




Explainability and causability in digital pathology

Markus Plass¹ , Michaela Kargl¹, Tim-Rasmus Kiehl², Peter Regitnig¹, Christian Geißler³, Theodore Evans³, Norman Zerbe², Rita Carvalho², Andreas Holzinger^{1,4}  and Heimo Müller^{1*} 

¹Diagnostic and Research Institute of Pathology, Medical University of Graz, Graz, Austria

²Charité-Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Pathology, Berlin, Germany

³DAI-Labor, Agent Oriented Technologies (AOT), Technische Universität Berlin, Berlin, Germany

⁴Human-Centered AI Lab, University of Natural Resources and Life Sciences Vienna, Vienna, Austria

*Correspondence to: Heimo Müller, Diagnostic and Research Institute of Pathology, Medical University of Graz, Neue Stiftingtalstraße 6, A-8010 Graz, Austria. E-mail: heimo.mueller@medunigraz.at

Abstract

The current move towards digital pathology enables pathologists to use artificial intelligence (AI)-based computer programmes for the advanced analysis of whole slide images. However, currently, the best-performing AI algorithms for image analysis are deemed black boxes since it remains – even to their developers – often unclear why the algorithm delivered a particular result. Especially in medicine, a better understanding of algorithmic decisions is essential to avoid mistakes and adverse effects on patients. This review article aims to provide medical experts with insights on the issue of explainability in digital pathology. A short introduction to the relevant underlying core concepts of machine learning shall nurture the reader's understanding of why explainability is a specific issue in this field. Addressing this issue of explainability, the rapidly evolving research field of explainable AI (XAI) has developed many techniques and methods to make black-box machine-learning systems more transparent. These XAI methods are a first step towards making black-box AI systems understandable by humans. However, we argue that an explanation interface must complement these explainable models to make their results useful to human stakeholders and achieve a high level of causability, i.e. a high level of causal understanding by the user. This is especially relevant in the medical field since explainability and causability play a crucial role also for compliance with regulatory requirements. We conclude by promoting the need for novel user interfaces for AI applications in pathology, which enable contextual understanding and allow the medical expert to ask interactive ‘what-if’-questions. In pathology, such user interfaces will not only be important to achieve a high level of causability. They will also be crucial for keeping the human-in-the-loop and bringing medical experts' experience and conceptual knowledge to AI processes.

Keywords: digital pathology; artificial intelligence; explainability; causability

Received 3 November 2022; Revised 17 February 2023; Accepted 16 March 2023

No conflicts of interest were declared.

Introduction

During the last decade, technological advancements in whole slide images (WSIs) and approval for clinical use by regulatory agencies in many countries have paved the way for implementing digital workflows in diagnostic pathology. This shift to digitisation enables pathologists to view, examine and annotate histopathological slides seamlessly in various magnifications on a computer screen and facilitates telepathology and obtaining a second opinion from (remote) colleagues. However, digital pathology is not just a modern

version of the conventional microscope but the foundation for computational pathology. It enables the usage of artificial intelligence (AI) to aggregate information from multiple sources of patient information including WSIs. Thus, this big-data approach to pathology opens up entirely new capabilities and will change the pathology practice [1,2].

Over the last few years, many computational approaches for the advanced analysis of WSIs have been developed. These applications aim to support pathologists with tedious routine tasks, improve the accuracy of diagnosis, and aid in exploring new

diagnostic and prognostic criteria in many pathology subspecialties, including breast pathology [3], lung pathology [4], prostate [5,6], musculoskeletal [7], and dermatopathology [8]. Many of these computational approaches deliver results comparable to human experts or may even outperform human pathologists in specific tasks [9]. Furthermore, computational WSI analysis enables exploring aspects beyond human capabilities using ‘sub-visual’ information. An example is the prediction of molecular genetic alterations from histopathologic morphology [10]. Most of these applications utilise AI, specifically machine learning (ML) techniques to achieve such promising results. However, these AI applications have become increasingly opaque, meaning that it is often impossible to retrace and understand how a result was generated. Therefore, there is a growing call for explainability, especially in high-stakes domains such as medicine.

This review aims to shed light on the issue of explainability in digital pathology:

1. To provide insights into the context of explainability in digital pathology, we briefly dive into the basic principles of ML.
2. We explain the notion of explainability and give an overview of the state-of-the-art in relevant explainability techniques.
3. We introduce the concept of causability and describe why explainability and causability are particularly important in pathology.
4. We discuss current research strands and open questions in this field.

ML: basic principles

The concept of explainability in digital pathology is closely linked to ML applications in this domain. To lay the ground for the following sections, we briefly introduce basic principles and limitations of ML without going into technical details.

