ANNOTATEANYCELL: OPEN-SOURCE AI FRAMEWORK FOR EFFICIENT ANNOTATION IN DIGITAL PATHOLOGY

A PREPRINT

Shourya Verma*[1], **Aditya Malusare***[2,5], **Mengbo Wang**[1,5], **Luopin Wang**[1,5], **Arpan Mahapatra**[1]

Abigail English[3], **Abigail Cox**[3], **Meaghan Broman**[3], **Simone De Brot**[6], **Grant N. Burcham**[3]

Deborah Knapp[4,5], **Deepika Dhawan**[4,5], **Mario Sola**[3], **Vaneet Aggarwal**[1,2,5], **Ananth Grama**[1]

Nadia Atallah Lanman[3,5]

* Equal Contribution
Purdue University [1] Computer Science, [2] Industrial Engineering, [3] Comparative Pathobiology,
[4] Veterinary Clinical Sciences; [5] Purdue Institute for Cancer Research; **University of Bern** [6] Veterinary Pathology

January 29, 2026

Key Points: **Question:** Can a semi-supervised deep learning framework reduce annotation burden for pathologist while maintaining accuracy in digital pathology? **Findings:** In this methods' study evaluating eight canine urothelial carcinoma slides, the AnnotateAnyCell framework achieved 96% accuracy for mitotic classification and 98% for nucleoli classification using just 1075 labeled samples, while reducing annotation time by 25% compared to sequential annotation. **Meaning:** This open-source framework enables efficient large-scale cellular annotation in histopathology, potentially accelerating cancer research and clinical application.

ABSTRACT

Importance: Manual annotation of histopathological images by pathologists is effort-intensive and represents a major challenge for computational analysis and clinical AI deployment. **Objective:** To develop and validate an open-source semi-supervised deep learning framework for accurate and efficient annotation and analysis of cellular structures in whole slide images. **Design:** Methods development study using active contrastive learning with iterative human-in-the-loop feedback. The framework integrates cell segmentation, embedding-based visualization, and semi-supervised classification. **Setting:** Digital pathology computational analysis platform with intuitive web-based annotation interface. **Methods:** Five whole slide images of canine invasive urothelial carcinoma digitized at 40× magnification, representing low, intermediate, and high histological grades with diverse morphological patterns. The AnnotateAnyCell framework processing cells through Cellpose segmentation, UMAP dimensionality reduction for interactive annotation, and contrastive learning-based classification with pseudolabel generation. **Main Outcomes and Measures:** Primary outcomes were accurate classification for nuclear morphology features (mitotic figures, nucleoli, chromatin, shape) and annotation efficiency measured as time per cell. Secondary outcomes included inter-annotator agreement and model performance scaling with training data size. **Results:** The framework processed 8 slides containing hundreds of thousands of cells. Latent space clustering-based annotation required 47 minutes versus 63 minutes for sequential annotation (25% reduction, 95% CI 18%-32%). Classification accuracy reached 96.3% ± 1.2% for mitotic figures and 98.3% ± 1.4% for nucleoli with 1075 labeled samples. Nuclear shape classification achieved 59.5% ± 2.1% accuracy. Inter-annotator agreement was highest for chromatin (100%) and nucleoli (95%,

κ = 0.95), moderate for mitotic figures (64%, κ = 0.58), and lowest for shape (81%, κ = 0.36). Performance scaled efficiently, with nucleoli classification reaching 95.5% ± 1.5% accuracy using only 215 samples. **Conclusions and Relevance:** This semi-supervised active learning framework substantially reduces annotation burden while achieving expert-level accuracy for well-defined morphological features. The open-source tool accelerates histopathological dataset curation for cancer research and enables deployment of AI-assisted diagnostics in resource-constrained settings. **Github:** https://github.com/shouryaverma/AnnotateAnyCell

Keywords Digital Pathology, Cell Annotation, Active Learning

1 Introduction

The integration of AI into digital pathology presents opportunities for systematic analysis of histopathological images at unprecedented scale [12, 29]. Modern whole slide images at 40× magnification contain hundreds of thousands of cells, typically taking several gigabytes of storage, and often exceeding the capacity for comprehensive manual review [28, 20]. While deep learning methods have shown promise for automated cellular analysis, their clinical deployment is limited by the requirement for extensive manual annotations by expert pathologists, who must spend hundreds of hours delineating cellular boundaries and assigning morphological classifications. Current approaches primarily use fully supervised learning, which requires extensive labeling of training datasets. This annotation burden is exacerbated by poor model generalization [25] across institutions due to variations in tissue preparation protocols and scanner characteristics, necessitating institution-specific retraining. Semi-supervised learning methods can leverage abundant unlabeled data alongside limited expert annotations. However, existing platforms require substantial computational infrastructure, lack intuitive interfaces for pathologist workflows, and do not provide iterative feedback loops to build clinical trust [1, 9, 19]. We developed AnnotateAnyCell, an open-source semi-supervised framework combining deep learning with an interactive web-based interface designed for pathologist workflows. The system enables annotation within learned embedding spaces, where morphologically similar cells naturally cluster, integrating active learning to intelligently select informative samples for expert review. We evaluated this framework on canine invasive urothelial carcinoma specimens, a translational model for human muscle-invasive bladder cancer [11], to determine annotation efficiency and classification accuracy for diagnostically relevant nuclear morphology features.

2 Background

Cell Annotation. Desktop applications like QuPath [2] and Orbit [23] provide efficient tile-based visualization with manual annotation capabilities, while commercial platforms including HALO (Indica Labs) and Visiopharm dominate clinical deployment with FDA-cleared algorithms, validated workflows, and enterprise support infrastructure. Web-based platforms such as Cytomine [13] and Digital Slide Archive [8] enable collaborative annotation through RESTful APIs and distributed architectures, though they require substantial server infrastructure. Recent AI-assisted tools have dramatically improved annotation efficiency [27]: Quick Annotator [16] demonstrated significant speedups over manual annotation by training U-Net models concurrently during annotation and using UMAP-based patch selection for active learning, while PatchSorter [26] achieved improvements through deep learning-based object labeling [3].

