GMR
The Original

INSTANCE-LEVEL STACKING OF DEEPLABV3+ AND SEGFORMER ENSEMBLES WITH EXPLAINABLE FEATURE FUSION FOR NUCLEUS PHENOTYPE CLASSIFICATION IN IMMUNOHISTOCHEMISTRY IMAGES OF LYMPHOMA

Piyush Kumar Gupta[1*], Mumtaz Ahmed[2]

[1]Department of Computer Engineering, Faculty of Engineering and Technology, Jamia Millia Islamia, New Delhi, INDIA
Email: piyushbhaartiya@gmail.com ORCID ID: 0000-0001-7740-6605
[2]Department of Computer Engineering, Faculty of Engineering and Technology, Jamia Millia Islamia, New Delhi, INDIA Email: mahmed1@jmi.ac.in
ORCID ID: 0000-0001-6241-4733

ABSTRACT
Accurate nucleus phenotype classification in immunohistochemistry (IHC) images is essential for biomarker quantification and characterization of the tumor microenvironment, yet automated analysis remains difficult because staining intensity, nuclear morphology, overlapping structures, and tissue heterogeneity vary considerably across specimens. To address these challenges, this study presents an explainable instance-level stacking framework that combines convolutional and transformer-based segmentation models with feature-level meta-learning. DeepLabV3+ with a ResNet-50 encoder and SegFormer-B2 were selected because they capture complementary aspects of tissue organization. Both architectures were trained using five-fold cross-validation, and fold predictions were averaged to obtain ensemble probability maps. Instead of performing direct pixel-level fusion, probability information was aggregated for each annotated nucleus using instance masks. This enabled the extraction of morphological descriptors, probability statistics, and features describing agreement and disagreement between the two models. These complementary descriptors were integrated through an XGBoost stacking classifier. Experiments on the LyNSeC dataset containing 86,997 annotated nuclei showed that the proposed framework outperformed individual models and simpler fusion strategies. The centroid-independent model achieved 95.94% accuracy and a ROC-AUC of 99.24%. SHAP analysis indicated that SegFormer-derived probabilities and inter-model agreement features contributed most strongly to classification, while ablation experiments confirmed the benefit of combining multiple feature categories.

KEYWORDS: Digital Pathology; Immunohistochemistry; Nucleus Phenotype Classification; DeepLabV3+; SegFormer; Ensemble Learning; Instance-Level Stacking; XGBoost; Explainable Artificial Intelligence; SHAP; Computational Pathology

1. INTRODUCTION
Histopathological examination remains central to cancer diagnosis and biomarker assessment, while immunohistochemistry (IHC) provides additional information by revealing protein expression patterns through antigen–antibody interactions [1]–[3]. At the cellular level, identifying positive and negative nuclei directly influences biomarker quantification and characterization of the tumor microenvironment [4]. In routine practice, pathologists often evaluate thousands of nuclei within a single whole-slide image, making the process labor-intensive and susceptible to variability arising from staining intensity, overlapping nuclei, tissue artifacts, and subjective interpretation [5], [6]. Deep learning has transformed computational pathology. Convolutional architectures such as U-Net and DeepLabV3+ are effective at capturing local texture and boundary information, whereas transformer-based models exploit long-range spatial relationships through self-attention [7]. SegFormer has emerged as an attractive alternative because its hierarchical design provides broader contextual awareness without excessive computational complexity. Practical experience indicates that CNNs and transformers often fail in different regions and therefore provide complementary rather than competing representations.
Most existing approaches rely either on single segmentation architecture or on direct pixel-level fusion strategies [8], [9]. Such approaches overlook information contained in the agreement and disagreement between heterogeneous models. Interpretability also remains a concern, since predictive performance alone is insufficient in pathology if the factors driving decisions cannot be related to biological characteristics [10]. To address these limitations, this study proposes an explainable instance-level stacking framework for nucleus phenotype classification. DeepLabV3+ and SegFormer-B2 are trained using five-fold cross-validation and their ensemble probability maps are aggregated at the nucleus level using instance masks. Morphological descriptors, probability statistics, and agreement-based features are subsequently integrated through an XGBoost meta-learner. SHAP analysis and ablation experiments are used to investigate feature contributions, while stratified group cross-validation prevents image-level data leakage and provides realistic estimates of model performance.