ML, a subfield of AI, comprises computer programmes that can automatically learn and improve from ‘experience’ without following explicit instructions. This distinguishes ML from traditional software engineering, where the developers explicitly define and encode the rules to determine how the software handles the input data for a specific output. In ML, the algorithm approximates input–output relations from data using numerical optimisation and statistics. Most ML approaches define a solution space (also called ‘model’), a loss function and a learning algorithm that searches for solutions that minimise the loss function. In unsupervised learning, a

sub-discipline of ML, the algorithm attempts to find patterns in the input data. Unsupervised ML is used for clustering and automatically recovering structure in data or feature extraction for other ML types. In supervised learning, the algorithm receives a so-called training dataset, which includes samples of input data and corresponding output data (often called labels). The ML algorithm autonomously derives a model that best describes input–output relations for these training data. The goal is to train an ML model which does not only perform well on the training data but can also make highly accurate predictions for previously unseen data – an ability called generalisation. A test dataset, not previously seen by the model, is used to test the generalisation-ability of the ML model.

Current ML applications for classification tasks in pathology, where the algorithm predicts a discrete class (e.g. tumour/non-tumour) and for regression tasks, where the algorithm predicts a continuous value (e.g. likelihood of tumour metastasis), are mostly based on training with labelled data [11]. Unsupervised ML techniques are mainly used for assisting tasks along the ML pipeline. For example, clustering can be used to propagate labels from a subset of annotated data to remaining unlabelled data [12].

Generally, the ML model after training on images, is a complex high-dimensional non-linear mathematical function that can become quite complicated, depending on the complexity of the task and the ML algorithms used. For example, deep learning (DL) algorithms, a family of ML algorithms using ‘deep’ (i.e. multi-layered) neural networks, yield ML models comprising millions of parameters and are so complex that humans cannot understand relationships amongst variables to form the model’s prediction. Such ML models are called ‘black-box’ models [13].

An ML model, which reaches high prediction accuracy for the test dataset, can still show decreased prediction accuracy in a real-world setting. Two common reasons for prediction failures of ML models are out-of-distribution (OOD) data and spurious correlations.

Out-of-distribution data

The prediction performance of an ML model is approximated under the assumption of being applied to identically distributed data as encountered during the training. For OOD data, the performance may be substantially different. An illustrative example for such OOD input data would be to present an immunohistochemistry (IHC) image to an ML model that was trained solely with haematoxylin and eosin stained slides. However, especially in digital pathology, issues

with OOD data are often more subtle. For example, staining variations between pathology laboratories can result in decreased performance of an ML application when trained on WSIs of a single laboratory and applied to WSIs of another laboratory [14].

Spurious correlations

The predictions of an ML model are based on the input–output relations it has learned from the training data. However, the input–output relations learned by the ML model are not causal relationships but only correlations. An ML model may also base its predictions on spurious correlations, i.e. ones that occur purely by chance. Confounding variables, which cause spurious correlations, may be artefacts visible to humans. For example, predictions of an ML model for the recognition of melanoma in dermoscopic images have shown associations with surgical skin markings [15]. However, confounding variables in digital pathology may also be hidden variables such as patient age, preparation date or origin of histology slide, or WSI scanner [16].

In practice, such decreased prediction accuracy and prediction failures of ML models are a problem if they cannot be identified and recognised by the human who relies on the machine's decision. Users must be aware of the limitations of AI applications and consider their potential flaws [17]. Therefore, if an AI system is supposed to contribute to a medical decision, where wrong decisions can have significant adverse effects on patients, medical experts must have the means to retrace and understand machine decision processes [13]. The lack of explainability of black-box models is often named as a significant obstacle to introducing these applications in high-stakes areas such as medicine [18].

Explainability and approaches of explainable AI

'Explainability' refers to a characteristic of an AI system which enables humans to understand why the AI system delivered the presented predictions [17]. Technically speaking, explainability highlights those elements of input data and ML model which contributed to the AI system's output [19].

It can be differentiated into global explanations, which aim to provide insights into the inner workings of the ML model and make the model transparent as a whole, and local explanations, which aim to explain a specific prediction of the ML model [20,21]. Furthermore, two types of AI systems can be distinguished

concerning the time when explainability is obtained: ante hoc explainable systems and post hoc explainable systems [22]. Ante hoc explainable systems are inherently interpretable by design since they use ML algorithms that generate models which are directly human-interpretable. Classical examples include decision trees, linear regression and fuzzy inference systems [23], but there are also approaches to create interpretable DL algorithms [24,25]. Post hoc explanation methods generate explanations for so-called black-box ML models, which are too complex to be directly interpretable. Post hoc explanation methods usually do not aim to explain how an ML model functions, but they aim to explain why it made a specific prediction [21].

During the last decade, the research topic of explainable AI (XAI), a subfield of ML, has developed many methods and mechanisms to provide human-interpretable explanations for complex ML models. Several review papers (e.g. [25–27]) provide detailed overviews of current XAI techniques.

It can be distinguished between XAI methods aiming to explain the inner workings of the ML model, XAI methods seeking to explain a specific prediction of the ML model and XAI methods aiming to estimate the uncertainty of the ML model's prediction [25]. So-called model-ground explanations, i.e. XAI methods that aim to explain the inner workings of the ML model, are mainly targeted at the developers of the ML system who want to gain insights on how to improve the system [28]. In contrast, explanations of the ML model's predictions and uncertainty are highly relevant to the users of the ML application.

