is to integrate different types of imaging, omics, and clinical data, all with substantially different information structure and noise characteristics. Moreover, multimodal explanations need to be fused.

Recent research has shown that ML models can successfully predict clinical outcomes—a potentially valuable instrument for clinical trials and companion diagnostics. However, a word of caution seems necessary. ML models can overfit; therefore, a meticulous separation between trial data and model is mandatory in order to avoid overoptimistic predictions. While it is highly important to maintain fully independent validation, strategies for calibration from one clinical trial to another are needed.

Another challenge is interdisciplinarity, as ML experts and pathologists need to learn to interact and find a common language. Current curricula rarely cover the coursework necessary for students to appropriately reach a level at which they can seamlessly cooperate with members of the other discipline. Novel academic curricula, such as digital clinician scientist programs, are required at levels ranging from specialized minors to postgraduate education.

Finally, data privacy concerns need to be addressed with respect to both training and model deployment. Federated learning has recently shown promise in this regard (169).

The opportunities arising from AI are manifold, and AI has the potential to transform pathology. At the same time, current AI is far from being on par with human intelligence and should be seen as another, although mighty, ancillary technique that will improve diagnostics with respect to accuracy and predictive capabilities in the hands of expert pathologists. The full potential of AI, however, will unfold if it succeeds in integrating complex multimodal data for the development of novel diagnostics and the generation of novel biological insights.

DISCLOSURE STATEMENT

F.K. and K.R.M. are cofounders of the AI spin-off Aignostics GmbH, which develops AI algorithms for pathology. M.A. is employed as the CTO of Aignostics. L.R. is an employee of Aignostics. The other authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

AUTHOR CONTRIBUTIONS

F.K., G.M., and K.R.M. planned and supervised the writing of the review. F.K. and K.R.M. wrote the introduction, discussion, and conclusion. G.M., J.D., L.R., and M.A. wrote the sections on ML and XAI. G.M. and F.K. wrote the systematic view section. F.K., O.B., A.M., M.B., P.J., and P.K. wrote the computational pathology sections. A.M. wrote the part on multimodal data integration. The examples were written by P.K., J.D., L.R., and M.A. All authors participated in preparing tables and figures and reviewed the whole article.

ACKNOWLEDGMENTS

K.R.M. was supported in part by the German Ministry for Education and Research within BIFOLD and through grants 01IS14013A-E, 01GQ1115, 01GQ0850, 01IS18025A, 031L0207D, and 01IS18037A. K.R.M. was also partly supported by Institute of Information & Communications Technology Planning & Evaluation grants funded by the Korean government (2019-0-00079, AI Graduate School Program, Korea University, and 2022-0-00984). F.K. was supported in part by the German Ministry for Education and Research within BIFOLD, the German Consortium for Cancer Research, Berlin and Munich partner sites, under grants 031L0207B and 01IS21069D.

LITERATURE CITED

1. Krizhevsky A, Sutskever I, Hinton GE. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, ed. PL Bartlett, FCN Pereira, CJC Burges, L Bottou, KQ Weinberger, pp. 1106–14. Red Hook, NY: Curran
2. Ronneberger O, Fischer P, Brox T. 2015. U-net: convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015): 18th International Conference, Proceedings, Part III*, pp. 234–41. Berlin: Springer
3. Ren S, He K, Girshick RB, Sun J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39(6):1137–49
4. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33*, ed. H Larochelle, M Ranzato, R Hadsell, MF Balcan, H Lin, pp. 1877–901. Red Hook, NY: Curran
5. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596(7873):583–89
6. Unke OT, Chmiela S, Sauceda HE, Gastegger M, Poltavsky I, et al. 2021. Machine learning force fields. *Chem. Rev.* 121(16):10142–86
7. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, et al. 2017. A survey on deep learning in medical image analysis. *Med. Image Anal.* 42:60–88
8. Lengauer T, Sander O, Sierra S, Thielen A, Kaiser R. 2007. Bioinformatics prediction of HIV coreceptor usage. *Nat. Biotechnol.* 25(12):1407–10
9. Bottou L. 2010. Large-scale machine learning with stochastic gradient descent. In *19th International Conference on Computational Statistics (COMPSTAT)*, ed. Y Lechevallier, G Saporta, pp. 177–86. Heidelberg, Ger.: Physica
10. Kingma DP, Ba J. 2015. Adam: a method for stochastic optimization. arXiv:1412.6980 [cs.LG]

11. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, et al. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542(7639):115–18
12. Chen T, Kornblith S, Norouzi M, Hinton G. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1597–607. New York: ACM
13. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. 2022. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR 2022*), pp. 16000–9. Piscataway, NJ: IEEE
14. James G, Witten D, Hastie T, Tibshirani R. 2013. *An Introduction to Statistical Learning*. New York: Springer. 2nd ed.
15. Cortes C, Vapnik V. 1995. Support-vector networks. *Mach. Learn.* 20(3):273–97
16. Müller KR, Mika S, Rätsch G, Tsuda K, Schölkopf B. 2001. An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Netw.* 12(2):181–201
17. Zien A, Rätsch G, Mika S, Schölkopf B, Lengauer T, Müller KR. 2000. Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics* 16(9):799–807
18. Schölkopf B, Smola A, Müller KR. 1998. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.* 10(5):1299–319
19. LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* 521(7553):436–44
20. Schmidhuber J. 2015. Deep learning in neural networks: an overview. *Neural Netw.* 61:85–117
21. Goodfellow I, Bengio Y, Courville A. 2016. *Deep Learning*. Cambridge, MA: MIT Press
22. Chen RJ, Krishnan RG. 2021. *Self-supervised vision transformers learn visual concepts in histopathology*. Paper presented at Learning Meaningful Representations of Life Workshop, 35th Conference on Neural Information Processing Systems (NeurIPS 2021), online, Dec. 13–14
23. Krishnan R, Rajpurkar P, Topol EJ. 2022. Self-supervised learning in medicine and healthcare. *Nat. Biomed. Eng.* 6:1346–52
24. Goyal A, Bengio Y. 2022. Inductive biases for deep learning of higher-level cognition. *Proc. R. Soc. A* 478(2266):20210068
25. He K, Zhang X, Ren S, Sun J. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR 2016*), pp. 770–78. Piscataway, NJ: IEEE
26. Hochreiter S, Schmidhuber J. 1997. Long short-term memory. *Neural Comput.* 9(8):1735–80
27. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (*NeurIPS'17*), pp. 6000–10. Red Hook, NY: Curran
28. Rumelhart DE, Hinton GE, Williams RJ. 1986. Learning representations by back-propagating errors. *Nature* 323(6088):533–36
29. LeCun Y, Bottou L, Orr GB, Müller KR. 2012. Efficient BackProp. In *Neural Networks: Tricks of the Trade*, ed. G Montavon, GB Orr, KR Müller, pp. 9–48. Berlin: Springer
30. Shao X, Liao J, Lu X, Xue R, Ai N, Fan X. 2020. scCATCH: automatic annotation on cell types of clusters from single-cell RNA sequencing data. *iScience* 23(3):100882
31. Hastie T, Tibshirani R, Friedman J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer. 2nd ed.
32. Ilse M, Tomczak J, Welling M. 2018. Attention-based deep multiple instance learning. *Proc. Mach. Learn. Res.* 80:2127–36
33. Redmon J, Divvala SK, Girshick RB, Farhadi A. 2016. You only look once: unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR 2016*), pp. 779–88. Piscataway, NJ: IEEE
34. Bengio Y, Courville A, Vincent P. 2013. Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 35(8):1798–828
35. Schölkopf B, Locatello F, Bauer S, Ke NR, Kalchbrenner N, et al. 2021. Toward causal representation learning. *Proc. IEEE* 109(5):612–34
36. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, et al. 2021. A comprehensive survey on transfer learning. *Proc. IEEE* 109(1):43–76

