IDC Image Preprocessing and Augmentation Pipeline

This document explains the preprocessing and augmentation pipeline used for the IDC (Invasive Ductal Carcinoma) histopathology image dataset. The goal was to enhance tissue visibility, balance class representation, and ensure model robustness.

---

1. IMAGE PREPROCESSING

The raw IDC dataset contains 50x50 RGB histopathology patches. The preprocessing steps below were applied to all images and saved into the directory '/content/final_preprocessed'.

STEPS:

1. Mild CLAHE (Contrast Limited Adaptive Histogram Equalization)
   - clipLimit = 1.0
   - tileGridSize = (8,8)
   - Purpose: Enhance local contrast without over-sharpening the tissue texture.

2. Soft Circular Blend
   - feather_px = 10-12, softness = 0.85-0.9
   - Removes hard borders and softly blends tissue into background, focusing the network on central tissue regions.

3. Professional Color Balance
   - Mild lift in white, pink, and purple tones.
   - Purpose: Enhance histopathological features (nuclei, stroma, adipose regions) while preserving natural H&E stain distribution.

---

2. ON-THE-FLY AUGMENTATION

Augmentation is applied dynamically during training - images are NOT saved after augmentation. This ensures the model sees new, unique variations every epoch.

AUGMENTATIONS USED:

1. Random Flip
   - Type: Horizontal and Vertical
   - Purpose: Simulate microscope slide rotation or flipping.

2. Random Rotation
   - Range: +-8% (~+-30 degrees)
   - Purpose: Improve rotational invariance.

3. Random Zoom
   - Range: +-8%
   - Purpose: Emulate magnification variations under microscope.

---
3. WHY THIS APPROACH?

- Medical datasets often suffer from class imbalance and limited variability.
- On-the-fly augmentation generates effectively infinite diverse samples.
- Preprocessing improves clarity and reduces background noise.
- The combination improves model generalization and recall for IDC-positive samples.

---
4. OUTPUT

- Preprocessed dataset stored in '/content/final_preprocessed'

- Training data augmented dynamically via TensorFlow during model fit()
- Final CNN trained with improved tissue visibility and pattern diversity.

Summary:
This preprocessing + augmentation pipeline mimics a real-world pathology workflow, producing visually enhanced, stain-balanced, and rotation-invariant images.
It forms the foundation for our robust 5-layer CNN classification performance.