In pathology, where typical tasks for ML applications are related to WSI analysis, explanations are best conveyed to the user by visualization on the input image or synthetic visualisations [29]. Popular presentation modalities of XAI methods in the imaging domain include saliency maps, concept attribution, prototypes, counterfactuals, and trust scores (see Figure 1) [29,30].

Saliency maps show the estimated relevance of each (group of) pixel(s) of the input image to the ML model's prediction in the form of a visual heatmap overlay on the input image. Various techniques are available to estimate the relevance of input pixels to the ML model's prediction. Often, these techniques result in divergent explanations for the same input image [30,31]. For non-experts in the XAI field, who often do not have sufficient knowledge and information to understand which technique has been applied to create a specific saliency map and what the limitations of that specific approximation technique are, saliency maps might provide a misleading explanation.

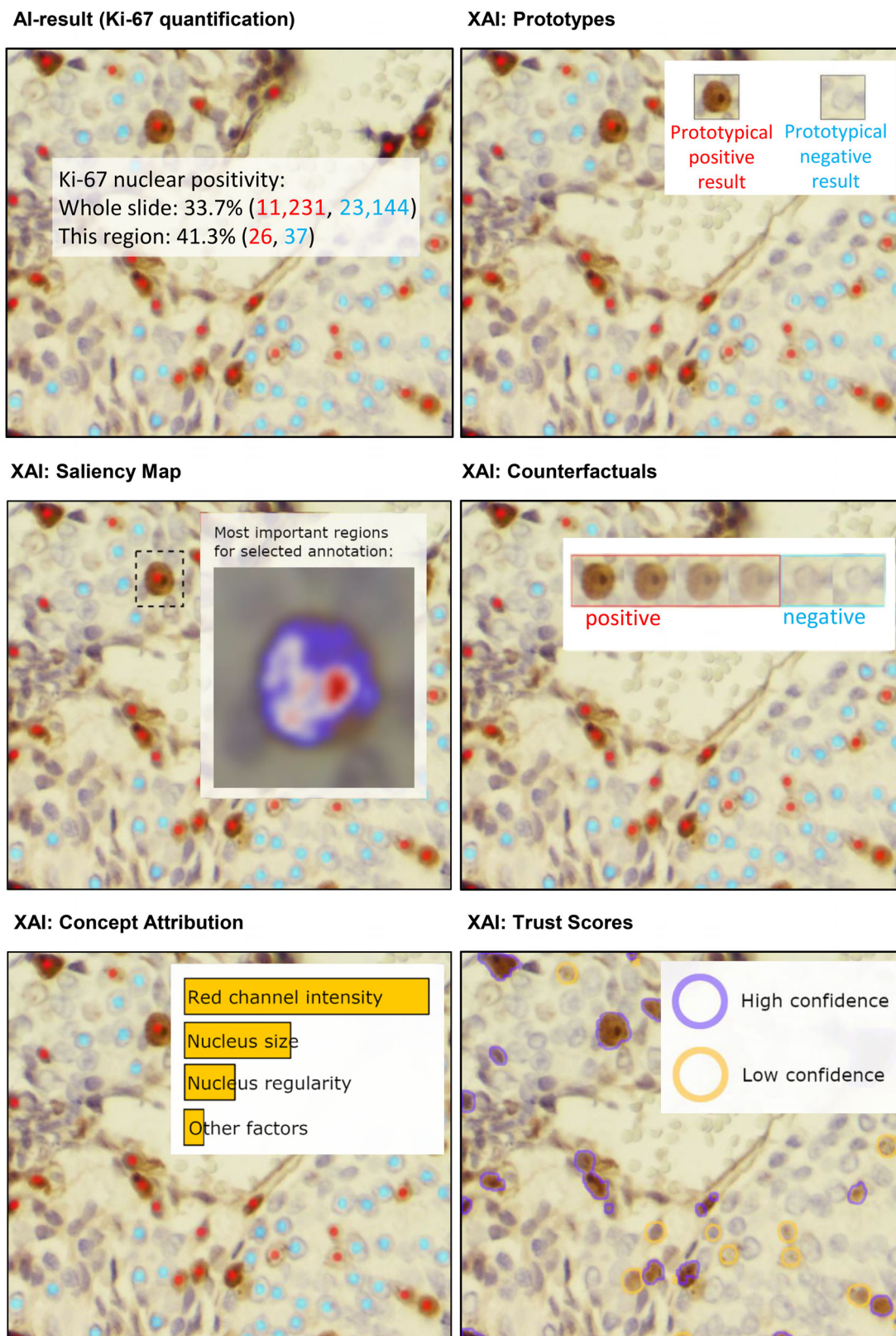


Figure 1. Legend on next page.

Thus, although they are one of the most widespread XAI methods, saliency maps should be applied and interpreted with caution [21].