37. Ciga O, Xu T, Martel AL. 2022. Self supervised contrastive learning for digital histopathology. *Mach. Learn. Appl.* 7:100198
38. Chen RJ, Chen C, Li Y, Chen TY, Trister AD, et al. 2022. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR 2022*), pp. 16123–34. Piscataway, NJ: IEEE
39. Ma S, Song X, Huang J. 2007. Supervised group Lasso with applications to microarray data analysis. *BMC Bioinform.* 8:60
40. Jurmeister P, Bockmayr M, Seegerer P, Bockmayr T, Treue D, et al. 2019. Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases. *Sci. Transl. Med.* 11(509):eaaw8513
41. Capper D, Jones DTW, Sill M, Hovestadt V, Schrimpf D, et al. 2018. DNA methylation–based classification of central nervous system tumours. *Nature* 555(7697):469–74
42. Nguyen L, Van Hoeck A, Cuppen E. 2022. Machine learning–based tissue of origin classification for cancer of unknown primary diagnostics using genome-wide mutation features. *Nat. Commun.* 13:4013
43. Garg M, Couturier DL, Nsengimana J, Fonseca NA, Wongchenko M, et al. 2021. Tumour gene expression signature in primary melanoma predicts long-term outcomes. *Nat. Commun.* 12:1137
44. Keyl J, Kasper S, Wiesweg M, Götze J, Schönrock M, et al. 2022. Multimodal survival prediction in advanced pancreatic cancer using machine learning. *ESMO Open* 7(5):100555
45. Zhang Y, Feng T, Wang S, Dong R, Yang J, et al. 2020. A novel XGBoost method to identify cancer tissue-of-origin based on copy number variations. *Front. Genet.* 11:585029
46. Li Q, Yang H, Wang P, Liu X, Lv K, Ye M. 2022. XGBoost-based and tumor-immune characterized gene signature for the prediction of metastatic status in breast cancer. *J. Transl. Med.* 20(1):177
47. Moncada-Torres A, van Maaren MC, Hendriks MP, Siesling S, Geleijnse G. 2021. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Sci. Rep.* 11:6968
48. Keyl P, Bockmayr M, Heim D, Dernbach G, Montavon G, et al. 2022. Patient-level proteomic network prediction by explainable artificial intelligence. *npj Precis. Oncol.* 6(1):35
49. Keyl P, Bischoff P, Dernbach G, Bockmayr M, Fritz R, et al. 2023. Single-cell gene regulatory network prediction by explainable AI. *Nucleic Acids Res.* 51(4):e20
50. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, et al. 2018. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* 36(10):983–87
51. Mostavi M, Chiu YC, Huang Y, Chen Y. 2020. Convolutional neural network models for cancer type prediction based on gene expression. *BMC Med. Genom.* 13(Suppl. 5):44
52. Boža V, Brejová B, Vinař T. 2017. DeepNano: deep recurrent neural networks for base calling in MinION nanopore reads. *PLOS ONE* 12(6):e0178751
53. Liu Q, Fang L, Yu G, Wang D, Xiao CL, Wang K. 2019. Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. *Nat. Commun.* 10:2449
54. Kim HH, Lim YS, Seo SI, Lee KJ, Kim JY, Shin WG. 2021. A deep recurrent neural network–based explainable prediction model for progression from atrophic gastritis to gastric cancer. *Appl. Sci.* 11(13):6194
55. Xu Y, Zhang Z, You L, Liu J, Fan Z, Zhou X. 2020. scIGANs: single-cell RNA-seq imputation using generative adversarial networks. *Nucleic Acids Res.* 48(15):e85
56. Binder A, Bockmayr M, Hägele M, Wienert S, Heim D, et al. 2021. Morphological and molecular breast cancer profiling through explainable machine learning. *Nat. Mach. Intell.* 3(4):355–66
57. Dietrich S, Oleś M, Lu J, Sellner L, Anders S, et al. 2018. Drug-perturbation-based stratification of blood cancer. *J. Clin. Investig.* 128(1):427–45
58. Schulte-Sasse R, Budach S, Hnisz D, Marsico A. 2021. Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. *Nat. Mach. Intell.* 3(6):513–26
59. Pfeifer B, Baniecki H, Saranti A, Biecek P, Holzinger A. 2022. Multi-omics disease module detection with an explainable Greedy Decision Forest. *Sci. Rep.* 12:16857