Active Learning. Active learning has emerged as a critical approach to address the substantial annotation burden in digital pathology. Recent work has demonstrated significant efficiency gains through intelligent sample selection strategies. DADA (Diversity-Aware Data Acquisition) combined uncertainty and diversity clustering to achieve comparable AUC with 96% fewer patches for tumor-infiltrating lymphocyte classification [15]. Human-in-the-loop systems like Quick Annotator [16] have reduced annotation time through iterative review-and-revise workflows, where models provide hypothesized annotations that pathologists can rapidly verify. Uncertainty quantification methods, particularly Monte Carlo dropout [7], have enabled high-confidence predictions robust to domain shift. Despite these advances, most active learning systems operate at the patch or region level rather than individual cells, and few leverage modern embedding spaces for direct interactive annotation. AnnotateAnyCell significantly extends these approaches by combining active learning with contrastive learning-based embeddings in UMAP [14] space, enabling pathologists to annotate cells interactively across learned feature representations rather than spatial coordinates alone.

3 Methods

Our framework implements a semi-supervised active learning pipeline (Figure 1) for efficient cellular structure annotation through four iterative stages. First, deep learning models segment H&E-stained tissue images to extract individual cells as multi-modal representations: raw patches, isolated regions, and semantic masks. Second, extracted

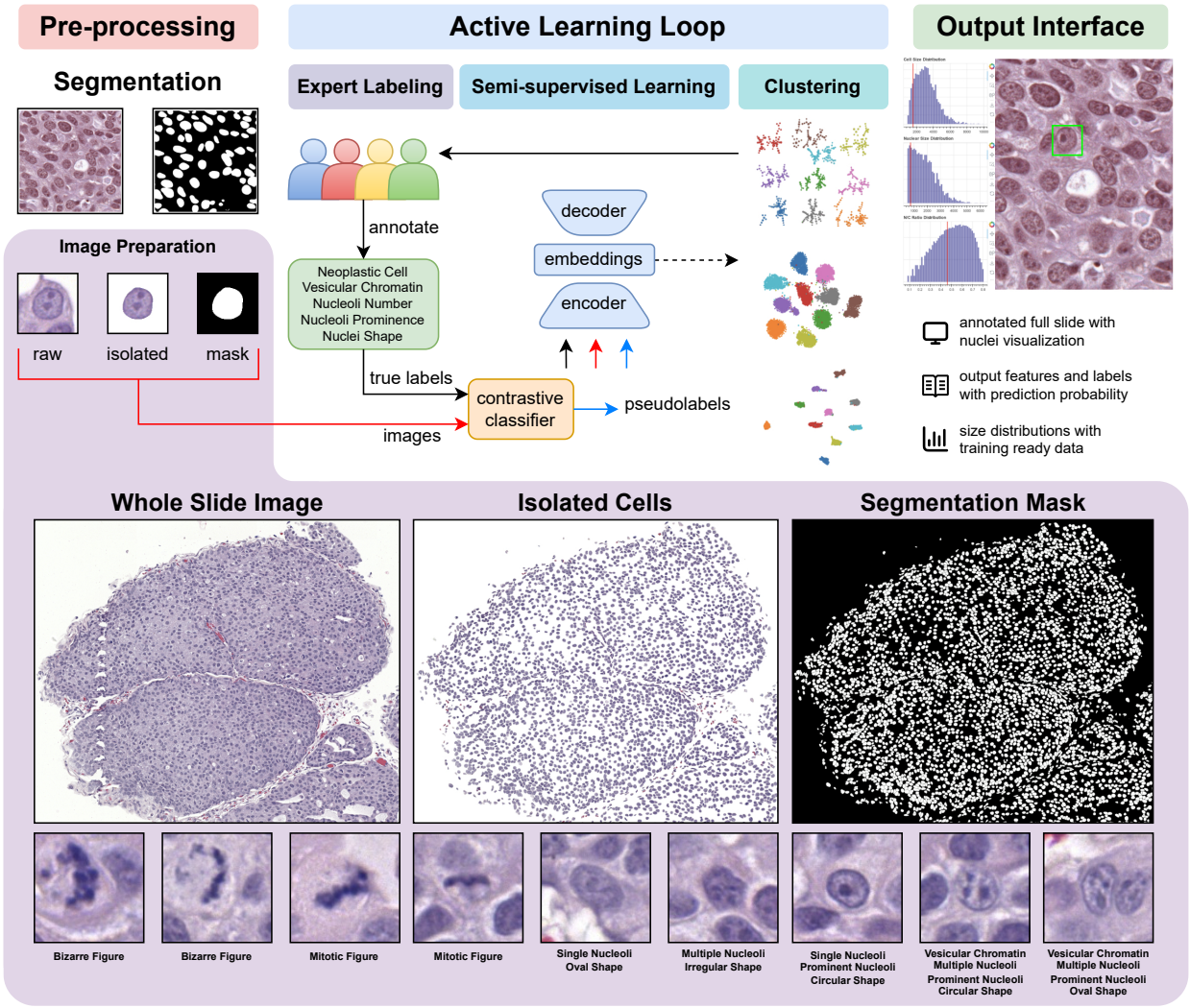

Figure 1: Overview of the "AnnotateAnyCell" framework showing the complete pipeline from image preprocessing to output generation: **Pre-processing stage:** Raw H&E-stained tissue images undergo segmentation using Cellpose models to generate nuclear masks. Each detected nucleus is extracted as three complementary representations: raw image patches, isolated nuclear regions, and semantic masks; **Active Learning Loop:** Expert pathologists (represented by colored figures) provide initial annotations for nuclear morphology features of neoplastic cells, including mitotic figures, vesicular chromatin, and prominent nucleoli. These labeled samples train a contrastive classifier that generates embeddings and pseudolabels for unlabeled data. A convolutional autoencoder processes the multi-modal inputs (raw, isolated, mask) to learn compact latent representations; **Output Interface:** The system produces annotated whole slide images with nuclei visualization, quantitative feature distributions, and prediction probabilities for downstream analysis. The iterative process refines the model's understanding of cellular morphologies through expert feedback. **Image Preparation:** Left: Representative whole-slide image (WSI). Center: Nuclear segmentation results using pretrained Cellpose models. Right: Binary segmentation mask showing all detected valid cells. (Bottom row) Nine examples of 128×128 pixel cellular tiles illustrating morphological diversity across expert-defined annotation categories. Examples include bizarre mitotic figures showing irregular mitotic configurations, canonical mitotic figures demonstrating typical mitotic chromatin condensation patterns, nucleolar variations (single vs. multiple nucleoli) combined with nuclear shape classifications (circular, oval, irregular), and chromatin texture patterns (vesicular chromatin).

cells are presented to pathologists in a learned embedding space where morphologically similar cells cluster for efficient phenotype labeling. Third, a contrastive learning framework integrates expert labels and unlabeled data to generate high-confidence pseudolabels and prioritize informative samples for review. Fourth, a multi-modal autoencoder processes labeled and pseudolabeled data to learn compact latent embeddings capturing fine-grained morphology and higher-level context, producing updated visualizations. This loop iterates until sufficient annotation coverage is achieved, progressively improving model accuracy and expert efficiency.

