

ContriMix

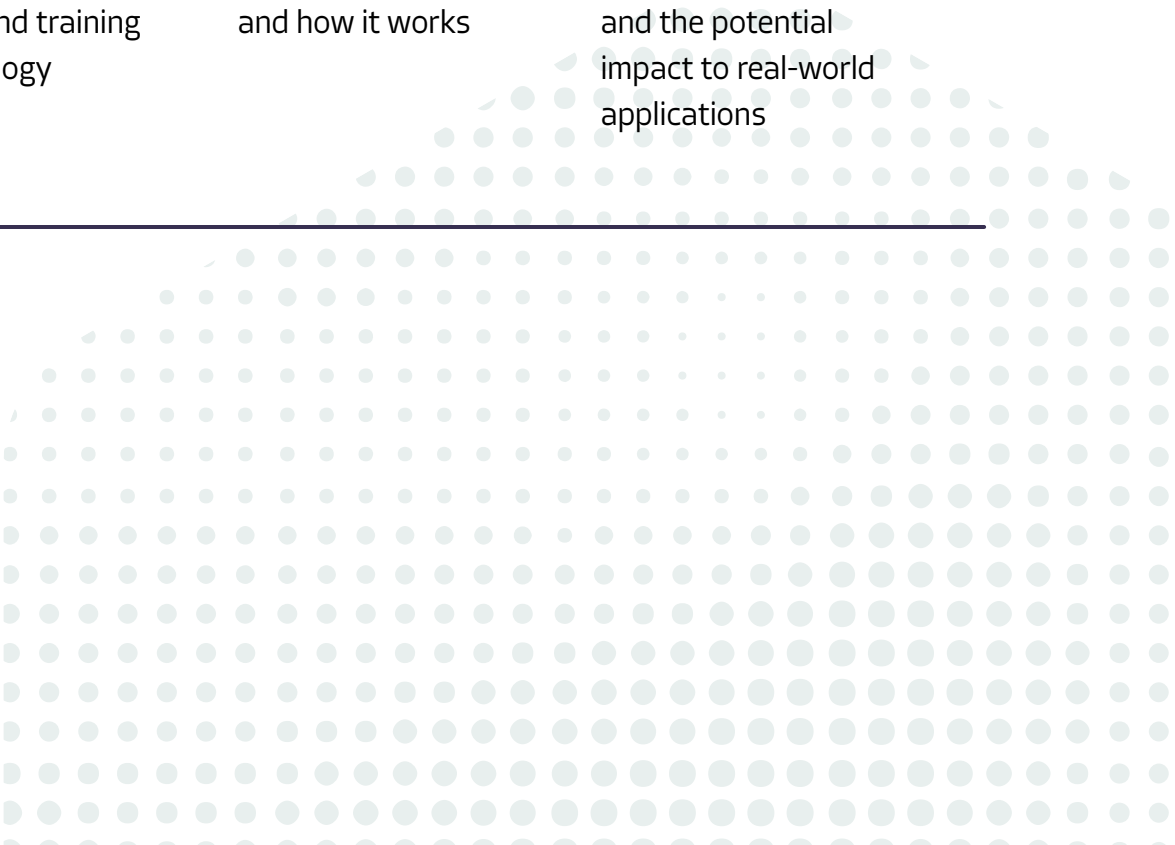
ContriMix is a powerful machine learning technique developed by PathAI to enable the development of more accurate and generalizable digital pathology algorithms at greater speed and scale.

This article provides an overview of:

The challenges with developing and training digital pathology algorithms

What Contrimix is and how it works

Results from Contrmix and the potential impact to real-world applications



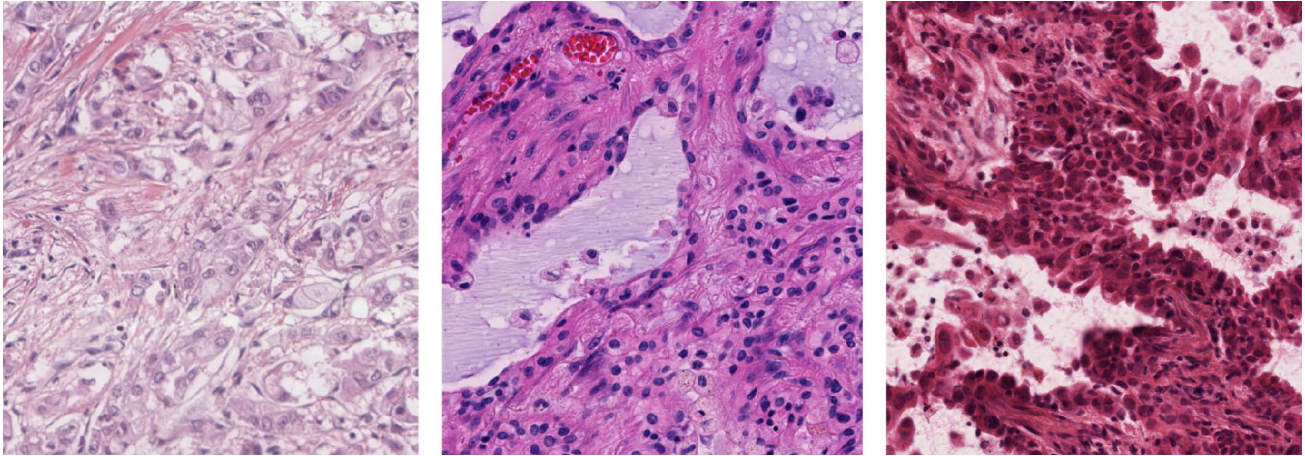


Figure 1

Figure 1: Differences in staining and preparation characteristics may result in algorithms confusing biological vs. non-biological features, which can reduce the ability of AI systems to make accurate and reproducible predictions for research or clinical care. The three images in Figure 1 all come from TCGA Lung Adenocarcinoma slides, at the same scanning resolution, and scanned on the same scanner, yet wide variation remains in the image characteristics that aren't captured simply by labeled domains and not all images have all domains readily available.

I. Challenges in digital pathology algorithm development

Training AI digital pathology algorithms today requires an immense amount of inputs to capture all the permutations of staining and scanning variability so that models can distinguish between actual biological information vs. incidental image characteristics that aren't related to the biology.

These limitations have resulted in: Timely and costly algorithm development, such as requiring the collection of pathologist annotations on slides and costs to source a wide array of images and inputs Specific scanner and stain compatibility, which limits real-world use of algorithms to workflows that utilize those specific requirements. This ultimately prevents our ability scale algorithms and maximize impact to patients.

II. PathAI's Solution: ContriMix (Unsupervised Content and Attribute Mixing for Domain Generalization)

1. How Does this work?

ContriMix learns to separate the biological content of images from the noise or attribute by leveraging specially designed architectures and losses. The architecture enforces the concept that images are a mix of noise (from differences in staining and scanning techniques) and content signals and that these are naturally separable, consistent with human performance. ContriMix's specifically-designed losses enforce that the biological content extracted from the image is consistent regardless of the noise it is mixed with, and that it remains faithful to the biological content of the original image. We also enforce that the attribute or noise is consistent when mixed with various content examples. This combination of consistency and reconstruction losses pushes the network to learn automatically what are the noise signals which are present across batches. ContriMix performs this disentanglement without needing to know the origin of the image (e.g, scanner or hospital), similar to a human domain expert.

Interested in more technical explanation of ContriMix?
Check out our recent publication [here](#)

