A Novel AI-Powered Preprocessing Framework for Maximizing Whole-Slide Image Dataset Quality for Machine Learning Applications

Rathinaraja Jeyaraj[1], Barathi Subramanian[1], Songmi Noh[2], Mitch Peterson[1], Terry Guo[1], George Fisher[3], Jeanne Shen[1]

[1]Department of Pathology and [3]Department of Medicine, Stanford University, USA, [2]Department of Diagnostic Pathology, CHA University College of Medicine, Republic of Korea

Contact: rajaj@stanford.edu, jeannes@stanford.edu

Background

- Tissue identification in H&E-stained whole-slide images (WSIs) is a critical first step for downstream machine learning (ML) workflows for both research and clinical applications.
- A large portion of WSI contain less-informative regions (e.g., non-tissue background, pen marks, artifacts, low-tissue content, as shown in Figure 1), which degrade ML model performance [1].

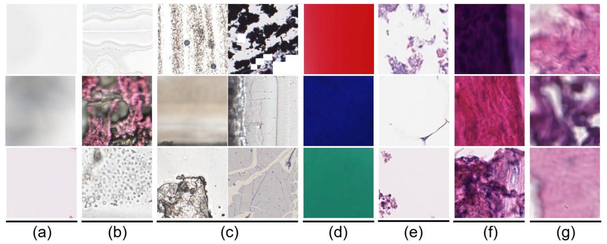


Figure 1: (a) non-tissue background, (b) air bubble, (c) dust/debris, (d) pen marks, (e) low-tissue content, (f) tissue folds, and (g) blur

- Dataset quality is essential for accurate and efficient ML training

Research gap: Existing pre-processing methods [2], [3], [4], [5] lack generalizability and rely on manual RGB-based filtering, which fails to consistently eliminate less-informative regions.

Proposed Framework

An AI-powered preprocessing framework with four primary phases (Figure 2) is proposed to enhance ML model performance:

- **Phase 1** – Image tiles are extracted from WSI, discarding the pen mark (red, green, and blue, etc.)-containing tiles
- **Phase 2** – Applies optical density analysis to remove tiles with mean eosin values below a defined threshold
- **Phase 3** – Utilizes an AutoEncoder (Figure 3) to remove artifacts and organize tiles based on reconstruction loss into four groups:
  - Informative_set1 – high-quality (tissue-rich) tiles (>55% tissue/tile)
  - Informative_set2 – partial tissue content
  - Informative_set3 – adipose-like tissue
  - Less_informative – Low-quality (artifacts) and low-tissue content
- **Phase 4** – A classifier recovers adipose and mucin tiles excluded during Phase 2

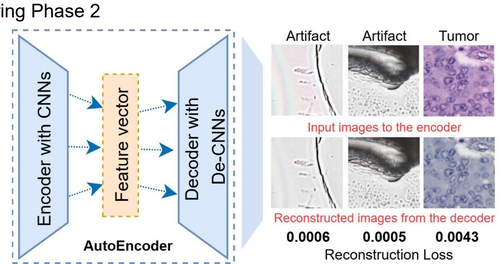


Figure 3: AutoEncoder to reject less informative tiles.

Flowchart of the Proposed Framework

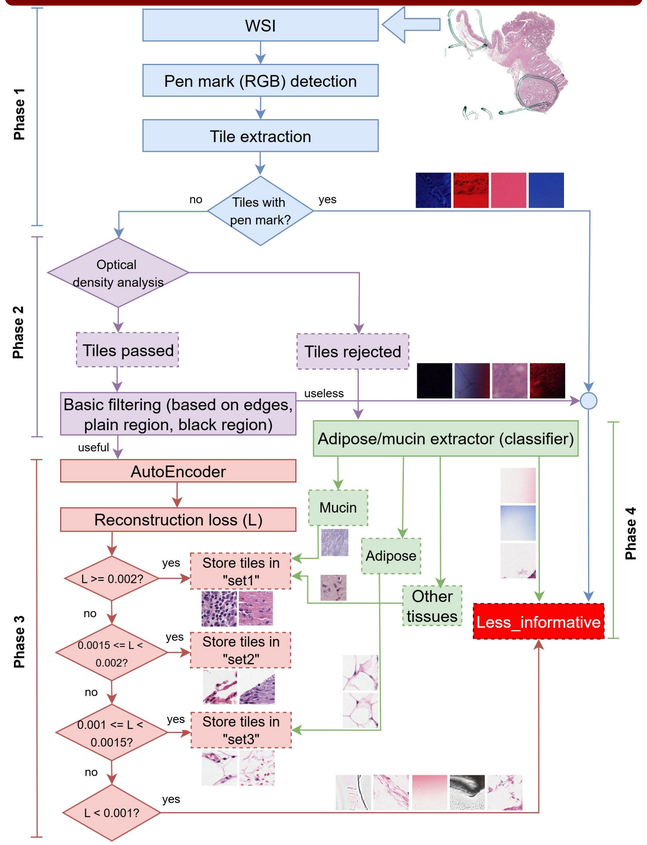


Figure 2: Overall workflow of the proposed framework.

Results

- Pre-processing performance is evaluated on a survival analysis task using Stanford and TCGA-COAD-READ datasets [2].
- SOTA models: ABMIL, CLAM-MB, TransMIL, MambaMIL, DTFD-MIL, and PatchGCN use CLAM and/or basic pre-processing.