Dataset and Sample Collection. We evaluated our framework on canine invasive urothelial carcinoma (IncUC), a spontaneous tumor model [11, 6] that recapitulates human muscle-invasive bladder cancer biology for translational research. Tissue samples from client-owned dogs with naturally occurring IncUC were collected at Purdue University Veterinary Teaching Hospital under informed consent and institutional protocols. Formalin-fixed paraffin-embedded blocks were sectioned at 4 μm and H&E-stained using automated protocols (Leica Autostainer XL). Slides were digitized at 40× magnification (0.25 μm/pixel, Aperio AT2), producing whole slide images of 80,000×100,000 pixels. Cases were selected for diagnostic quality (minimal artifacts, adequate fixation, uniform staining) and tumor cellularity (>30% neoplastic nuclei). Eight representative slides from cystoscopy or necropsy samples captured morphological diversity with mitotic activity ranging from 5 to >30 mitoses per 10 high-power fields and papillary or solid architectural patterns. ACVP Board-certified veterinary pathologists defined annotation categories based on diagnostically relevant nuclear features: mitotic figures (canonical and atypical configurations), nucleolar characteristics (absent, single, multiple, prominent), chromatin texture (vesicular vs. hyperchromatic), and nuclear shape (circular, oval, irregular). These features were selected for their prognostic significance in canine and human urothelial carcinoma [17, 18], relevance to grade assignment, and reliable inter-observer agreement.

User Interface. The interactive user interface was built with BokehJS (JavaScript with Python bindings) to enable real-time multi-user annotation. The server maintains persistent sessions with checkpoint functionality for saving progress and reverting states. The primary workspace (Fig. 2) comprises three coordinated panels. The central panel displays a UMAP [14] embedding as an interactive scatter plot where each point represents a cellular tile, color-coded by status (blue: unlabeled, green: labeled, orange: selected). Hovering reveals the original image tile, segmented regions, and WSI coordinates. The left panel provides stepwise protocol guidance with interactive analytics: histograms of nuclear area and nuclear-to-cytoplasmic ratios dynamically update with selections, with red markers highlighting measurements relative to population statistics. The right panel presents 512×512 high-resolution previews with annotation controls for pathologist-defined morphological classes (mitotic figures, chromatin patterns, nucleoli, nuclear shape). Each annotation records confidence scores and timestamps for inter-observer variability analysis. Real-time progress tracking displays dataset coverage (e.g., "300/300 labeled points"). The "Update Clusters" button triggers asynchronous model retraining with live progress display; upon completion, the embedding automatically updates to reveal refined clusters and morphological relationships. The results interface (Fig. 3) reuses the three-panel layout: the left panel summarizes annotation counts per morphology, the central panel displays the WSI with cyan markers for annotated cells supporting pan/zoom exploration, and the right panel enables detailed inspection of individual annotations. Data export via CSV and advanced filters enable hypothesis generation and pattern discovery. Eleven board-certified veterinary pathologists tested the interface by labeling 300 cells each to evaluate usability and inter-annotator agreement.

Image Pre-processing. High-quality input data requires precise preprocessing to segment, extract, and normalize cellular and nuclear features from whole-slide images (WSIs). We use pre-trained Cellpose models [22]: cyto3 [21] for cell and nuclear segmentation. Segmentation masks and boundaries are filtered to remove artifacts, partial objects, and merged cells. Cell and nuclear masks are integrated, enforcing containment (each nucleus within a cell body); cells without detected nuclei are treated entirely as nuclei for consistency. From segmented regions, we extract tile patches: 128×128 pixels for training and 512×512 pixels for expert inspection. Three versions per tile are generated: (i) raw WSI image, (ii) semantic mask, and (iii) isolated cell/nucleus with surrounding tissue removed. All tiles are bilinearly interpolated to standard dimensions. Metadata including nuclear-to-cytoplasmic ratio, nuclear and cellular areas, and centroid coordinates are stored for analysis and annotation context. UMAP [14] is applied to 128×128 patches for dimensionality reduction, enabling visualization and clustering of morphologically similar cells.

Active Learning Loop. To minimize annotation burden, we developed a semi-supervised active learning pipeline iteratively selecting informative samples for expert review. The process begins with UMAP embeddings of 128×128 cellular tiles creating a 2D morphological feature space for pathologists to explore and annotate representative tiles. Annotations are stored in a MongoDB database with secure access, version control, and multi-expert collaboration support. Next, a contrastive learning model with convolutional backbone, projection head, and classification head is trained. The objective combines supervised classification loss on labeled data with unsupervised contrastive loss on all samples to embed similar morphologies together while separating dissimilar ones. The trained model generates class-balanced pseudolabels by selecting top-N highest-confidence samples per class to avoid bias toward dominant morphologies. For subsequent rounds, uncertainty sampling with diversity promotion identifies informative examples. A convolutional autoencoder trained on annotated and pseudolabeled data processes raw images, nuclear features,

Figure 2: Main annotation interface of the "AnnotateAnyCell" system demonstrating the embedding-based annotation workflow. Left panel: Displays workflow guide for pathologists, nuclear size distribution histogram, and nuclear-to-cell ratio distribution with red indicators showing the currently selected cell's measurements. Center panel: Interactive UMAP embedding visualization, where each point represents a cellular tile extracted from the whole slide image. Cells are color-coded based on annotation status (blue: unlabeled, green: labeled, orange: selected) and naturally cluster based on morphological similarities. Right panel: Shows a high-resolution preview of the selected cell tile (Tile 710) with morphological annotation options including mitotic figures, chromatin patterns, nucleoli characteristics, and nuclear shape classifications. The interface includes progress tracking (300/300 labeled points), version control, and model retraining capabilities ("Update Clusters" button).

and semantic masks simultaneously. With skip connections and a variational bottleneck, the autoencoder captures fine-grained morphology and contextual structure in its latent space. UMAP projection of these embeddings yields updated visualizations clustering morphologically similar cells while preserving expert-defined boundaries. Experts use this refined space to validate or correct pseudolabels and annotate uncertain regions. This loop refines the nuclear morphology space, improving classification accuracy, annotation efficiency (performance gain per label), and novel phenotype discovery. The following subsections detail key pipeline components.