This study introduces an explainable instance-level stacking framework that combines convolutional and transformer segmentation ensembles through nucleus-level probability aggregation and XGBoost-based feature fusion. Morphological descriptors, probability statistics, and agreement-based features are jointly analysed using SHAP under a leakage-resistant stratified group cross-validation protocol.

2. RELATED WORK

2.1 Deep Learning and Transformer-Based Histopathology

Early computational pathology systems relied heavily on handcrafted descriptors describing color, texture, and morphology. The emergence of convolutional neural networks shifted the field toward hierarchical feature learning and enabled end-to-end analysis of histopathological images [11], [12]. Architectures such as U-Net and DeepLabV3+ became widely adopted because they preserve spatial information while incorporating multi-scale contextual representations [13], [14]. Benchmark frameworks such as nnU-Net further demonstrated that carefully configured segmentation pipelines can generalize effectively across diverse biomedical datasets [15].

Transformer-based models were introduced to address limitations associated with local receptive fields. Vision Transformer, Swin Transformer, and SegFormer exploit self-attention mechanisms to model long-range spatial dependencies and broader tissue context [16]–[18]. Hybrid architectures, including Medical Transformer and TransUNet, further demonstrated that convolutional and transformer representations can complement one another in medical image segmentation [19], [20]. More recently, foundation models such as SAM, MedSAM, UNI, CONCH, and other large-scale pathology models have shown promising transferability across different computational pathology tasks [21]–[28]. Despite these advances, their application to nucleus phenotype classification remains relatively limited.

2.2 Nucleus Segmentation and Ensemble Learning

Several methods have been proposed for nucleus segmentation and classification. HoVer-Net, StarDist, and Cellpose demonstrated strong performance in separating touching nuclei and identifying cellular phenotypes [29]–[31]. Public datasets such as Lizard, NuInsSeg, PanNuke, and CoNIC have facilitated comparative evaluation of nucleus analysis algorithms across multiple tissue types [32]–[35]. Studies focusing on lymphoma histopathology have also highlighted the value of specialized architectures for segmentation and phenotype characterization [36], [37]. In addition, dense interaction modelling has emphasized the importance of contextual relationships among neighbouring nuclei [38].

Ensemble learning has become a common strategy for reducing prediction variance and improving stability. Most existing approaches, however, employ either a single architecture or combine predictions through pixel-level fusion. Such strategies do not explicitly exploit the information contained in the agreement and disagreement between heterogeneous models, particularly when convolutional and transformer architectures capture different characteristics of tissue organization and staining patterns.

2.3 Explainability and Research Gap

Interpretability has become increasingly important as machine learning systems move closer to clinical deployment. SHAP provides feature-level explanations based on cooperative game theory and has become one of the most widely adopted approaches for interpreting complex prediction models [39]. Nevertheless, explainability in nucleus phenotype classification is often treated as a post hoc analysis rather than being integrated into the feature fusion process itself.

Several limitations therefore remain. CNN and transformer architectures are typically evaluated independently, disagreement between models is rarely considered as an informative feature, and most ensemble frameworks rely on pixel-level fusion. These observations motivated the development of an explainable instance-level stacking framework that integrates morphological descriptors, probability statistics, and agreement-based features through XGBoost-based meta-learning.

3. MATERIALS AND DATASET DESCRIPTION

3.1 LyNSeC Dataset

This study was conducted using the publicly available Lymphocyte Nucleus Segmentation and Classification (LyNSeC) dataset [42], which was developed to support nucleus segmentation and phenotype classification in immunohistochemistry images. Among its three subsets, only LyNSeC-1 was used because it provides explicit phenotype annotations required for supervised classification. The subset contains 379 annotated image tiles extracted from IHC-stained tissue specimens. Each image is represented as a five-channel array of size $512 \times 512 \times 5$. The first three channels correspond to RGB information, whereas the remaining channels contain annotation maps. Channel four stores the nucleus instance map, in which each nucleus is assigned a unique identifier, while the fifth channel contains phenotype labels encoded as background (0), negative nucleus (1), and positive nucleus (2). This annotation structure enables simultaneous analysis of segmentation quality and phenotype classification at the instance level.

