SIIM CAIMI26 Conference on Artificial Intelligence in Medical Imaging
OCT 26-27 | University of Pennsylvania, Philadelphia, PA

Analysis of histopathology data drift in an eight-year cohort reveals laboratory and instrumentation-induced variability

Mathew Francis, PhD, Perelman School of Medicine, University of Pennsylvania; Arian Arab, PhD; Rohan G. Chandraghatgi; Wenchao Han, PhD; Christopher A. Garcia, MD; David E. Breen, PhD; Fernando U. Garcia, MD; Mark Zarella, PhD

Introduction/Background

AI in digital pathology is vulnerable to data drift (DD) caused by changes in laboratory preparation, image acquisition, and other factors. Continuous monitoring is therefore necessary, yet reliable methods for tracking DD in histopathology remain limited. Many prior studies rely on retrospectively scanned slides, making it difficult to distinguish true DD from glass-slide degradation over time.

Methods/Intervention

We analyzed 5,772 Ki67-stained whole-slide images (WSIs) from breast cancer patients collected over eight years within a real-world clinical workflow. Two symmetric autoencoders were trained—using 128- and 512-bottleneck architectures—on a dataset of 494 WSIs from 2016 (original dataset, OD), Figure 1a. These models were then applied to WSIs from 2017 to 2023 (n = 4,785, experimental dataset), Figure 1b. Wasserstein distance (WD) of perceptual loss was used to quantify deviations from the OD distribution. As negative controls, we used 493 randomly held-out slides from 2016. A positive control consisted of 52 additional WSIs scanned with a different slide scanner to isolate scanner effects. Additionally, WSIs stained with Ki67, ER, and HER2 (921 each) from another laboratory using the same scanner were included as external references to assess inter-laboratory and staining variability. Clinically significant DD was defined as a non-overlapping 95% confidence interval of the bootstrapped WD distribution.

Results/Outcome

Both autoencoders reconstructed images effectively, with the 512-bottleneck model capturing more detail (Figure 2). The positive control was clearly distinguished from the negative control, demonstrating the model's ability to detect meaningful differences (Figure 3). The WD of perceptual loss increased significantly after 2021, indicating clinically relevant DD compared to the 2016 OD. This drift was comparable to variability observed in reference datasets stained with Ki67 or ER but much smaller than variation introduced by scanner differences.

Conclusion

This study provides the first direct, model-agnostic demonstration of DD in digital pathology. The findings show that clinical data may move outside the distribution used for AI validation, effectively turning prospective use into an unvalidated test, thus necessitating continuous monitoring and periodic revalidation to maintain reliable, robust clinical AI performance.

Statement of Impact

Implementing continuous, model-agnostic surveillance strategies in digital pathology holds promise for enabling reliable, robust AI deployment in clinical practice.