Pseudolabel Generation. Our semi-supervised learning framework integrates supervised and unsupervised techniques through three components: a convolutional classifier for feature extraction, a contrastive learning module for representation enhancement, and a selective pseudolabeling strategy for label propagation. The base CNN extracts features using four convolutional blocks (Conv2D with 3×3 kernels, ReLU, max-pooling), increasing channel depth (16→128) while reducing spatial dimensions, followed by four fully connected layers (1024→128 neurons) with batch normalization and 50% dropout. The final layer maps to expert-defined nuclear morphology classes. A projection head maps CNN features to a 128-dimensional normalized embedding space for contrastive learning. The supervised classification objective combines with InfoNCE loss [4]:

$$L_{\text{InfoNCE}} = -\log \frac{\exp(sim(q, k_+)/\tau)}{\sum_{i=0}^{K} \exp(sim(q, k_i)/\tau)} \tag{1}$$

where q is the query embedding, k_+ the positive key, k_i negative keys, $sim(u, v)$ cosine similarity, τ temperature, and K number of negatives. The total objective is:

$$L_{\text{total}} = \alpha L_{CE} + \beta L_{\text{InfoNCE}} \tag{2}$$

with α and β controlling loss balance. For pseudolabeling, the trained model uses class-balanced selection, choosing top-N most confident predictions per class (e.g., $N = 20$) to ensure balanced representation and mitigate bias toward

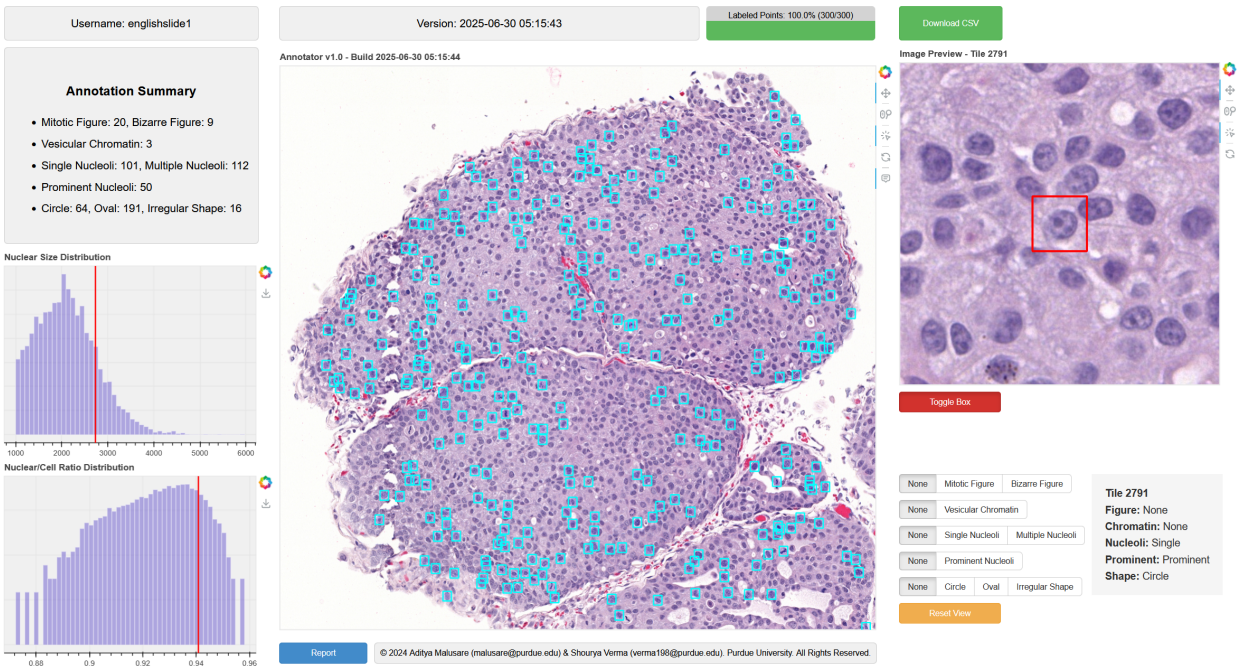

Figure 3: Results and output interface showing the final annotated dataset and analysis capabilities. Left panel: Displays annotation summary statistics including counts for different morphological features along with distribution histograms. Center panel: Whole slide image view with cyan markers indicating the spatial locations of all annotated cells across the tissue sample, enabling pathologists to assess spatial distribution patterns and tissue architecture context. Right panel: Individual cell inspection tool showing detailed annotations for the selected cell (Tile 2791) with its specific morphological classifications. The interface includes data export functionality (Download CSV), version information, and the ability to filter and visualize cells based on specific morphological criteria. This output interface enables analysis of annotation results and supports research workflows requiring quantitative cellular phenotype data.

dominant classes. Unselected samples retain null labels and contribute only to unsupervised learning. Experimental validation [24] shows this approach outperforms fully supervised baselines and naive pseudolabeling.