Concept attribution techniques depict the estimated relevance of human-defined concepts (e.g. colour, area, architecture of nuclear features, mitosis) to the ML model's prediction in the form of relevance scores [32].

Prototypes are a so-called 'explanation by example' technique. The prototypes method aims to convey the common features of a specific prediction class by showing representative input examples for the respective prediction [30]. Prototypes can either be examples picked from the model's training dataset or synthetic examples [33,34].

Counterfactuals are also an 'explanation by example' technique. The counterfactuals method displays examples of input data, which are similar to the specific input but result in a different output of the ML model. Counterfactuals aim to reveal necessary minimal changes in the input so that one would obtain a contrastive output of the ML model [30,35].

Trust scores provide estimates for the ML model's uncertainty and help to better understand the limitations of the ML model resulting from the model's training data and training scheme (i.e. reducible [epistemic] uncertainty) as well as the limitations resulting from the fact that any ML model is only a simplified emulation of the real world (i.e. irreducible [aleatoric] uncertainty) [25]. Uncertainty measures can be visualised as overlays on the input image to highlight areas where the ML model's uncertainty is high [25]. Alternatively, they can be used to calculate confidence intervals of the model's output [36].

Since post hoc explanations created by XAI techniques are simplified approximations of complex ML models, it is vital to be aware of their limitations. Users must be informed where, i.e. for which domain and task, the underlying approximations are valid and reliable and where such an explanation might become inaccurate and misleading [21]. In addition to the risk that the recipient of a post hoc explanation is not aware of its limitations, there is also a risk that (ambiguous) post hoc explanations unintentionally support positive confirmation bias and, thus, false interpretations by the recipients [29]. Furthermore,

post hoc explanations can be intentionally misused to foster inadequate trust, for example, by falsely attributing an AI decision to an irrelevant but acceptable feature [21,37].

From explainability to causability

AI applications in high-stakes domains such as pathology must perform well and reliably. In addition, they should also be interpretable and explainable for a human expert. Indeed, various stakeholder groups in pathology have different requirements regarding explainability [38]: in clinical use, explainability should enable the pathologist to make an informed decision on whether or not to rely on the system's recommendations, specifically in cases where the AI system and the medical expert disagree [17]. Explainability should support the pathology laboratory's quality assurance in understanding the limitations of the AI application and the ML model's ability to generalise to the specific framework conditions of that laboratory [25]. Finally, medical researchers also require explainability, as they seek to uncover new insights from AI and thus need to understand the causality of patterns learned by an ML model [13]. However, current best-performing AI approaches typically rely on statistical correlations and cannot build causal models to support human understanding. There is a need to disentangle correlation from causation to avoid providing misleading explanations [39].

As described in the previous section, XAI methods are a first step towards making black-box models better understandable by humans. However, an explainable model (XAI) is not sufficient. To make the results gained by that model useful to the human stakeholder, an explanation interface must complement the explainable model (see Figure 2) [13,40]. Ideally, as shown recently [29], the explanation interface for an AI system in digital pathology should be interactive, enabling the user to retrieve explanations on demand and to ask 'why' and 'what-if' questions to refine the understanding of explanatory factors underlying the ML model's prediction. Holzinger *et al* [13] have coined the term

Figure 1. Popular presentation modalities of XAI methods are prototypes (show representative input examples for the predicted classes), counterfactuals (show how minimal changes in the input would lead to contrastive output), trust scores (highlight areas where the ML model's uncertainty is high), saliency map (visualises estimated relevance of input pixels to the ML model's output) and concept attribution (depict the estimated relevance of human-defined concepts to the ML model's prediction). Here these presentation modalities of XAI methods are exemplified through AI results for Ki-67 quantification (graphic adapted from [29]).