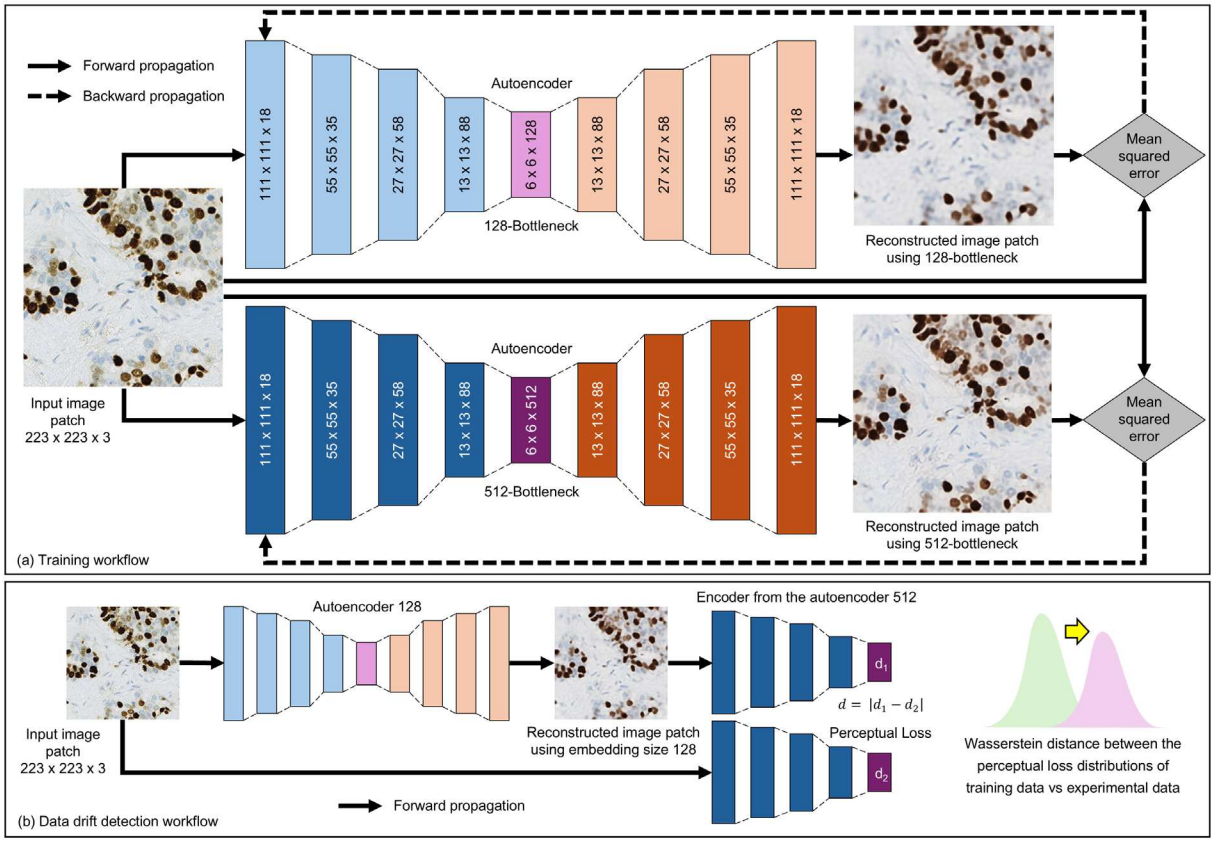

Figure 1: Schematic diagram of the autoencoder training and data drift detection workflows. (a) Two autoencoders with 128- and 512-bottleneck architectures were trained using approximately half a million Ki67-stained patches from 494 original whole-slide images (WSI) collected in 2016. (b) Inference pipeline using the 512-bottleneck encoder to compute perceptual loss between raw inputs and 128-bottleneck reconstructions across 8.29 million patches. Wasserstein distances were calculated between the original dataset and different datasets.

