Generating Morphologically Diverse, Class-Balanced Training Datasets for Enhanced Histopathologic Tissue-Type Classification

[1]Barathi Subramanian, [2]Rathinaraja Jeyaraj, [3]Songmi Noh, [4]Mitch Peterson, [5]Terry Guo, [6]George Fisher, [7]Jeanne Shen

([1,2,4,5,7]Department of Pathology, [6]Department of Medicine), Stanford University, USA [3]Department of Diagnostic Pathology, CHA University College of Medicine, Republic of Korea

Contact: [1]barathi1@stanford.edu, [7]jeannes@stanford.edu

Background

- With the growing adoption of digital pathology, computational approaches, leveraging deep learning, will play an important role in pathology workflows.
- Multi-class tissue classification represents one such foundational task in pathology [1], enabling models to distinguish between diverse tissue types.
- However, publicly available datasets used to build tissue classifiers often suffer from insufficient morphologic diversity, class imbalance, and low-quality image tiles, limiting downstream model performance and generalizability.
- To address this research gap, we propose a framework (Figure 1) to build a large-scale dataset for multi-class tissue classification, reducing redundancy and selecting maximally representative tiles for downstream machine learning tasks.

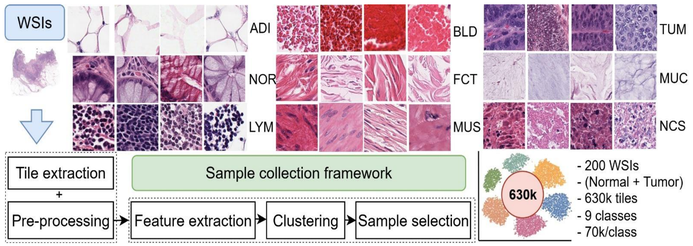


Figure 1: Overview of large-scale dataset construction. adipose tissue (ADI), lymphoid tissue (LYM), muscle (MUS), fibroconnective tissue (FCT), mucin (MUC), necrosis (NCS), blood (BLD), tumor (TUM), and normal mucosa (NOR)

Research Gap

- Publicly available CRC datasets, such as NCT-CRC-HE-100K [2], HMU-GC-HE-30K [3], and TCGA COAD and READ [4], have advanced research but remain insufficient for generalizable models due to:
  - Limited morphologic diversity/severe class imbalance
  - Low-quality or mislabeled tiles, as shown in Figure 2 [5].
- Techniques like cross-entropy uncertainty and probabilistic local outlier detection can improve label quality, no cohesive pipeline exists for large-scale, balanced dataset curation.
- Creating new datasets is labor-intensive, with no standardized method to ensure diversity or quality, posing a major challenge to robust tissue classification model development impacting model accuracy.

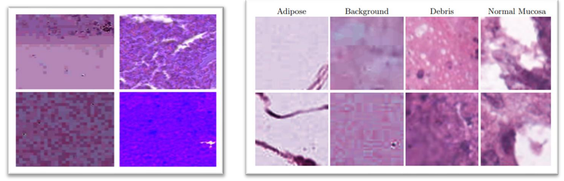


Figure 2: [5] Corrupt images (left) and low quality images (right) such as blurry, artifact like pen marks, and non-representative tiles from NCT-CRC-HE-100K dataset.

Results

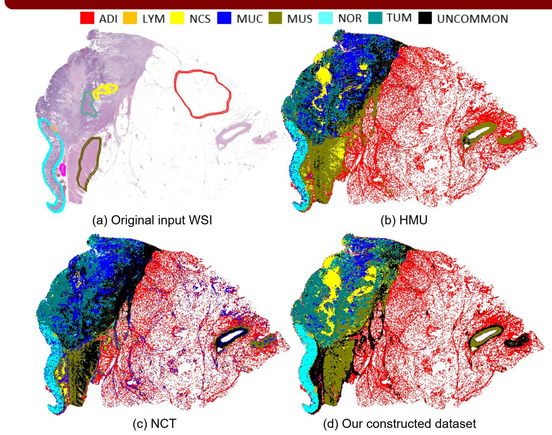


Figure 4: Tile-level predictions from models trained on different datasets are overlaid on WSIs for common tissue types.

- Across all common tissue classes (Figure 4), models trained on our dataset consistently outperformed those trained on HMU or NCT, achieving >98% accuracy (Table 1).
- Our models also produced significantly more accurate and contextually relevant predictions, particularly in complex tissue regions (Figure 4).

Models trained on NCT and HMU, our framework constructed dataset showed:

- >45% and >90% higher accuracy for NCS tissue identification.
- >80% reduction in misclassification among classes (NOR, TUM, and MUC).
- >85% accuracy on mixed tissue-type tiles, outperforming HMU (55%) and NCT (42%).

Model Name	NCT (%)	HMU (%)	Our dataset (%)
EfficientNet-B7	82.47	84.45	**98.80**
SwinTrans.	79.05	91.88	**98.79**
Ctranspath	79.05	91.88	**99.00**
UNI	80.43	91.80	**98.26**
CNN	64.21	68.10	**97.81**

Table 1: Performance of top five diverse model architectures trained on three datasets

- The significant impact of training data quality on model feature activations for multi-class tissue classification is illustrated in Figure 5.