60. Chereda H, Bleckmann A, Menck K, Perera-Bel J, Stegmaier P, et al. 2021. Explaining decisions of graph convolutional neural networks: patient-specific molecular subnetworks responsible for metastasis prediction in breast cancer. *Genome Med.* 13(1):42
61. Bourgeais V, Zehraoui F, Hamdoune MB, Hanczar B. 2021. Deep GONet: self-explainable deep neural network based on gene ontology for phenotype prediction from gene expression data. *BMC Bioinform.* 22(Suppl. 10):455
62. Baehrens D, Schroeter T, Harmeling S, Kawanabe M, Hansen K, Müller KR. 2010. How to explain individual classification decisions. *J. Mach. Learn. Res.* 11:1803–31
63. Lauritsen SM, Kristensen M, Olsen MV, Larsen MS, Lauritsen KM, et al. 2020. Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nat. Commun.* 11:3852
64. Zeiler MD, Fergus R. 2014. Visualizing and understanding convolutional networks. In *Computer Vision: 13th European Conference* (*ECCV 2014*), ed. DJ Fleet, T Pajdla, B Schiele, T Tuytelaars, pp. 818–33, Berlin: Springer
65. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE* 10(7):e0130140
66. Arras L, Arjona-Medina JA, Widrich M, Montavon G, Gillhofer M, et al. 2019. Explaining and interpreting LSTMs. See Ref. 100, pp. 231–38
67. Preuer K, Klambauer G, Rippmann F, Hochreiter S, Unterthiner T. 2019. Interpretable deep learning in drug discovery. See Ref. 100, pp. 331–45
68. Schnake T, Eberle O, Lederer J, Nakajima S, Schütt KT, et al. 2022. Higher-order explanations of graph neural networks via relevant walks. *IEEE Trans. Pattern Anal. Mach. Intell.* 44(11):7581–96
69. Kim B, Wattenberg M, Gilmer J, Cai CJ, Wexler J, et al. 2018. Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). *Proc. Mach. Learn. Res.* 80:2673–82
70. Chormai P, Herrmann J, Müller KR, Montavon G. 2022. Disentangled explanations of neural network predictions by finding relevant subspaces. arXiv:2212.14855 [cs.LG]
71. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. 2015. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In *KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1721–30. New York: ACM
72. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, et al. 2020. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* 2(1):56–67
73. Wachter S, Mittelstadt B, Russell C. 2018. Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harvard J. Law Technol.* 31(2):841–87
74. Verma S, Dickerson JP, Hines K. 2020. Counterfactual explanations for machine learning: a review. arXiv:2010.10596 [cs.LG]
75. Zhou B, Khosla A, Lapedriza À, Oliva A, Torralba A. 2016. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR 2016*), pp. 2921–29. Piscataway, NJ: IEEE
76. Chen RJ, Lu MY, Williamson DFK, Chen TY, Lipkova J, et al. 2022. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell* 40(8):865–78
77. Strumbelj E, Kononenko I. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.* 41(3):647–65
78. Covert I, Lundberg SM, Lee S. 2021. Explaining by removing: a unified framework for model explanation. *J. Mach. Learn. Res.* 22:209
79. Lundberg SM, Lee S. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (*NeurIPS'17*), pp. 4765–74. Red Hook, NY: Curran
80. Fong RC, Vedaldi A. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *IEEE International Conference on Computer Vision* (*ICCV 2017*), pp. 3449–57. Piscataway, NJ: IEEE
81. Sundararajan M, Taly A, Yan Q. 2017. Axiomatic attribution for deep networks. *Proc. Mach. Learn. Res.* 70:3319–28
82. Ribeiro MT, Singh S, Guestrin C. 2016. "Why should I trust you?": explaining the predictions of any classifier. In *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–44. New York: ACM