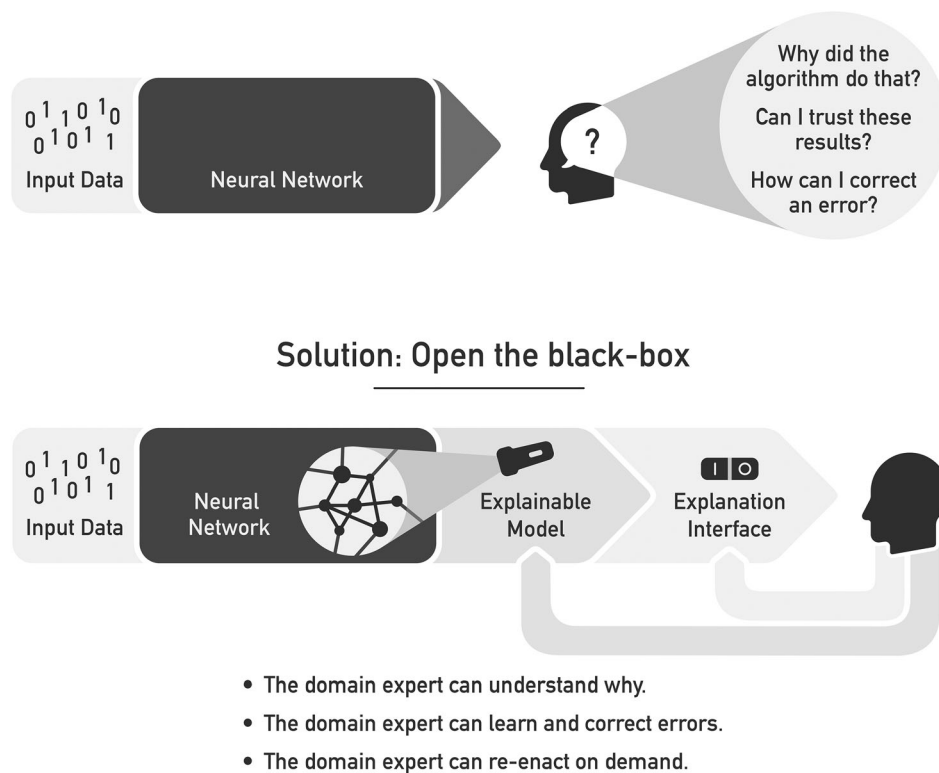


Figure 2. XAI techniques help to make black-box ML models more transparent and explainable. However, these explainable models must be complemented by an explanation interface to deliver results that are useful to the users and achieve high causability.

‘causability’ as a measure of the usability of such explanation interfaces:

Causability is the extent to which an explanation of a statement to a human expert achieves a specified level of causal understanding with effectiveness, efficiency and satisfaction in a specified context of use. [13]

Thus, similar to the well-known term usability, which is widely utilised in human–computer interaction to measure the quality of use, causability measures the quality of explanations in human–AI interaction. As such, causability is critical for successfully designing, developing and evaluating human–AI interfaces [41]. In digital pathology, there is a need for effective and efficient human–AI interfaces to empower human experts to take responsibility for their AI-supported decision-making by helping them better understand the underlying factors for an ML model’s prediction. In addition, human–AI interfaces should also enable humans to complement AI with conceptual knowledge, experience, and contextual understanding [42]. For example, the graphical user interface of an AI solution for automatic Ki-67 quantification in digital pathology (as shown in Figure 1) should not only

visualise in an overlay on the WSI all cells the AI has classified as Ki-67-positive tumour cells, but should also enable the pathologist to easily provide corrective feedback, for example by clicking on cells misclassified by the AI. Moreover, human–AI interfaces play a crucial role in keeping the human always in control, which is essential in the medical field for social, ethical and legal reasons [41].

Explainability and causability in the regulatory context

The regulatory landscape for market approval of medical devices is different across continents and countries and it is beyond the scope of this article to give a detailed and comprehensive overview. Instead, in the following paragraphs, the importance of explainability and causability in the regulatory context is illustrated based on the example of the European *in vitro* diagnostic medical devices regulation (IVDR).

In Europe, AI applications for WSI analysis in diagnostics are regarded as IVD as defined in article 2 (2) of the European IVDR [43]. To apply for market

approval of an IVD in the European Union, the IVDR requests manufacturers to provide evidence of scientific validity, analytical performance and clinical performance of the IVD.

Role of explainability and causability in demonstrating scientific validity

To provide evidence for the scientific validity of an AI application, it must be demonstrated that there is a scientifically proven link between the output of the AI application and the targeted physiological state or clinical condition as defined in the intended purpose of the AI application. Demonstrating such an association of the output of the AI application with the targeted clinical condition or physiological state is impossible with a black-box ML approach. Still, it requires explainability and causability to answer the question: *Why does the AI application work in general and generate reliable results for the intended purpose?* [44]. However, for demonstration of scientific validity, it must be distinguished between (1) an AI algorithm that supports or replaces an existing diagnostic method that scientific studies have already validated and (2) an AI algorithm that itself contributes significantly to the scientific and diagnostic approach [44].

An illustrative example for the first case would be an AI algorithm that assists pathologists with quantifying the Ki-67 labelling index (i.e. the percentage of Ki-67 IHC-stained tumour nuclei), a well-established indicator for cellular proliferation used for diagnosis and prognosis assessment in various cancers [45]. Manufacturers of the AI algorithm may refer to the existing scientific studies providing evidence for the scientific validity of Ki-67 as a cellular proliferation index. However, in addition, the algorithm must demonstrate that its results (i.e. the Ki-67-values determined by the algorithm) are indeed based on the percentage of Ki-67-stained tumour cell nuclei of the input WSI data. To show this, the algorithm should: (1) mark the identified tumour cell nuclei as Ki-67-positive or -negative on the WSI and (2) compute the Ki-67 result as the percentage of positive tumour cells in the region of interest. This visualisation of the algorithm's outcome makes the algorithm's prediction process retractable and verifiable by a medical expert.