Autoencoder Model. The autoencoder enables unsupervised feature learning, dimensionality reduction, and visualization by processing preprocessed tiles to learn compact representations integrating expert labels with intrinsic data structure. We implement a multi-modal convolutional autoencoder with three input channels: raw cellular images, nuclear regions, and semantic masks. Each modality passes through parallel encoders (four Conv2D layers: $3\times16 \rightarrow 16\times32 \rightarrow 32\times64 \rightarrow 64\times128$, ReLU). Skip connections preserve spatial detail. Encoded features are flattened, passed through fully connected layers (1024→512, batch normalization), and concatenated with one-hot pseudolabels. The integrated embedding is processed through additional layers (1024→512). Following the VAE framework [10], mean and logvar projections map to a 256-dimensional latent space with reparameterization trick applied during training. Decoders mirror encoder structure with transposed convolutions. The training objective combines reconstruction and feature prediction losses:

$$L_{\text{recon}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x_i})^2 \tag{3}$$

where x_i is input and $\hat{x_i}$ reconstruction, and feature prediction uses cross-entropy:

$$L_{\text{feat}} = -\frac{1}{m}\sum_{j=1}^{m}\sum_{c=1}^{C} y_{j,c}\log(\hat{y}_{j,c}) \tag{4}$$

The total loss is:

$$L_{\text{total}} = \alpha L_{\text{recon}} + \beta L_{\text{feat}} \tag{5}$$

with α, β controlling trade-offs. The encoder transforms $128\times128\times9$ input tensors into 256-dimensional embeddings capturing supervised annotation signals and unsupervised data structure, projected to 2D via UMAP [14] for interpretable visualization. This multi-modal variational autoencoder improves representation quality, reconstruction fidelity, and

clustering performance, supporting efficient annotation through modality integration, dimensionality reduction, and continuous visualization updates.

4 Results

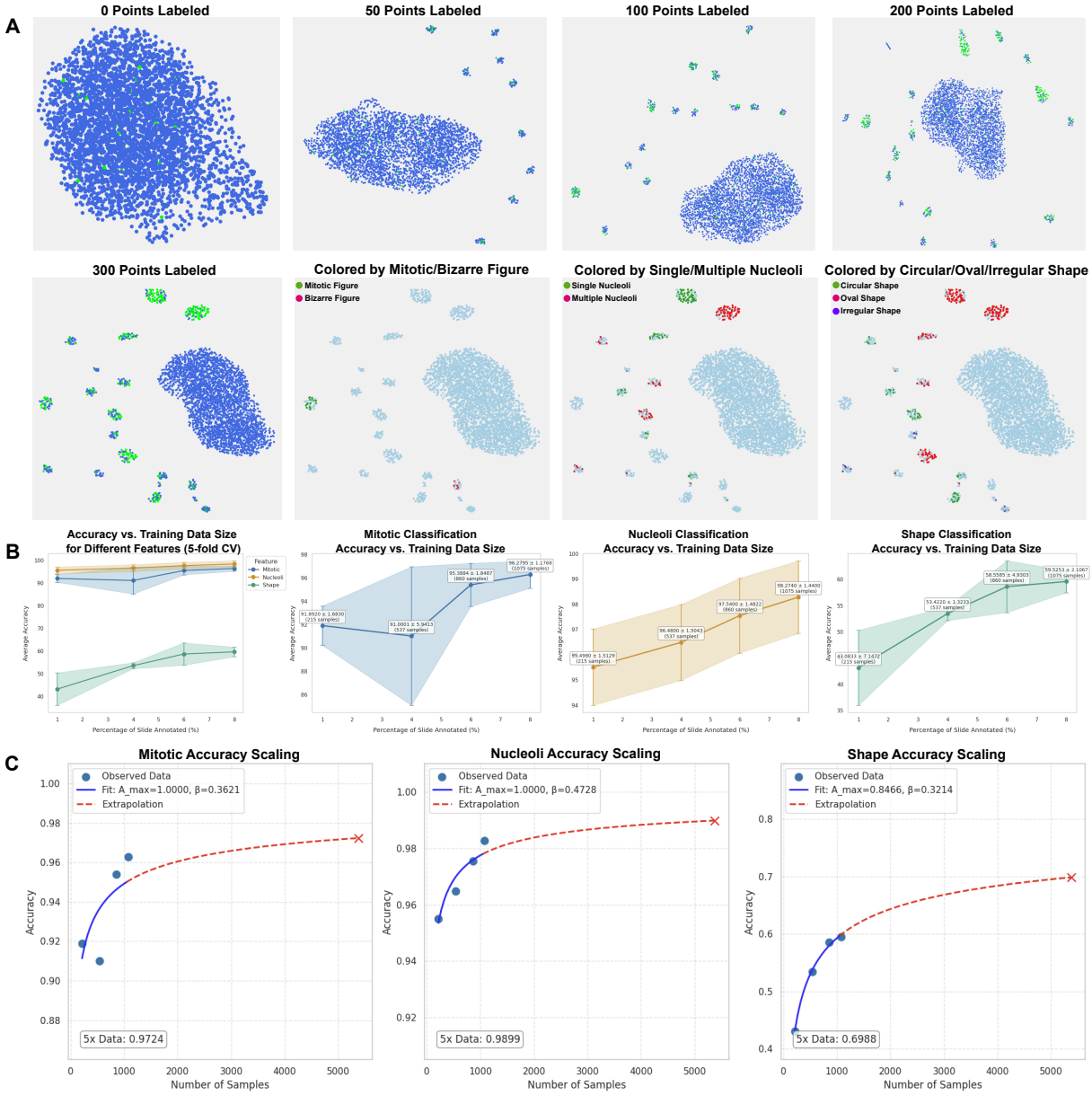

Figure 4: **A.** Illustrates sequential snapshots of the learned embedding space at 0, 50, 100, 200, and 300 labeled cells, showing the progression of expert annotations (green) within the unlabeled cellular population (blue). **B.** Reports classification evaluation of mitotic figures, nucleoli, and nuclear shape across four training sizes (1%, 4%, 6%, 8% of slide coverage; 215, 537, 860, 1075 labeled samples) using 5-fold cross-validation. **C.** Illustrates how scaling behaviors highlight fundamental differences in data efficiency, suggesting that clinically acceptable performance (>90%) can be achieved with only 200-500 samples for nucleoli, 800-1000 samples for mitotic figures, and >5000 samples for shape.

Annotator Cohort. The annotator cohort consisted of ACVP (American College of Veterinary Pathologists) and FMH (Foederatio Medicorum Helveticorum) board certified pathologists with more than 10 years of experience each. Annotations were also provided by the pathology residents in training, which were verified by the certified pathologists.