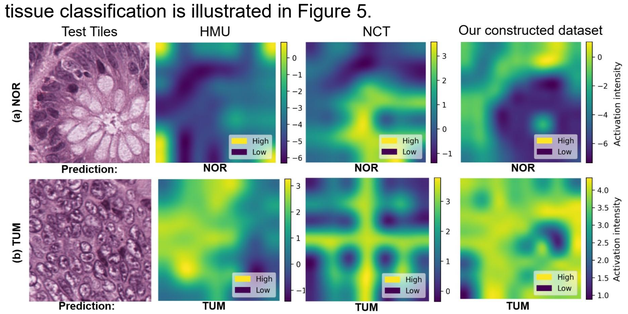


Figure 5: Feature map visualizations for the best models trained on HMU, NCT, and our constructed dataset for (a) normal and (b) tumor tissue type.

Large-scale Dataset Construction Framework

We developed a semi-automated, multi-phase pipeline (Figure 3) to build a large-scale dataset for multi-class tissue classification across nine tissue types (Figure 1).

Phase 1: Autoencoder Training

- Tiles are extracted from whole-slide images (WSI) with advanced pre-processing.
- Trained an AutoEncoder (AE_CRC) on 100,000 CRC tiles for building domain-specific feature extractor.
- The encoder–decoder in AE_CRC is trained with SSIM loss to preserve structural details.

Phase 2: Clustering Tiles (K-means)

- Features from tiles are extracted using AE_CRC encoder and reduced to 256-dim latent vectors using PCA.
- Clustered tiles using K-means (in which $K = \sqrt{n}$), where n is the number of tiles in a WSI for morphology-based grouping.
- Preserved local morphologic coherence across adjacent clusters.

Phase 3: Sampling with Equal Frequency Binning

- Computed and normalized centroid distances for each cluster.
- Applied equal-frequency binning with number of groups (g) to capture diverse tissue patterns.
- Sampled 20% (s) from each group/bin (g), balancing class representation.

Phase 4: Pathologist Verification & Assembly

- Sampled tiles were mapped back to WSIs automatically using a custom interface for easy pathologist verification.
- Verified tissue classification with expert pathologists.
- Final dataset: 630,000 balanced, high-quality tiles.

The parameters s and g can be adjusted based on the number of samples needed to construct the dataset.

For the CRC multi-tissue classification task as a use case, we constructed a dataset of 630,000 tiles uniformly sampled across nine clinically relevant tissue classes (~70,000 tiles each) from 200 H&E-stained WSIs at Stanford University School of Medicine.

Advantages of the proposed framework:

- Eliminates redundant tissue tiles into a concise set.
- Lowers computational complexity.
- Builds high-quality datasets without manual effort.

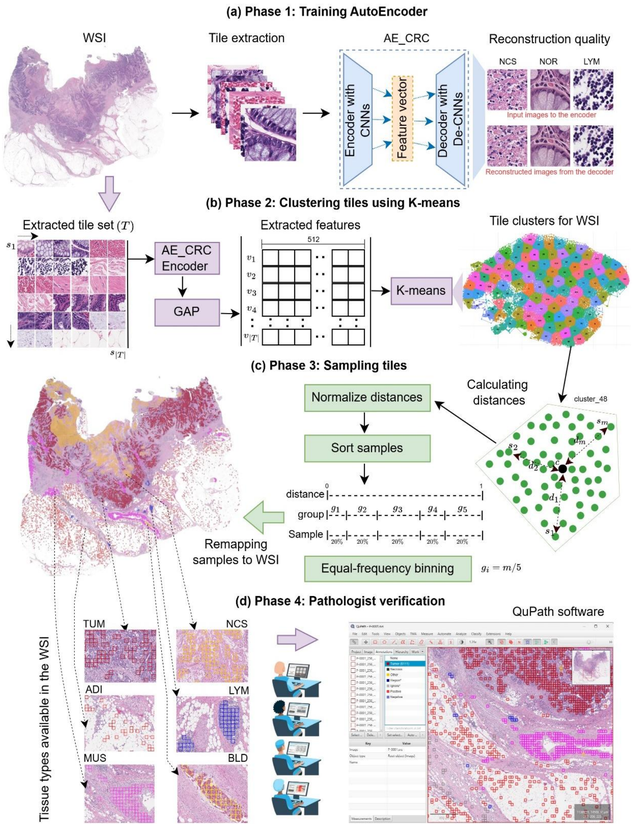


Figure 3: Overall workflow of dataset generation framework.

Conclusions

- We developed a scalable, semi-automated framework to efficiently construct and sufficiently large, balanced, diversified dataset for multi-tissue classification.
- Models trained on our constructed dataset achieved over 98% classification accuracy across various models, demonstrating strong generalization on independent validation sets.

References

1. Q. Ke, W. S. Yap, Y. K. Tee, Y. C. Hum, H. Zheng, and Y. J. Gan, “Advanced deep learning for multi-class colorectal cancer histopathology: integrating transfer learning and ensemble methods,” Quant Imaging Med Surg, vol. 15, no. 3, pp. 2329–2346, Mar. 2025,doi:10.21037/qims-24-1641.Journal Article, Name of Journal
2. J. N. Kather et al., “Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study,” PLoS Med, vol. 16, no. 1, 2019, doi:10.1371/journal.pmed.1002730.
3. S. Lou et al., “A large histological images dataset of gastric cancer with tumour microenvironment annotation for AI,” Sci Data, vol. 138, no. 12, 2025. Available: www.nature.com/scientificdata.
4. D. M. Muzny et al., “Comprehensive molecular characterization of human colon and rectal cancer,” Nature, vol. 487, no. 7407, pp. 330–337, Jul. 2012, doi:10.1038/nature11252.
5. A. Ignatov and G. Malivenko, “NCT-CRC-HE: Not All Histopathological Datasets Are Equally Useful,” ArXiv, Sep. 2024.