83. Levy JJ, Titus AJ, Petersen CL, Chen Y, Salas LA, Christensen BC. 2020. MethylNet: an automated and modular deep learning approach for DNA methylation analysis. *BMC Bioinform.* 21(1):108
84. Liu B, Liu Y, Pan X, Li M, Yang S, Li SC. 2019. DNA methylation markers for pan-cancer prediction by deep learning. *Genes* 10(10):778
85. Modhukur V, Sharma S, Mondal M, Lawarde A, Kask K, et al. 2021. Machine learning approaches to classify primary and metastatic cancers using tissue of origin-based DNA methylation profiles. *Cancers* 13(15):3768
86. Zhou K, Arslanturk S, Craig DB, Heath E, Draghici S. 2021. Discovery of primary prostate cancer biomarkers using cross cancer learning. *Sci. Rep.* 11:10433
87. Lemsara A, Ouadfel S, Fröhlich H. 2020. PathME: pathway based multi-modal sparse autoencoders for clustering of patient-level multi-omics data. *BMC Bioinform.* 21(1):146
88. Böhle M, Fritz M, Schiele B. 2022. B-cos networks: Alignment is all we need for interpretability. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR 2022*), pp. 10329–38. Piscataway, NJ: IEEE
89. Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller KR. 2019. Unmasking Clever Hans predictors and assessing what machines really learn. *Nat. Commun.* 10:1096
90. Anders CJ, Weber L, Neumann D, Samek W, Müller KR, Lapuschkin S. 2021. Finding and removing Clever Hans: using explanation methods to debug and improve deep models. *Inf. Fusion* 77:261–95
91. Hägele M, Seegerer P, Lapuschkin S, Bockmayr M, Samek W, et al. 2020. Resolving challenges in deep learning–based analyses of histopathological images using explanation methods. *Sci. Rep.* 10:6423
92. Sanneman L, Shah J. 2023. Transparent Value Alignment. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (*HRI 2023*), ed. G Castellano, LD Riek, M Cakmak, I Leite, pp. 557–60. New York: ACM
93. Srivastava N, Hinton GE, Krizhevsky A, Sutskever I, Salakhutdinov R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15:1929–58
94. Jurmeister P, Glöß S, Roller R, Leitheiser M, Schmid S, et al. 2022. DNA methylation–based classification of sinonasal tumors. *Nat. Commun.* 13:7148
95. Leibig C, Allken V, Ayhan MS, Berens P, Wahl S. 2017. Leveraging uncertainty information from deep neural networks for disease detection. *Sci. Rep.* 7:17816
96. Luecken MD, Theis FJ. 2019. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* 15(6):e8746
97. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, et al. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* 11(10):733–39
98. Calude CS, Longo G. 2016. The deluge of spurious correlations in big data. *Found. Sci.* 22(3):595–612
99. Gunning D, Aha DW. 2019. DARPA's explainable artificial intelligence (XAI) program. *AI Mag.* 40(2):44–58
100. Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR, eds. 2019. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Berlin: Springer
101. Holzinger A, Goebel R, Fong R, Moon T, Müller KR, Samek W, eds. 2022. *XXAI—Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020. Revised and Extended Papers*. Berlin: Springer
102. Samek W, Montavon G, Lapuschkin S, Anders CJ, Müller KR. 2021. Explaining deep neural networks and beyond: a review of methods and applications. *Proc. IEEE* 109(3):247–78
103. Zhang Y, Weng Y, Lund J. 2022. Applications of explainable artificial intelligence in diagnosis and surgery. *Diagnostics* 12(2):237
104. Montavon G, Binder A, Lapuschkin S, Samek W, Müller KR. 2019. Layer-wise relevance propagation: an overview. See Ref. 100, pp. 193–209
105. Lipton ZC. 2018. The mythos of model interpretability. *Commun. ACM* 61(10):36–43
106. Rudin C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1(5):206–15
107. Letham B, Rudin C, McCormick TH, Madigan D. 2015. Interpretable classifiers using rules and Bayesian analysis: building a better stroke prediction model. *Ann. Appl. Stat.* 9(3):1350–71