Model Performance. In Figure 4 **A.** unsupervised UMAP projection of 128×128 pixel cellular tiles creates a 2D morphological landscape. At initialization (0 points), all cells are unlabeled. With 50 labeled points, annotations scatter sparsely as pathologists explore diversity and establish class boundaries. At 100-200 points, annotation density increases in coherent neighborhoods via uncertainty sampling with diversity promotion. At 3000 points, labeled cells (green) populate discrete clusters and boundaries, with a large central cluster of unlabeled cells not yet confidently assigned and peripheral clusters capturing annotated dominant and rare phenotypes. The framework aims to break the large unlabeled cluster into phenotypically necessary sub-clusters. Bottom row shows feature-specific visualization of the 300-point embedding colored by morphological classifications, revealing feature-dependent clustering quality.

Figure 4 **B.** shows nucleoli classification achieved highest accuracy (98.3% at 1075 samples), mitotic reached 96.3%, and shape plateaued at 59.5%. Mitotic classification showed strong baseline (91.9% $\pm$ 1.7% at 215 samples), transient dip at 4% coverage (91.0% $\pm$ 5.9%), recovery to 95.4% $\pm$ 1.8% at 6%, and plateau at 96.3% $\pm$ 1.2% with 1075 samples. Mid-range instability likely reflects class imbalance or sampling variability. Above 800 samples, performance stabilized with narrow bounds, exceeding the 64% inter-annotator agreement. Nucleoli classification started at 95.5% $\pm$ 1.5% with 215 samples and improved monotonically to 98.3% $\pm$ 1.4%. Consistently narrow confidence intervals mirror the 95% inter-annotator agreement, with final accuracy approaching the theoretical ceiling. Shape classification proved most difficult, rising from 43.1% $\pm$ 7.1% to 59.5% $\pm$ 2.1% at full data. Despite the largest absolute gain (16.4 points), performance remained lowest due to subjectivity of categorical shape labels and dependence on continuous morphological spectra with complex contextual factors. Figure 4 **C.** Shows power-law fits highlight distinct scaling regimes: nucleoli and mitotic figures exhibit early saturation with diminishing returns, whereas shape maintains a steep learning trajectory, suggesting substantially more labeled data would be required to approach parity.

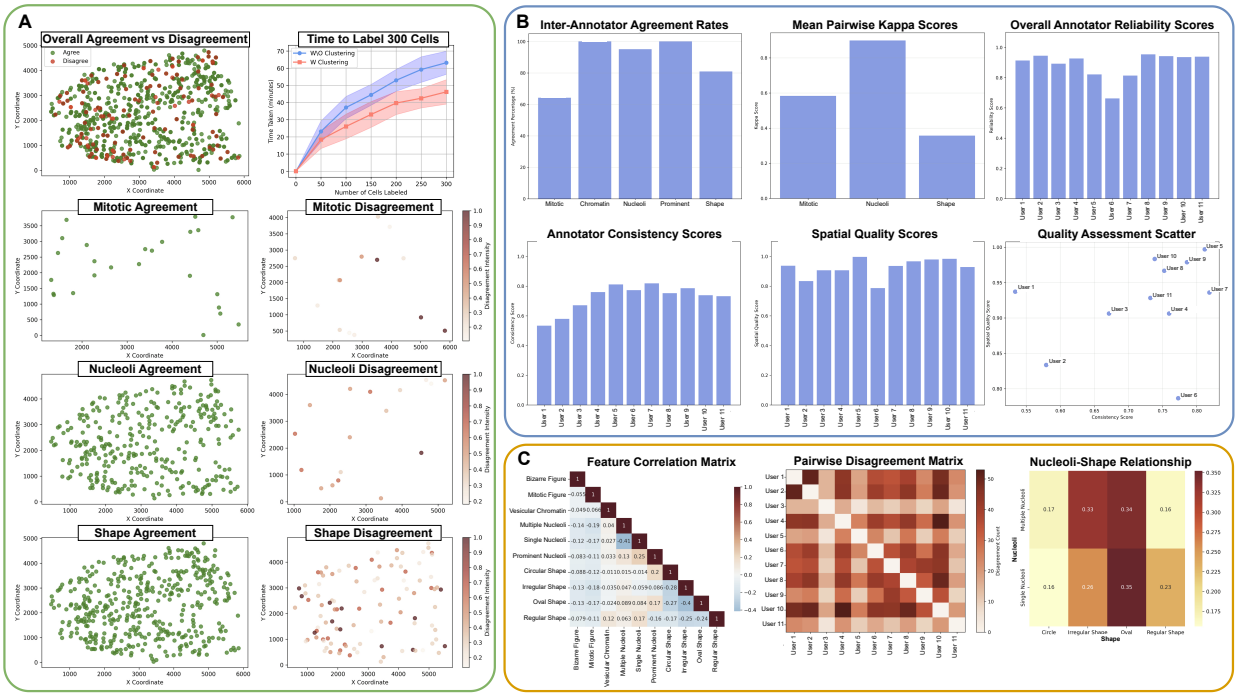

Figure 5: **A.** Illustrates labelling time with and without clustering, along with spatial distributions of agreement (green) and disagreement (brown) reveal localized cell-level ambiguities, with disagreement intensity encoded by saturation. **B.** Presents the annotation workflow metrics like agreement rates, reliability scores, consistency and spatial coverage scores across 11 annotators. **C.** Illustrates feature correlation demonstrating nuclear features are largely independent dimensions of morphology, and pairwise disagreement analyses showing some users might disagree more than others.

User Performance. Figure 5 summarizes spatial agreement maps, temporal efficiency, feature-level metrics, annotator reliability, inter-feature correlations, and disagreement structures. In Figure 5 **A.** clustering-based active learning substantially outperformed sequential annotation, requiring 47 vs. 63 minutes for 300 cells (25% reduction; 16 minutes absolute; 9.4 vs. 12.6 sec/cell) averaged over 11 users. Improvements were consistent across volumes: 22% at 50 cells (18 vs. 23 min), 30% at 100 cells (26 vs. 37 min), and 25% at 150 cells (33 vs. 44 min), all scaling linearly without fatigue effects. Clustering showed narrower confidence intervals, reflecting consistent performance. Efficiency gains stem from embedding-based grouping of 5-10 morphologically similar cells, reducing cognitive load and context

switching. Interactive UMAP visualization, real-time statistics, and uncertainty- and diversity-based sampling ensure informative morphologies are prioritized, suggesting the 25% time savings underestimates true efficiency when accounting for information content per annotation hour. Spatial annotation patterns revealed feature-specific reliability. Mitotic figures were sparse (38 sites: 25 concordant, 13 disagreements at 0.2-1.0 intensity), reflecting difficulty distinguishing mitotic events from apoptotic bodies, with sparse distribution suggesting false negatives. Nucleoli achieved highest density with uniform coverage and minimal disagreements (16 sites), mostly borderline single vs. multiple classifications, confirming high reliability. Nuclear shape showed highest absolute disagreement across the full intensity range, with moderate disagreements (0.4-0.6) clustered regionally and sporadic high-intensity disagreements (≥ 0.8) reflecting divergent classification strategies. Disagreements showed no regional clustering across features, indicating cell-level ambiguity rather than slide-level artifacts.