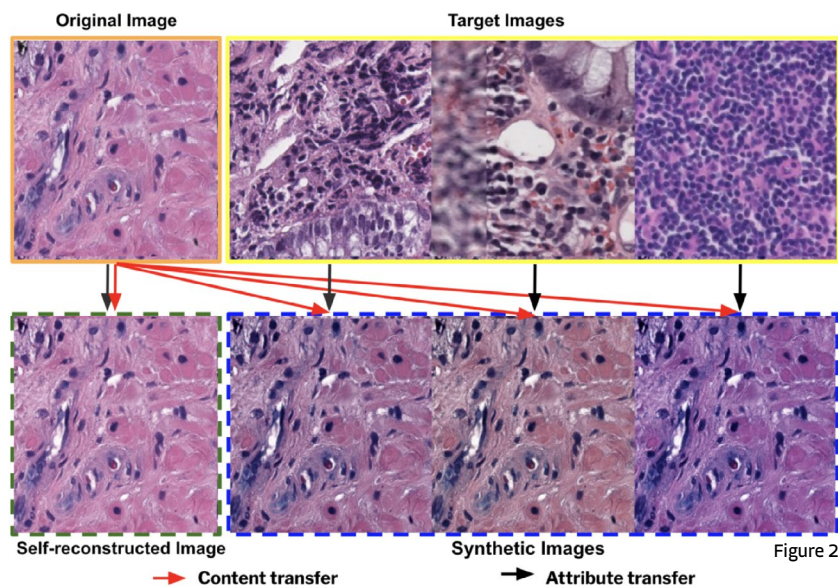


Figure 2: The “content” of the H&E patch on the upper left was mixed with the “attributes” of three other patches on the right, which arise from different slides. The result is four patches with the same biological content, but a subtly different appearance. Human pathologists learn to ignore these non-biological differences in visual presentation; by generating these variations, we train our models to do the same. Contrimix was able to correctly transfer the color even when the input image is heavily affected by blur (3rd column, top row).

III. Results & Impact to Digital Pathology

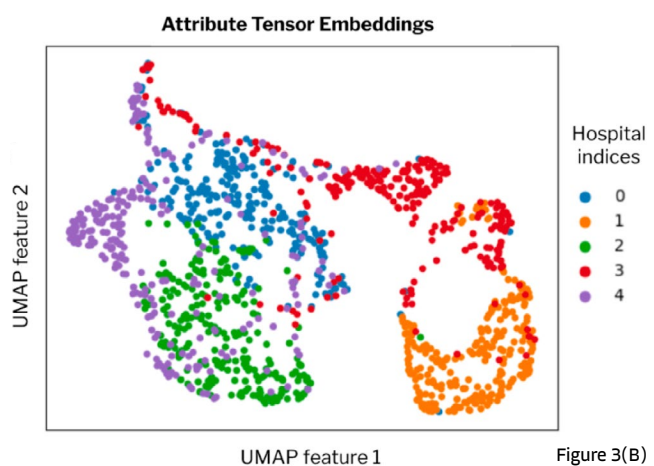
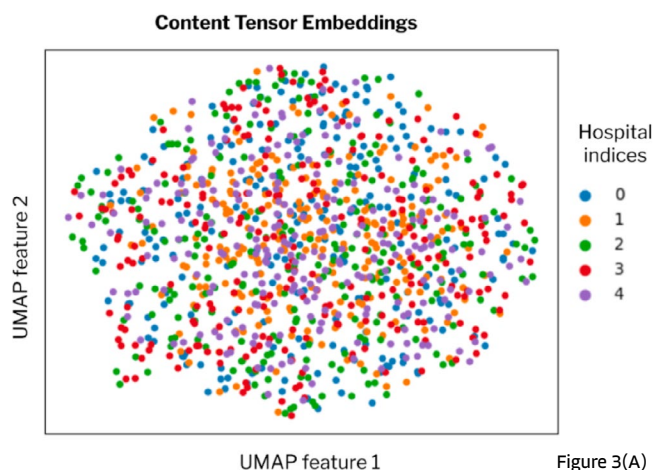
When applying Contrimix to whole slide images, features that Contrimix classifies as “biological” (content) have no relationship to the hospital where the slide was prepared. This is exactly what we want to see as there should be complete randomness across which biological features are contained across samples from various institutions. See figure 3a.

But the features that Contrimix classifies as “noise” (attribute) show that there are distinct clusters of characteristics attributable to each institution. In other words, we have the ability to predict common staining and scanning characteristics to the institution where the slide was prepared. See figure 3b.

Contrimix has ranked first among algorithms developed for digital pathology data on the public [Stanford WILDS Camelyon17](#) dataset.

Why this matters

Contrimix is powerful in that it allows us to build models that “see” the underlying biology without undue influence by incidental image characteristics – and to do so more rapidly and with more varied unlabeled data. Ultimately, improved generalization to the varied data in the real world is critical for achieving our goal of using AI-powered pathology to improve patient outcomes.



Interested in learning more? Set up time with our machine learning team by reaching out to BD@pathai.com