108. Ali A, Schnake T, Eberle O, Montavon G, Müller KR, Wolf L. 2022. XAI for transformers: better explanations through conservative propagation. *Proc. Mach. Learn. Res.* 162:435–51
109. Hosseini MS, Bejnordi BE, Trinh VQH, Hasan D, Li X, et al. 2023. Computational pathology: a survey review and the way forward. arXiv:2304.05482 [eess.IV]
110. Baxi V, Edwards R, Montalto M, Saha S. 2022. Digital pathology and artificial intelligence in translational medicine and clinical practice. *Mod. Pathol.* 35(1):23–32
111. Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, et al. 2017. QuPath: open source software for digital pathology image analysis. *Sci. Rep.* 7:16878
112. Pocock J, Graham S, Vu QD, Jahanifar M, Deshpande S, et al. 2022. TIAToolbox as an end-to-end library for advanced tissue image analytics. *Commun. Med.* 2:120
113. Cifci D, Veldhuizen GP, Foersch S, Kather JN. 2023. AI in computational pathology of cancer: improving diagnostic workflows and clinical outcomes? *Annu. Rev. Cancer Biol.* 7:57–71
114. Schmidt U, Weigert M, Broaddus C, Myers G. 2018. Cell detection with star-convex polygons. arXiv:1806.03535 [cs]
115. Kather JN, Pearson AT, Halama N, Jäger D, Krause J, et al. 2019. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* 25(7):1054–56
116. Warnat-Herresthal S, Schultze H, Shastry KL, Manamohan S, Mukherjee S, et al. 2021. Swarm learning for decentralized and confidential clinical machine learning. *Nature* 594(7862):265–70
117. Saldanha OL, Quirke P, West NP, James JA, Loughrey MB, et al. 2022. Swarm learning for decentralized artificial intelligence in cancer histopathology. *Nat. Med.* 28(6):1232–39
118. Leiter C, Zhang R, Chen Y, Belouadi J, Larionov D, et al. 2023. ChatGPT: a meta-analysis after 2.5 months. arXiv:2302.13795 [cs]
119. Ruff L, Kauffmann JR, Vandermeulen RA, Montavon G, Samek W, et al. 2021. A unifying review of deep and shallow anomaly detection. *Proc. IEEE* 109(5):756–95
120. Zehnder P, Feng J, Fuji RN, Sullivan R, Hu F. 2022. Multiscale generative model using regularized skip-connections and perceptual loss for anomaly detection in toxicologic histopathology. *J. Pathol. Inform.* 13:100102
121. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, et al. 2018. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* 24(10):1559–67
122. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Krauss Silva VW, et al. 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* 25(8):1301–9
123. Wang S, Wang T, Yang L, Yang DM, Fujimoto J, et al. 2019. ConvPath: a software tool for lung adenocarcinoma digital pathological image analysis aided by a convolutional neural network. *eBioMedicine* 50:103–10
124. Kiani A, Uyumazturk B, Rajpurkar P, Wang A, Gao R, et al. 2020. Impact of a deep learning assistant on the histopathologic classification of liver cancer. *npj Digit. Med.* 3:23
125. Bao L, Zhou M, Cui Y. 2005. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Res.* 33:W480–82
126. Sammut SJ, Crispin-Ortuzar M, Chin SF, Provenzano E, Bardwell HA, et al. 2022. Multi-omic machine learning predictor of breast cancer therapy response. *Nature* 601(7894):623–29
127. Alharbi WS, Rashid M. 2022. A review of deep learning applications in human genomics using next-generation sequencing data. *Hum. Genom.* 16(1):26
128. Moore LD, Le T, Fan G. 2013. DNA methylation and its basic function. *Neuropsychopharmacology* 38(1):23–38
129. Lokk K, Modhukur V, Rajashekar B, Märtens K, Mägi R, et al. 2014. DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biol.* 15(4):r54
130. Kim M, Costello J. 2017. DNA methylation: an epigenetic mark of cellular memory. *Exp. Mol. Med.* 49(4):e322
131. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, et al. 2016. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol.* 17(1):208