Figure 5 **B.** shows agreement rates varied substantially. Chromatin texture and prominent nucleoli achieved perfect concordance (100%). Nucleoli classification showed high agreement (95%, Cohen's $\kappa = 0.95$) [5], shape was variable (81%, $\kappa = 0.36$), and mitotic activity most challenging (64%, $\kappa = 0.58$). Reliability across 11 annotators ranged from 0.65-0.95: Users 2, 8, 9, 10 achieved ≥ 0.95, most clustered at 0.80-0.95, and User 6 lowest (0.65). Decision consistency ranged from 0.52-0.82 (User 5: 0.82; Users 1, 2: 0.52-0.57). Spatial quality scores were 0.80-1.00 (User 5: 1.00; User 6: 0.80). High performers (Users 5, 9, 10) combined strong consistency (0.75-0.82) with near-perfect spatial coverage (0.95-1.00), while Users 2 and 6 emerged as retraining candidates. Chromatin and nucleoli are robust for clinical and computational pipelines, while mitotic activity and shape require standardized protocols or noise-robust learning.

Figure 5 **C.** shows pairwise correlations were weak ($|r| < 0.3$), with strongest between single and prominent nucleoli ($r = 0.25$). Oval shape correlated weakly with single ($r = 0.084$) and prominent nucleoli ($r = 0.17$). Irregular shape correlated negatively with circular ($r = -0.27$) and oval ($r = -0.40$). Multiple and single nucleoli were mutually exclusive ($r = -0.41$). Mitotic figures ($r < 0.20$) and vesicular chromatin ($r < 0.10$) showed minimal associations. Cross-tab analysis revealed biologically meaningful associations: multiple nucleoli enriched in oval (0.34) and irregular (0.33) nuclei; single nucleoli also frequent in oval (0.35) and irregular (0.26). Both nucleolar states elevated in irregular nuclei, suggesting nuclear irregularity as an independent atypia marker. These support a multidimensional diagnostic model. Inter-annotator disagreement revealed systematic variability: Users 1, 4, 10 showed highest discord (>40 with multiple annotators); Users 3, 9 showed strong concordance; User 8 maintained low-to-moderate disagreement. Asymmetric patterns (e.g., User 1 vs. 4, User 1 vs. 10) revealed two annotation "schools": concordant (Users 2, 3, 6, 9) and discordant (Users 1, 4, 10). Results emphasize targeted calibration, consensus reviews, and benchmark evaluations to harmonize practices. For machine learning, features with near-perfect agreement (chromatin, nucleoli) provide robust ground-truth, while ambiguous features (mitotic figures, shape) require consensus labeling or noise-tolerant algorithms. Incorporating disagreement intensity as spatial uncertainty enables principled annotation weighting, preventing model penalization for inherently ambiguous cases.

5 Discussion

We presented a semi-supervised active learning framework for efficient nuclear morphology annotation in histopathological images, integrating contrastive learning, class-balanced pseudolabeling, and multi-modal variational autoencoders. Applied to canine invasive urothelial carcinoma, the framework achieved 25% reduction in annotation time (47 vs. 63 minutes for 300 cells) through clustering-based sample selection and interactive UMAP visualization. Classification performance reached 98.3% for nucleoli, 96.3% for mitotic figures, and 59.5% for nuclear shape, with nucleoli achieving 95.5% accuracy using only 215 labeled samples. The InfoNCE loss successfully leverages unlabeled data to establish morphological feature spaces where similar cells cluster, enabling robust classification from sparse annotations. The class-balanced pseudolabel generation proved effective, with performance improving without biased pseudolabel degradation. The multi-modal autoencoder processing raw images, nuclear features, and semantic masks through a 128-dimensional normalized embedding space captures diverse morphological characteristics. Integration of labeled and pseudolabeled data through the variational bottleneck creates smooth latent representations supporting both supervised classification and unsupervised structure learning. The iterative active learning loop demonstrates steep learning curves for mitotic and shape features, with updated embeddings revealing refined morphological relationships guiding subsequent annotation rounds.

Inter-annotator analysis revealed feature-dependent reliability: chromatin texture and prominent nucleoli achieved perfect agreement (100%), while mitotic activity (64%, $\kappa = 0.58$) and nuclear shape (81%, $\kappa = 0.36$) showed higher variability. Weak pairwise feature correlations ($|r| < 0.3$) support a multidimensional diagnostic model where features contribute complementary information. For clinical deployment, differential performance suggests tiered confidence thresholds: nucleoli and mitotic classifications (96-98% accuracy) achieve reliability for automated screening or diagnostic support, while shape classifications (59% accuracy) require human verification or uncertainty quantification.

Clinical workflows could automatically accept high-confidence predictions while flagging ambiguous cases for expert review. Data efficiency demonstrated by achieving 95.5% accuracy with 215 samples supports rapid institutional adaptation. Pathologists could annotate 200-500 representative local cells to generate institution-specific models capturing local preparation protocols and scanner characteristics while leveraging general morphological knowledge from unsupervised pretraining. This democratizes computational pathology for resource-constrained settings without prohibitive annotation infrastructure. Future work should address label noise tolerance for ambiguous morphologies, extend the framework to multi-tissue generalization, and validate clinical utility in prospective diagnostic workflows. This human-in-the-loop approach demonstrates that combining expert knowledge with semi-supervised learning achieves expert-level performance while substantially reducing annotation burden.

References

[1] Esther Abels, Liron Pantanowitz, Famke Aeffner, Mark D Zarella, Jeroen Van der Laak, Marilyn M Bui, Venkata NP Vemuri, Anil V Parwani, Jeff Gibbs, Emmanuel Agosto-Arroyo, et al. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the digital pathology association. *The Journal of pathology*, 249(3):286–294, 2019.