Model	Stanford (C-index %)		TCGA-COAD-READ (C-index %)	
	Basic/CLAM	Infor_set1	Basic/CLAM	Infor_set1
ABMIL	0.541±0.023	0.565±0.038	0.547±0.087	0.582±0.087
CLAM-MB	0.529±0.010	0.540±0.031	0.516±0.069	0.557±0.056
TransMIL	0.528±0.023	0.552±0.029	0.503±0.076	0.662±0.070
PatchGCN	0.534±0.002	0.556±0.018	0.529±0.035	0.581±0.024
DTFD-MIL	0.501±0.019	0.527±0.030	0.502±0.042	0.587±0.280
MambaMIL	0.539±0.031	0.564±0.052	0.536±0.059	0.655±0.081

Table 1: Survival analysis on Stanford and TCGA-COAD-READ datasets.

Output Visualization with Masked Overlay

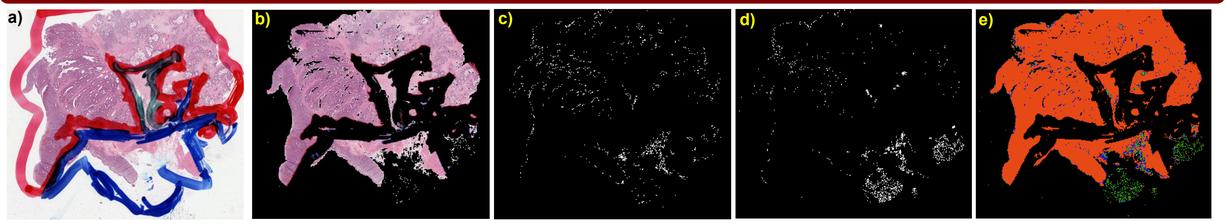


Figure 4: Proposed framework: (a) Input WSI, (b) Informative_set1, (c) Informative_set2, (d) Informative_set3, (e) Full mask with orange color (set1), blue color (set2), and green color (set3)

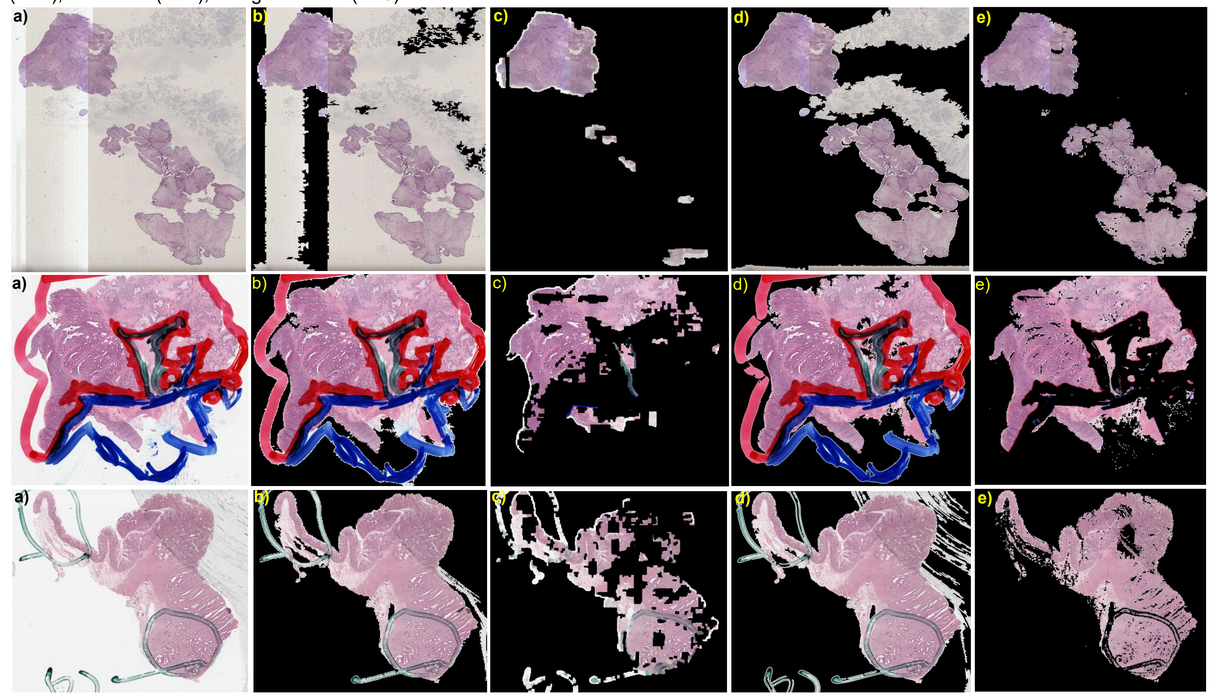


Figure 5: Masked overlay comparison: (a) Input WSI, (b) CLAM (2021), (c) Trident (2024), (d), HistoLab (2025) (e) Proposed informative_set1.

- Tiles from basic or CLAM-based preprocessing (commonly used in most models) and Informative_set1 (from our framework) were used for evaluation (Table 1).
- We observed empirically that tiles in Informative_set1 are the primary contributors to downstream ML applications.
- Our proposed framework improved prognostic accuracy, boosting the concordance index (C-index) by over 2% (Stanford) and 4% (TCGA-COAD-READ), when compared to existing preprocessing methods.

Conclusion

- The proposed AI-based preprocessing framework improves WSI dataset quality by removing less-informative tiles.
- This framework maximizes image dataset quality for ML applications such as survival prognostication and is potentially applicable to a broad range of H&E-based computational pathology tasks.

References

1. Kanwal, N., et al. (2022). The devil is in the details. IEEE Access.
2. Janowczyk, A., et al. (2019). HistoQC. JCO Clinical Cancer Infor.
3. Pati, P., et al. (2022). PathML. Nature Communications.
4. Graham, S., et al. (2019). Hover-Net. Medical Image Analysis.
5. Lu, M.Y., et al. (2021). CLAM. Nat Biomed Eng.