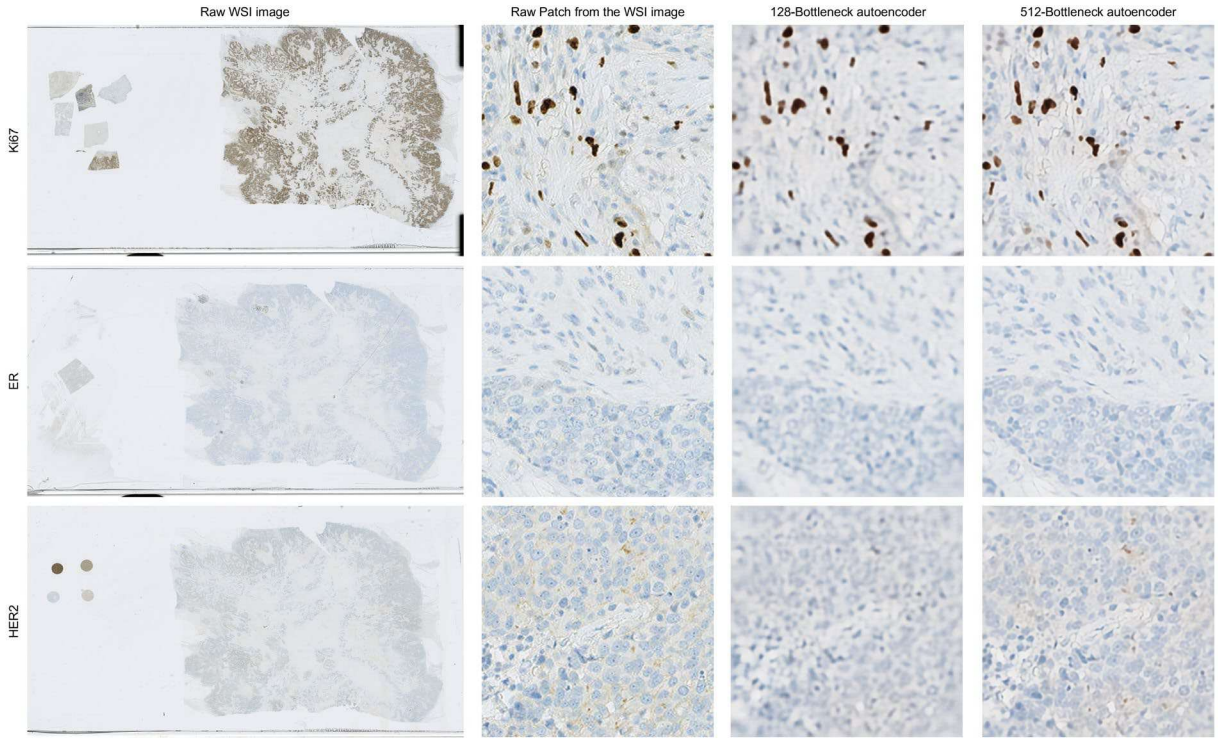

Figure 2: Examples of raw whole-slide images and corresponding autoencoder reconstructions from both 128- and 512-bottleneck models. Rows correspond to Ki67, ER, and HER2 stains, respectively. Columns show raw WSI, extracted patches, and reconstructions from the 128- and 512-bottleneck encoders.

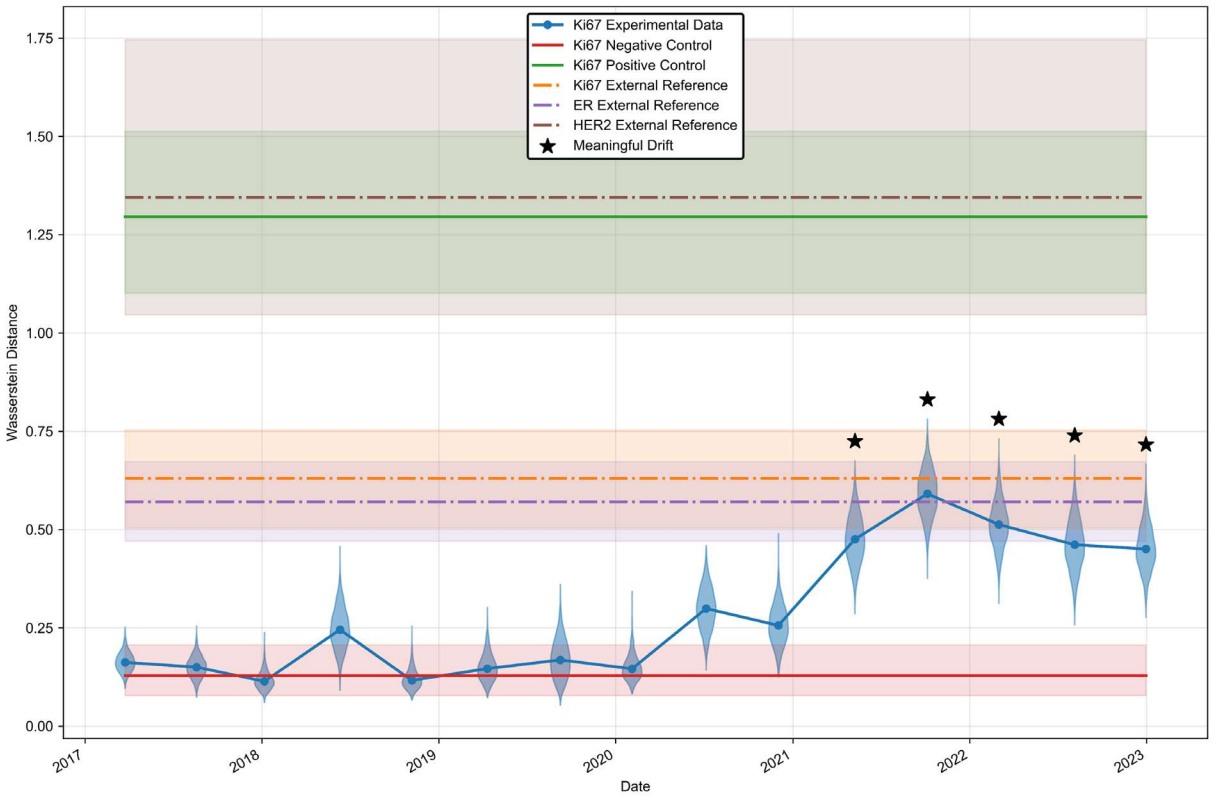

Figure 3: Chronological bootstrapped Wasserstein distances between the original dataset (2016) and experimental datasets from 2017 to 2023. Significant drift becomes apparent after 2021, illustrated in relation to the negative control, positive control, and external reference datasets (Ki67, ER, and HER2).

Keywords

Data Drift; Digital Pathology; Deep Learning; Medical Imaging