[2] Peter Bankhead, Maurice B. Loughrey, José A. Fernández, Yvonne Dombrowski, Darragh G. McArt, Philip D. Dunne, Stephen McQuaid, Ronan T. Gray, Liam J. Murray, Helen G. Coleman, Jacqueline A. James, Manuel Salto-Tellez, and Peter W. Hamilton. Qupath: Open source software for digital pathology image analysis. *Scientific Reports*, 7(1), December 2017.

[3] Samuel Budd, Emma C Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical image analysis*, 71:102062, 2021.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[5] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[6] Deepika Dhawan, Noah M Hahn, Jose A Ramos-Vara, and Deborah W Knapp. Naturally-occurring canine invasive urothelial carcinoma harbors luminal and basal transcriptional subtypes found in human muscle invasive bladder cancer. *PLoS genetics*, 14(8):e1007571, 2018.

[7] James M. Dolezal, Andrew Srisuwananukorn, Dmitry Karpeyev, Siddhi Ramesh, Sara Kochanny, Brittany Cody, Aaron S. Mansfield, Sagar Rakshit, Radhika Bansal, Melanie C. Bois, Aaron O. Bungum, Jefree J. Schulte, Everett E. Vokes, Marina Chiara Garassino, Aliya N. Husain, and Alexander T. Pearson. Uncertainty-informed deep learning models enable high-confidence predictions for digital histopathology. *Nature Communications*, 13(1), November 2022.

[8] David A Gutman, Mohammed Khalilia, Sanghoon Lee, Michael Nalisnik, Zach Mullen, Jonathan Beezley, Deepak R Chittajallu, David Manthey, and Lee AD Cooper. The digital slide archive: a software platform for management, integration, and analysis of histology for cancer research. *Cancer research*, 77(21):e75–e78, 2017.

[9] Mohammad Hossein Jarrahi, Vahid Davoudi, and Mohammad Haeri. The key to an effective ai-powered digital pathology: Establishing a symbiotic workflow between pathologists and machine. *Journal of pathology informatics*, 13:100156, 2022.

[10] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[11] Deborah W Knapp, Deepika Dhawan, Jose A Ramos-Vara, Timothy L Ratliff, Gregory M Cresswell, Sagar Utturkar, Breann C Sommer, Christopher M Fulkerson, and Noah M Hahn. Naturally-occurring invasive urothelial carcinoma in dogs, a unique model to drive advances in managing muscle invasive bladder cancer in humans. *Frontiers in Oncology*, 9:1493, 2020.

[12] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[13] Raphaël Marée, Loïc Rollus, Benjamin Stévens, Renaud Hoyoux, Gilles Louppe, Rémy Vandaele, Jean-Michel Begon, Philipp Kainz, Pierre Geurts, and Louis Wehenkel. Collaborative analysis of multi-gigapixel imaging data using cytomine. *Bioinformatics*, 32(9):1395–1401, January 2016.

[14] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[15] André L S Meirelles, Tahsin Kurc, Jun Kong, Renato Ferreira, Joel Saltz, and George Teodoro. Effective and efficient active learning for deep learning-based tissue image analysis. *Bioinformatics*, 39(4), March 2023.

[16] Runtian Miao, Robert Toth, Yu Zhou, Anant Madabhushi, and Andrew Janowczyk. Quick annotator: an open-source digital pathology based rapid image annotation tool. *The Journal of Pathology: Clinical Research*, 7(6):542–547, 2021.

[17] Ji Hoon Oh and Je-Yoel Cho. Comparative oncology: overcoming human cancer through companion animal studies. *Experimental & Molecular Medicine*, 55(4):725–734, 2023.

[18] Melissa Paoloni and Chand Khanna. Translation of new cancer treatments from pet dogs to humans. *Nature Reviews Cancer*, 8(2):147–156, 2008.

[19] Tilman T Rau, William Cross, Ricardo R Lastra, Regina C-L Lo, Andres Matoso, and C Simon Herrington. Closing the loop–the role of pathologists in digital and computational pathology research, 2024.

[20] Paola Chiara Rizzo, Ilaria Girolami, Stefano Marletta, Liron Pantanowitz, Pietro Antonini, Matteo Brunelli, Nicola Santonicco, Paola Vacca, Nicola Tumino, Lorenzo Moretta, et al. Technical and diagnostic issues in whole slide imaging published validation studies. *Frontiers in oncology*, 12:918580, 2022.

[21] Carsen Stringer and Marius Pachitariu. Cellpose3: one-click image restoration for improved cellular segmentation. *Nature methods*, 22(3):592–599, 2025.

[22] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature methods*, 18(1):100–106, 2021.

[23] Manuel Stritt, Anna K. Stalder, and Enrico Vezzali. Orbit image analysis: An open-source whole slide image analysis tool. *PLOS Computational Biology*, 16(2):e1007313, February 2020.

[24] Hang Su, Zhaozheng Yin, Seungil Huh, Takeo Kanade, and Jun Zhu. Interactive cell segmentation based on active and semi-supervised learning. *IEEE transactions on medical imaging*, 35(3):762–777, 2015.

[25] David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen Van Der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical image analysis*, 58:101544, 2019.

[26] Cédric Walker, Tasneem Talawalla, Robert Toth, Akhil Ambekar, Kien Rea, Oswin Chamian, Fan Fan, Sabina Berezowska, Sven Rottenberg, Anant Madabhushi, Marie Maillard, Laura Barisoni, Hugo Mark Horlings, and Andrew Janowczyk. Patchsorter: a high throughput deep learning digital pathology tool for object labeling. *npj Digital Medicine*, 7(1), June 2024.

[27] Haoran Wang, Qiuye Jin, Shiman Li, Siyu Liu, Manning Wang, and Zhijian Song. A comprehensive survey on deep active learning in medical image analysis. *Medical Image Analysis*, 95:103201, 2024.

[28] Mark D Zarella, Douglas Bowman, Famke Aeffner, Navid Farahani, Albert Xthona, Syeda Fatima Absar, Anil Parwani, Marilyn Bui, and Douglas J Hartman. A practical guide to whole slide imaging: a white paper from the digital pathology association. *Archives of pathology & laboratory medicine*, 143(2):222–234, 2019.

[29] S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5):820–838, 2021.