An illustrative example for the second case, an AI algorithm that significantly contributes to the scientific and diagnostic approach, is a survival prediction algorithm for colorectal cancer which was not trained to identify already known features but used a weakly supervised learning approach to find novel features

[46]. By analysing the algorithm's results, medical experts found that small groups of tumour cells in the vicinity of fat cells, called 'tumour-associated fat' (TAF), constituted a feature of high-risk clusters found by the algorithm. However, scientific studies have not yet validated the diagnostic approach based on TAF. Thus, to provide evidence for the scientific validity of the AI algorithm, it would be necessary to conduct independent validation studies to demonstrate the correlation of the presence of TAF with the survival of colorectal cancer patients in independent cohorts.

Role of explainability and causability in analytical performance evaluation

Evaluation of the analytical performance should demonstrate the AI application's ability to reliably and accurately generate the intended analytical output from the input data over the whole range of the intended use [47]. Thus, it is especially important that test datasets used for the evaluation of the analytical performance of an AI application cover features of the entire intended purpose of the AI application. Such features include, amongst others, population and disease spectrum, specimen preparation and handling, preanalytical parameters (e.g. fixation, staining), scanning process (e.g. scanner type, resolution, compression, colour calibration) and WSI quality (e.g. stitching errors, air bubbles, and out-of-focus areas) [48].

XAI methods are important tools to consult when an AI application cannot deliver reliable results for a given input dataset. Specific reasons may be the low quality of input data or features of the input data which were not present in the training data. Furthermore, XAI methods should help to recognise so-called 'Clever Hans' predictors and detect if the model learned any shortcuts or biases [49,50]. However, in addition to the explainability of the ML model itself, for analytical performance evaluation of an AI application, explanations regarding the training dataset (including quantity, quality, uniqueness, annotation process, and scope and origin of the training data) are also needed to enable the identification of possible biases, gaps or shortcomings in the training-data, which might cause the AI-application to generate wrong results under real-world conditions.

Role of explainability and causability in clinical performance evaluation

The clinical performance of an AI application is evaluated in clinical studies where medical experts in diagnostics apply the AI application. This includes evaluating the human-AI interface for its ability to

support medical decision-making under real-world conditions. To achieve high causability, the human–AI interface should provide the user (on request) with information about the AI application’s scientific validation and performance evaluation, including comprehensible information about the algorithmic approach, the training and test datasets, as well as the reference dataset used for analytical performance evaluation. In addition, the human–AI interface shall also highlight relevant medical factors in the decision-making of the AI application to enable the medical expert to understand why the AI application generated a specific result. The human–AI interface should provide explanatory information without disturbing or decelerating the user’s primary task. It must ensure an effective mapping between explanations generated by an XAI method and the user’s previous knowledge and consider the needs and preferences of the target user group.

Conclusion and future outlook

Due to the large amounts of data available and increasing computing power, AI methods are becoming progressively successful in pathology and other fields. Currently, the most successful AI algorithms are based on the family of so-called DL algorithms. These ML algorithms, which function based on neural networks, are called ‘black boxes’ because the obtained results are practically untraceable, and it remains – even to the AI expert – often unclear why an algorithm made a particular decision.

Traditionally, AI experts evaluate ML algorithms with metrics such as accuracy and specificity. These metrics are easily measurable on test data, and if the algorithm works as expected, this evaluation is also practically sufficient. However, these metrics can be dramatically misleading if, for example, the training data are not independent and identically distributed – which is rarely the case in clinical data. As a result, problems in the relations learned by the AI can remain hidden and lead to unexpected errors. Especially in medicine, when it comes to sensitive data and decisions that directly affect people, a better understanding is necessary to avoid damage and mistakes. To make these black boxes transparent, the field of XAI has developed a set of useful methods to highlight decision-relevant parts that contributed to model accuracy in training or a particular prediction. However, a disadvantage is that these methods do not refer to a human model, i.e. a user.

For this reason, ‘causability’ was introduced in reference to the well-known term usability. Whilst XAI

means implementing transparency and traceability, causability is measuring the quality of explanations. Causability can thus be defined as ‘cause identifiability’ – the measurable extent to which an explanation of an AI’s decision to a user achieves a specified level of causal understanding with effectiveness, efficiency, and satisfaction in a specified context of use.

In the future, novel human–AI interfaces that enable contextual understanding and allow the domain expert to ask interactive ‘what-if’-questions are urgently required to achieve a high level of causability. A human-in-the-loop can (sometimes – not always) bring human experience and conceptual knowledge to AI processes – something that the current best AI algorithms (still) lack. Consequently, human medical professionals will remain important in the future and together with algorithms, will achieve better quality in their daily work. Digital pathology is a vital enabler in that direction.

Acknowledgements

Parts of this work have received funding from the Austrian Science Fund (FWF), Project: P-32554 (explainable artificial intelligence) and from the Austrian Research Promotion Agency (FFG) under grant agreement No. 879881 (EMPAIA). Parts of this work have received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 857122 (CY-Biobank), No. 824087 (EOSC-Life), No. 874662 (HEAP), and No. 826078 (Feature Cloud). Support for the writing of this article for co-authors CG, TE, TRK, NZ and RC was provided by the EMPAIA project, which is funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) under FKZ 01MK20002A. This publication reflects only the authors’ view and the European Commission is not responsible for any use that may be made of the information it contains.

