Counterfactual Diffusion Models Provide Interpretable Explanations of Artificial Intelligence Models in Pathology

Info

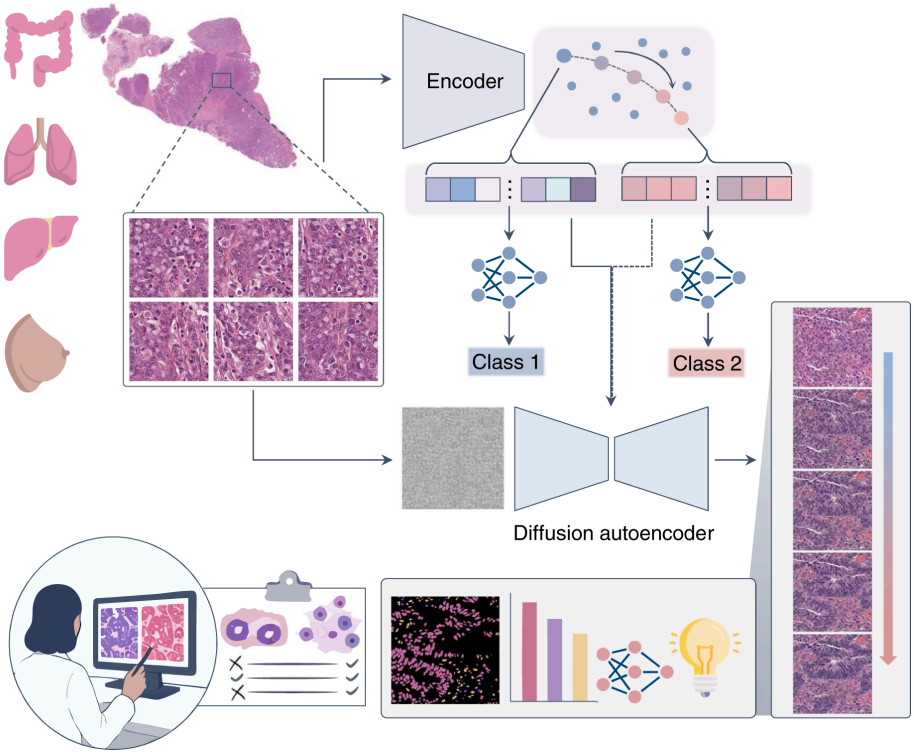
PUBLISHED DATE
2026-06-17

JOURNAL
Cancer Research

LINK
https://doi.org/10.1158/0008-5472.can-25-5645

Authors

Laura Žigutytė Tim Lenz Tianyu Han
Nic G. Reitsam Sebastian Foersch
Katherine J. Hewitt Moritz Jesinghaus
Zunamys I. Carrero Michaela Unger
Asier Rabasco Meneghetti Georg Lurje
Isabella Lurje Sophia Herda Justus Pein
Deniz Uluk Carolin V. Schneider
Alexander T. Pearson Daniel Truhn
Jakob Nikolas Kather



Abstract

Deep learning can extract predictive and prognostic biomarkers from histopathology whole-slide images. However, explainable artificial intelligence approaches widely used in digital pathology, such as attention heatmaps and class activation mapping, provide limited insight into the image features associated with classifier outputs. In this study, we developed Morphing histoPathology Diffusion (MoPaDi), a framework for generating counterfactual explanations for histopathology images that help identify morphologic or stain-

related features linked to model predictions. MoPaDi combined diffusion autoencoders with task-specific multiple instance learning classifiers to manipulate images and induce prediction shifts by modifying classifier-associated features. The framework was evaluated on multiple datasets spanning colorectal, breast, liver, and lung cancers, including tasks for tissue type, cancer subtype, and biomarker [microsatellite instability (MSI)] classification. MoPaDi generated perceptually realistic counterfactual histopathology images, enabling pathologists to identify morphologic features associated with changes in model predictions, complementing the conventional inspection of highly attended regions in digital pathology. In the MSI status prediction task, MoPaDi highlighted morphologic features linked to classifier predictions, including mucinous differentiation, altered glandular architecture, and lymphocytic infiltration, consistent with prior literature. Analyses separating stain-related from morphology-related components suggested that in this setting, prediction changes were predominantly associated with morphology-related rather than stain-related alterations. Overall, MoPaDi is a practical framework for counterfactual explanations in computational pathology that supports the evaluation of model-specific decision cues and hypothesis generation.