132. Hovestadt V, Remke M, Kool M, Pietsch T, Northcott PA, et al. 2013. Robust molecular subgrouping and copy-number profiling of medulloblastoma from small amounts of archival tumour material using high-density DNA methylation arrays. *Acta Neuropathol.* 125(6):913–16
133. Pajtler KW, Witt H, Sill M, Jones DT, Hovestadt V, et al. 2015. Molecular classification of ependymal tumors across all CNS compartments, histopathological grades, and age groups. *Cancer Cell* 27(5):728–43
134. Sturm D, Orr BA, Toprak UH, Hovestadt V, Jones DTW, et al. 2016. New brain tumor entities emerge from molecular classification of CNS-PNETs. *Cell* 164(5):1060–72
135. Bockmayr M, Harnisch K, Pohl LC, Schweizer L, Mohme T, et al. 2022. Comprehensive profiling of myxopapillary ependymomas identifies a distinct molecular subtype with relapsing disease. *Neuro-Oncology* 24(10):1689–99
136. Koelsche C, Schrimpf D, Stichel D, Sill M, Sahm F, et al. 2021. Sarcoma classification by DNA methylation profiling. *Nat. Commun.* 12:498
137. Louis DN, Wesseling P, Aldape K, Brat DJ, Capper D, et al. 2020. cIMPACT-NOW update 6: new entity and diagnostic principle recommendations of the cIMPACT-Utrecht meeting on future CNS tumor classification and grading. *Brain Pathol.* 30(4):844–56
138. Gündert M, Edelmann D, Benner A, Jansen L, Jia M, et al. 2019. Genome-wide DNA methylation analysis reveals a prognostic classifier for non-metastatic colorectal cancer (ProMCol classifier). *Gut* 68(1):101–10
139. Nassiri F, Mamatjan Y, Suppiah S, Badhiwala JH, Mansouri S, et al. 2019. DNA methylation profiling to predict recurrence risk in meningioma: development and validation of a nomogram to optimize clinical management. *Neuro-Oncology* 21(7):901–10
140. Jeschke J, Bizet M, Desmedt C, Calonne E, Dedeurwaerder S, et al. 2017. DNA methylation–based immune response signature improves patient diagnosis in multiple cancers. *J. Clin. Investig.* 127(8):3090–102
141. Safaei S, Mohme M, Niesen J, Schüller U, Bockmayr M. 2021. DIMEimmune: robust estimation of infiltrating lymphocytes in CNS tumors from DNA methylation profiles. *Oncoimmunology* 10(1):1932365
142. Leitheiser M, Capper D, Seegerer P, Lehmann A, Schüller U, et al. 2022. Machine learning models predict the primary sites of head and neck squamous cell carcinoma metastases based on DNA methylation. *J. Pathol.* 256(4):378–87
143. Bradley R, Braybrooke J, Gray R, Hills R, Liu Z, et al. 2021. 864 women in seven randomised trials. *Lancet Oncol.* 22(8):1139–50
144. Chapman PB, Hauschild A, Robert C, Haanen JB, Ascierto P, et al. 2011. Improved survival with vemurafenib in melanoma with *BRAF V600E* mutation. *N. Engl. J. Med.* 364(26):2507–16
145. Cercek A, Lumish M, Sinopoli J, Weiss J, Shia J, et al. 2022. PD-1 blockade in mismatch repair–deficient, locally advanced rectal cancer. *N. Engl. J. Med.* 386(25):2363–76
146. Keyl J, Hosch R, Berger A, Ester O, Greiner T, et al. 2023. Deep learning–based assessment of body composition and liver tumour burden for survival modelling in advanced colorectal cancer. *J. Cachexia Sarcopenia Muscle* 14(1):545–52
147. Delgado FM, Gómez-Vela F. 2019. Computational methods for gene regulatory networks reconstruction and analysis: a review. *Artif. Intell. Med.* 95:133–45
148. Fiers MWEJ, Minnoye L, Aibar S, Bravo González-Blas C, Kalender Atak Z, Aerts S. 2018. Mapping gene regulatory networks from single-cell omics data. *Brief. Funct. Genom.* 17(4):246–54
149. Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali TM. 2020. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* 17(2):147–54
150. Rao VS, Srinivas K, Sujini GN, Kumar GN. 2014. Protein-protein interaction detection: methods and analysis. *Int. J. Proteom.* 2014:147648
151. Raatz M, Shah S, Chitadze G, Brüggemann M, Traulsen A. 2021. The impact of phenotypic heterogeneity of tumour cells on treatment and relapse dynamics. *PLOS Comput. Biol.* 17(2):e1008702
152. Marusyk A, Janiszewska M, Polyak K. 2020. Intratumor heterogeneity: the Rosetta Stone of therapy resistance. *Cancer Cell* 37(4):471–84
153. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. 2010. Inferring regulatory networks from expression data using tree-based methods. *PLOS ONE* 5(9):e12776