Author contributions statement

All authors contributed to the concept and writing of this review article and approved the submitted and published versions.

References

1. Abels E, Pantanowitz L, Aeffner F, *et al.* Computational pathology definitions, best practices, and recommendations for regulatory

- guidance: a white paper from the Digital Pathology Association. *J Pathol* 2019; **249**: 286–294.
2. Kiehl TR. Digital and computational pathology: a specialty reimaged. In: *The Future Circle of Healthcare*, Ehsani S, Glauner P, Plugmann P, *et al.* (Eds). Springer International Publishing: Cham, 2022; 227–250.
3. Chang MC, Mrkonjic M. Review of the current state of digital image analysis in breast pathology. *Breast J* 2020; **26**: 1208–1212.
4. Viswanathan VS, Toro P, Corredor G, *et al.* The state of the art for artificial intelligence in lung digital pathology. *J Pathol* 2022; **257**: 413–429.
5. da Silva LM, Pereira EM, Salles PG, *et al.* Independent real-world application of a clinical-grade automated prostate cancer detection system. *J Pathol* 2021; **254**: 147–158.
6. Bulten W, Kartasalo K, Chen P-HC, *et al.* Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nat Med* 2022; **28**: 154–163.
7. Konnaris MA, Brendel M, Fontana MA, *et al.* Computational pathology for musculoskeletal conditions using machine learning: advances, trends, and challenges. *Arthritis Res Ther* 2022; **24**: 68.
8. Wells A, Patel S, Lee JB, *et al.* Artificial intelligence in dermatopathology: diagnosis, education, and research. *J Cutan Pathol* 2021; **48**: 1061–1068.
9. Liu X, Faes L, Kale AU, *et al.* A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* 2019; **1**: e271–e297.
10. Cifci D, Foersch S, Kather JN. Artificial intelligence to identify genetic alterations in conventional histopathology. *J Pathol* 2022; **257**: 430–444.
11. Rashidi HH, Tran NK, Betts EV, *et al.* Artificial intelligence and machine learning in pathology: the present landscape of supervised methods. *Acad Pathol* 2019; **6**: 2374289519873088.
12. McAlpine ED, Michelow P, Celik T. The utility of unsupervised machine learning in anatomic pathology. *Am J Clin Pathol* 2022; **157**: 5–14.
13. Holzinger A, Langs G, Denk H, *et al.* Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov* 2019; **9**: e1312.
14. Veta M, Heng YJ, Stathonikos N, *et al.* Predicting breast tumor proliferation from whole-slide images: the TUPAC16 challenge. *Med Image Anal* 2019; **54**: 111–121.
15. Winkler JK, Fink C, Toberer F, *et al.* Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatol* 2019; **155**: 1135–1141.
16. Schmitt M, Maron RC, Hekler A, *et al.* Hidden variables in deep learning digital pathology and their potential to cause batch effects: prediction model study. *J Med Internet Res* 2021; **23**: e23436.
17. Amann J, Blasimme A, Vayena E, *et al.* Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak* 2020; **20**: 310.
18. Cui M, Zhang DY. Artificial intelligence and computational pathology. *Lab Invest* 2021; **101**: 412–422.
19. Holzinger A. The next frontier: AI we can really trust. In: *Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Proceedings of the ECML PKDD 2021. Communications in Computer and Information Science*, Volume **1524**, Kamp M (Ed). Cham: Springer Nature, 2021; 427–440.
20. Mohseni S, Zarei N, Ragan ED. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Trans Interact Intell Syst* 2021; **11**: 1–45.
21. Mittelstadt B, Russell C, Wachter S. Explaining explanations in AI. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*; Atlanta, GA, USA, Boyd D, Morgenstern JH (Eds). Association for Computing Machinery: New York, 2019; 279–288.
22. Plass M, Kargl M, Evans T, *et al.* Human-AI interfaces are a central component of trustworthy AI. In: *Explainable AI: Foundations, Methodologies and Applications. Intelligent Systems Reference Library*, Volume **232**, Mehta M, Palade V, Chatterjee I (Eds). Cham: Springer International Publishing, 2022; 225–256.
23. Molnar C, Casalicchio G, Bischl B. Interpretable machine learning – a brief history, state-of-the-art and challenges. In: *ECML PKDD 2020 Workshops*, Koprinka I, Kamp M, Appice A, *et al.* (Eds). Springer International Publishing: Cham, 2020; 417–431.
24. Chen C, Li O, Tao D, *et al.* This looks like that: deep learning for interpretable image recognition. *Adv Neural Inf Process Syst* 2019; **32**: 8928–8939.
25. Pocevičiūtė M, Eilertsen G, Lundström C. Survey of XAI in digital pathology. In: *Artificial Intelligence and Machine Learning for Digital Pathology: State-of-the-Art and Future Challenges*, Holzinger A, Goebel R, Mengel M, *et al.* (Eds). Springer International Publishing: Cham, 2020; 56–88.
26. Samek W, Montavon G, Lapuschkin S, *et al.* Explaining deep neural networks and beyond: a review of methods and applications. *Proc IEEE* 2021; **109**: 247–278.
27. Holzinger A, Saranti A, Molnar C, *et al.* Explainable AI methods – A brief overview. In: *xxAI – Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, Holzinger A, Goebel R, Fong R, *et al.* (Eds). Springer International Publishing: Cham, 2022; 13–38.
28. Jiang J, Kahai S, Yang M. Who needs explanation and when? Juggling explainable AI and user epistemic uncertainty. *Int J Hum Comput Stud* 2022; **165**: 102839.
29. Evans T, Retzlaff CO, Geißler C, *et al.* The explainability paradox: challenges for xAI in digital pathology. *Future Gener Comput Syst* 2022; **133**: 281–296.
30. Bodria F, Giannotti F, Guidotti R, *et al.* Benchmarking and survey of explanation methods for black box models. *arXiv* 2021; <https://doi.org/10.48550/arxiv.2102.13076>.
31. Arras L, Osman A, Samek W. CLEVR-XAI: a benchmark dataset for the ground truth evaluation of neural network explanations. *Inf Fusion* 2022; **81**: 14–40.
32. Graziani M, Andrearczyk V, Marchand-Maillet S, *et al.* Concept attribution: explaining CNN decisions to physicians. *Comput Biol Med* 2020; **123**: 103865.
33. Gurumoorthy KS, Dhurandhar A, Cecchi GA, *et al.* Efficient data representation by selecting prototypes with importance weights. In: *2019 IEEE International Conference on Data*

