Reproducibility and explainability in digital pathology: The need to make black-box artificial intelligence systems more transparent

Gavino Faa[1], Matteo Fraschini[2] and Luigi Barberini[2] iD

Abstract

Artificial intelligence (AI), and more specifically Machine Learning (ML) and Deep learning (DL), has permeated the digital pathology field in recent years, with many algorithms successfully applied as new advanced tools to analyze pathological tissues. The introduction of high-resolution scanners in histopathology services has represented a real revolution for pathologists, allowing the analysis of digital whole-slide images (WSI) on a screen without a microscope at hand. However, it means a transition from microscope to algorithms in the absence of specific training for most pathologists involved in clinical practice. The WSI approach represents a major transformation, even from a computational point of view. The multiple ML and DL tools specifically developed for WSI analysis may enhance the diagnostic process in many fields of human pathology. AI-driven models allow the achievement of more consistent results, providing valid support for detecting, from H&E-stained sections, multiple biomarkers, including microsatellite instability, that are missed by expert pathologists.

Artificial intelligence (AI), and more specifically Machine Learning (ML) and Deep learning (DL), has permeated the digital pathology field in recent years, with many algorithms successfully applied as new advanced tools to analyze pathological tissues.[1]

The introduction of high-resolution scanners in histopathology services has represented a real revolution for pathologists, allowing the analysis of digital whole-slide images (WSI) on a screen without a microscope at hand. However, it means a transition from microscope to algorithms in the absence of specific training for most pathologists involved in clinical practice. The WSI approach represents a major transformation, even from a computational point of view. The multiple ML and DL tools specifically developed for WSI analysis may enhance the diagnostic process in many fields of human pathology. AI-driven models allow the achievement of more consistent results, providing valid support for detecting, from H&E-stained sections, multiple

biomarkers, including microsatellite instability, that are missed by expert pathologists.[2]

Despite all these possible advantages and promising results, the introduction of AI-driven tools in clinical practice needs to be revised. This is due to multiple reasons. The reproducibility of DL models applied to WSI analysis represents a crucial point, and often a barrier, for the transition of these models from research to clinical workflows. For a method to be widely adopted in clinical practice, it must be explainable and reproducible so that pathologists

[1]Dipartimento di Scienze mediche e sanità pubblica, Università degli Studi di Cagliari, Cagliari, Italia
[2]Dipartimento di Ingegneria Elettrica ed Elettronica, Università degli Studi di Cagliari, Cagliari, Italia

Corresponding author:
Luigi Barberini, Dipartimento di Ingegneria Elettrica ed Elettronica, Università degli Studi di Cagliari, via Ospedale, 54, Cagliari 09123, Italia.
Email: luigi.barberini@unica.it

can have confidence in its use.[3] Unfortunately, ML and DL models are characterized by crucial challenges regarding reusability and reproducibility.[4] It is time to rethink the approach of AI to pathology, aiming to help algorithms reach the levels of reproducibility and availability necessary for approval by national and international authorities, such as the European Medicines Agency (EMA) and the Food and Drugs Administration (FDA).

Here, we would start from the etymology of the term "algorithm." The word algorithm comes from the Muslim mathematician Muhammead Ibn Musa al-Khwarizmi, born in Uzbekistan around 780 CE, who is credited with inventing algebra and developing the concept of algorithms. These are systematic methods with a sequence of steps and rules, which end with solving mathematical problems. In their original definition, algorithms were characterized by their explainability and reproducibility. Unfortunately, the current status of the vast majority of algorithms applied to computational pathology is characterized by low levels of reusability and reproducibility. When applied to the local dataset, models with high specificity and sensibility often show lower performance when applied to external datasets, evidencing the inability of these models to explain the theory.

The analysis of the multiple steps utilized by AI models in the WSI analysis evidences multiple critical points: stain normalization of tissue sections, tissue type segmentation, type of patch extraction, whole-slide image-based classification versus patch-based analysis and mixed methods, hard negative mining, heatmap generation are among the multiple critical points that characterize the application of a DL model to histopathology in clinical workflows.[4] The failure to maintain high-level standards regarding data processing, an essential requisite for reproducibility, characterizes most studies on DL models in digital pathology.[3]

Another critical point is the absence of explainability and interpretability of these models, which appear as "black boxes."[5] Although convolutional neural networks (CNNs) have achieved impressive performance, it is more intriguing to understand how the models make decisions and how they learn to solve a given task.[6] The absence of algorithm elucidation hinders the medical acceptance of AI models.[7] From a practical point of view, in many countries, a clarification of how AI models work is required for governmental approval for use in clinical settings.[8]

In the medical community, understanding the decision-making process can be as important as the decision itself.[9] When dealing with a patient, a disease or a complex diagnosis, which can end with decisions that directly affect a human's health status and survival, a better understanding is necessary to avoid damage, adverse effects, and mistakes. For this reason, a better understanding of "algorithmic decisions" appears mandatory

for pathologists utilizing AI models, addressing the need for explainability in digital pathology.[10]

These data, taken together, are ready for the proposal of a high-quality, robust, easy-to-use and transparent processing pipeline, which can help ensure the validity and the explainability of AI models applied to histopathology in clinical workflows. The main goal of a new robust pipeline is to overcome the reproducibility crisis of AI models,[11] eventually allowing their faster applications in medicine and their acceptance by pathologists for clinical purposes.[12] To this end, novel pathologist-AI interfaces that refer to a human user enable contextual understanding and allow pathologists to ask interactive questions, overcoming the disadvantages of the actual AI models, which do not refer to a human model.

Declaration of conflicting interests

Funding

ORCID iD

Luigi Barberini https://orcid.org/0000-0002-3815-0392

References

1. Faa G, Castagnola M, Didaci L, et al. The quest for the application of artificial intelligence to whole slide imaging: unique prospective from new advanced tools. *Algorithms* 2024; 17(6): 254.
2. Reis-Filho JS and Kather JN. Overcoming the challenges to implementation of artificial intelligence in pathology. *J Natl Cancer Inst* 2023; 115(6): 608–612.
3. Fell C, Mohammadi M, Morrison D, et al. Reproducibility of deep learning in digital pathology whole slide image analysis. *PLoS Digit Heal* 2022; 1(12): 1–21.
4. Wagner SJ, Matek C, Shetab Boushehri S, et al. Built to last? Reproducibility and reusability of deep learning algorithms in computational pathology. *Mod Pathol* 2024; 37(1): 1–11. DOI: 10.1016/j.modpat.2023.100350
5. Rudin C. Stop explaining Black Box Machine Learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 2019; 1(5): 206–215.
6. Chen C-L, Chen C-C, Yu W-H, et al. An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. *Nat Commun* 2021; 12(1): 1193.
7. Ahmed AA, Abouzid M and Kaczmarek E. Deep learning approaches in histopathology. *Cancers* 2022; 14(21): 5264.
8. Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface* 2018; 15(141): 1–47.

9. Lee M. Recent advancements in deep learning using whole slide imaging for cancer prognosis. *Bioengineering* 2023; 10(8): 897.

10. Plass M, Kargl M, Kiehl T-R, et al. Explainability and causability in digital pathology. *J Pathol Clin Res* 2023; 9(4): 251–260.

11. Hutson M. Artificial intelligence faces reproducibility crisis. *Science* 2018; 359(6377): 725–726.

12. Stodden V, McNutt M, Bailey DH, et al. Enhancing reproducibility for computational methods. *Science* 2016; 354(6317): 1240–1241.