154. Moerman T, Aibar Santos S, Bravo González-Blas C, Simm J, Moreau Y, et al. 2019. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics* 35(12):2159–61
155. Medina-Martínez JS, Arango-Ossa JE, Levine MF, Zhou Y, Gundem G, et al. 2020. Isabl Platform, a digital biobank for processing multimodal patient data. *BMC Bioinform.* 21(1):549
156. Chen RJ, Lu MY, Weng WH, Chen TY, Williamson DFK, et al. 2021. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *2021 IEEE/CVF International Conference on Computer Vision* (*ICCV 2021*), pp. 3995–4005. Piscataway, NJ: IEEE
157. Chen RJ, Lu MY, Wang J, Williamson DFK, Rodig SJ, et al. 2022. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans. Med. Imaging* 41(4):757–70
158. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* 102(43):15545–50
159. Tan K, Huang W, Liu X, Hu J, Dong S. 2022. A multi-modal fusion framework based on multi-task correlation learning for cancer prognosis prediction. *Artif. Intell. Med.* 126:102260
160. Boehm KM, Khosravi P, Vanguri R, Gao J, Shah SP. 2022. Harnessing multimodal data integration to advance precision oncology. *Nat. Rev. Cancer* 22(2):114–26
161. Lipkova J, Chen RJ, Chen B, Lu MY, Barbieri M, et al. 2022. Artificial intelligence for multimodal data integration in oncology. *Cancer Cell* 40(10):1095–110
162. Kather JN, Halama N, Marx A. 2018. 100,000 histological images of human colorectal cancer and healthy tissue. *Zenodo*, April 7. **https://doi.org/10.5281/zenodo.1214456**
163. Anders CJ, Neumann D, Samek W, Müller KR, Lapuschkin S. 2021. Software for dataset-wide XAI: from local explanations to global insights with Zennit, CoRelAy, and ViRelAy. arXiv:2106.13200v2 [cs.LG]
164. Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, et al. 2009. A method for normalizing histology slides for quantitative analysis. In *Proceedings of the 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 1107–10. Piscataway, NJ: IEEE
165. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, et al. 2020. The reactome pathway knowledgebase. *Nucleic Acids Res.* 48(1):D498–503
166. Veta M, van Diest PJ, Willems SM, Wang H, Madabhushi A, et al. 2015. Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Med. Image Anal.* 20:237–48
167. Goodman B, Flaxman SR. 2017. European Union regulations on algorithmic decision-making and a "right to explanation." *AI Mag.* 38(3):50–57
168. Clough E, Barrett T. 2016. The Gene Expression Omnibus database. *Methods Mol. Biol.* 1418:93–110
169. Kaissis GA, Makowski MR, Rückert D, Braren RF. 2020. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* 2(6):305–11