- Mining (ICDM)*, Wang J, Shim K, Wu X (Eds). IEEE: New York, 2019; 260–269.
34. Li O, Liu H, Chen C, et al. Deep learning for case-based reasoning through prototypes: a neural network that explains its predictions. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 32, McIlraith SA, Weinberger KQ (Eds). AAAI Press: Palo Alto, CA, 2018; 3530–3537.
 35. Stepin I, Alonso JM, Catala A, et al. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access* 2021; **9**: 11974–12001.
 36. Tagasovska N, Lopez-Paz D. Single-model uncertainties for deep learning. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Wallach HM, Larochelle H, Beygelzimer A, d'Alche-Buc F, Fox EB (Eds). Curran Associates Inc.: Vancouver, 2019; article 576.
 37. Lipton ZC. The mythos of model interpretability. *Commun ACM* 2018; **61**: 36–43.
 38. Holzinger A, Kargl M, Kipperer B, et al. Personas for artificial intelligence (AI) an open source toolbox. *IEEE Access* 2022; **10**: 23732–23747.
 39. Chou Y-L, Moreira C, Bruza P, et al. Counterfactuals and causability in explainable artificial intelligence: theory, algorithms, and applications. *Inf Fusion* 2022; **81**: 59–83.
 40. Gunning D, Aha D. DARPA's explainable artificial intelligence (XAI) program. *AI Mag* 2019; **40**: 44–58.
 41. Holzinger A, Müller H. Toward human–AI interfaces to support explainability and causability in medical AI. *Computer* 2021; **54**: 78–86.
 42. Girardi D, Küng J, Kleiser R, et al. Interactive knowledge discovery with the doctor-in-the-loop: a practical example of cerebral aneurysms research. *Brain Inform* 2016; **3**: 133–143.
 43. European Commission. Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on in vitro diagnostic medical devices and repealing Directive 98/79/EC and Commission Decision 2010/227/EU, 2017.
 44. Müller H, Holzinger A, Plass M, et al. Explainability and causability for artificial intelligence-supported medical image analysis in the context of the European In Vitro Diagnostic Regulation. *N Biotechnol* 2022; **70**: 67–72.
 45. Li LT, Jiang G, Chen Q, et al. Ki67 is a promising molecular target in the diagnosis of cancer (review). *Mol Med Rep* 2015; **11**: 1566–1572.
 46. Wulczyn E, Steiner DF, Moran M, et al. Interpretable survival prediction for colorectal cancer using deep learning. *NPJ Digit Med* 2021; **4**: 71.
 47. Medical Device Coordination Group. MDCG 2020-1 Guidance on Clinical Evaluation (MDR)/Performance Evaluation (IVDR) of Medical Device Software, 2020.
 48. Homeyer A, Geißler C, Schwen LO, et al. Recommendations on compiling test datasets for evaluating artificial intelligence solutions in pathology. *Mod Pathol* 2022; **35**: 1759–1769.
 49. Geirhos R, Jacobsen J-H, Michaelis C, et al. Shortcut learning in deep neural networks. *Nat Mach Intell* 2020; **2**: 665–673.
 50. Lapuschkin S, Wäldchen S, Binder A, et al. Unmasking Clever Hans predictors and assessing what machines really learn. *Nat Commun* 2019; **10**: 1096.