3.2 Dataset Characteristics and Verification

Following dataset inspection and cleaning, a total of 86,997 annotated nuclei were identified across the 379 image tiles, including 47,966 negative nuclei and 39,031 positive nuclei. The class distribution is relatively balanced and therefore does not require aggressive resampling strategies. Morphological analysis revealed substantial variability in nuclear size, with a mean area of 354.48 pixels and a median area of 325 pixels. The smallest annotated nucleus occupied 11 pixels, whereas the largest covered 2,478 pixels, illustrating the heterogeneity typically encountered in histopathological specimens.

During data verification, five images (84_l1.npy, 91_l1.npy, 146_l1.npy, 190_l1.npy, and 215_l1.npy) were found to contain an unexpected phenotype value of 126, which is not part of the official annotation schema. Since valid labels are restricted to {0,1,2}, these anomalous pixels were treated as annotation artifacts and reassigned to the background class. The corrected files were subsequently revalidated to ensure label consistency across the dataset.

3.3 Instance-Level Representation

The proposed framework operates at the nucleus level rather than the pixel level. Each annotated nucleus is treated as an independent sample and represented using complementary feature groups. Morphological descriptors are extracted from instance masks, probability statistics are obtained from DeepLabV3+ and SegFormer ensemble probability maps, and agreement-based measures quantify consensus and disagreement between the two segmentation branches. This representation transforms pixel-wise predictions into biologically meaningful descriptors and provides a unified feature space for the subsequent XGBoost meta-learner.

4. PROPOSED METHODOLOGY

4.1 Framework Overview

The proposed framework integrates semantic segmentation, instance-level feature engineering, ensemble learning, and explainable machine learning to classify nucleus phenotypes in immunohistochemistry images. The workflow consists of six sequential stages: (i) DeepLabV3+ ensemble training, (ii) SegFormer-B2 ensemble training, (iii) generation of ensemble probability maps, (iv) nucleus-level feature extraction, (v) XGBoost-based stacking classification, and (vi) SHAP-based explainability analysis. Rather than combining segmentation outputs directly at the pixel level, probability information from the two segmentation branches is aggregated for each annotated nucleus using the corresponding instance masks. Figure 1 presents the overall architecture of the proposed framework.

The design was motivated by the observation that convolutional and transformer architectures often produce complementary predictions instead of identical ones. DeepLabV3+ captures local texture and boundary information through convolutional operations, whereas SegFormer-B2 exploits self-attention to model wider spatial context. These differences become particularly apparent in challenging IHC regions containing overlapping nuclei, irregular staining patterns, or heterogeneous tissue structures, where one model may assign high confidence while the other remains uncertain. Conventional ensemble strategies generally average pixel-wise probabilities or apply majority voting, thereby reducing prediction variance but discarding information about the degree of agreement between individual models. In this work, disagreement is treated as a measurable characteristic of each nucleus rather than as undesirable noise, allowing the classifier to learn situations in which complementary predictions provide additional discriminatory information.

Another important design choice is the transition from pixel-level predictions to nucleus-level representations. Clinical interpretation and biomarker assessment are performed at the level of individual nuclei rather than isolated pixels. Aggregating probability distributions within each annotated nucleus therefore produces descriptors that are more closely aligned with the biological entities evaluated by pathologists. This representation also enables the integration of heterogeneous information, including nuclear morphology, prediction confidence, and agreement statistics, into a single feature space suitable for supervised learning. Such descriptors are easier to interpret than dense probability maps and provide a direct link between segmentation outputs and phenotype classification.

The framework also separates segmentation from classification, allowing each stage to address a distinct objective. The segmentation models focus on estimating reliable class probabilities for every pixel, while the stacking classifier learns how different feature groups contribute to phenotype prediction across nuclei. This modular formulation offers practical advantages because individual segmentation networks or feature sets can be replaced without redesigning the entire pipeline. It also facilitates feature attribution through SHAP analysis, making it possible to determine whether predictions are primarily influenced by morphology, model confidence, or inter-model agreement. The following sections describe each component of the framework in detail, beginning with the construction of the DeepLabV3+ and SegFormer-B2 ensemble models.

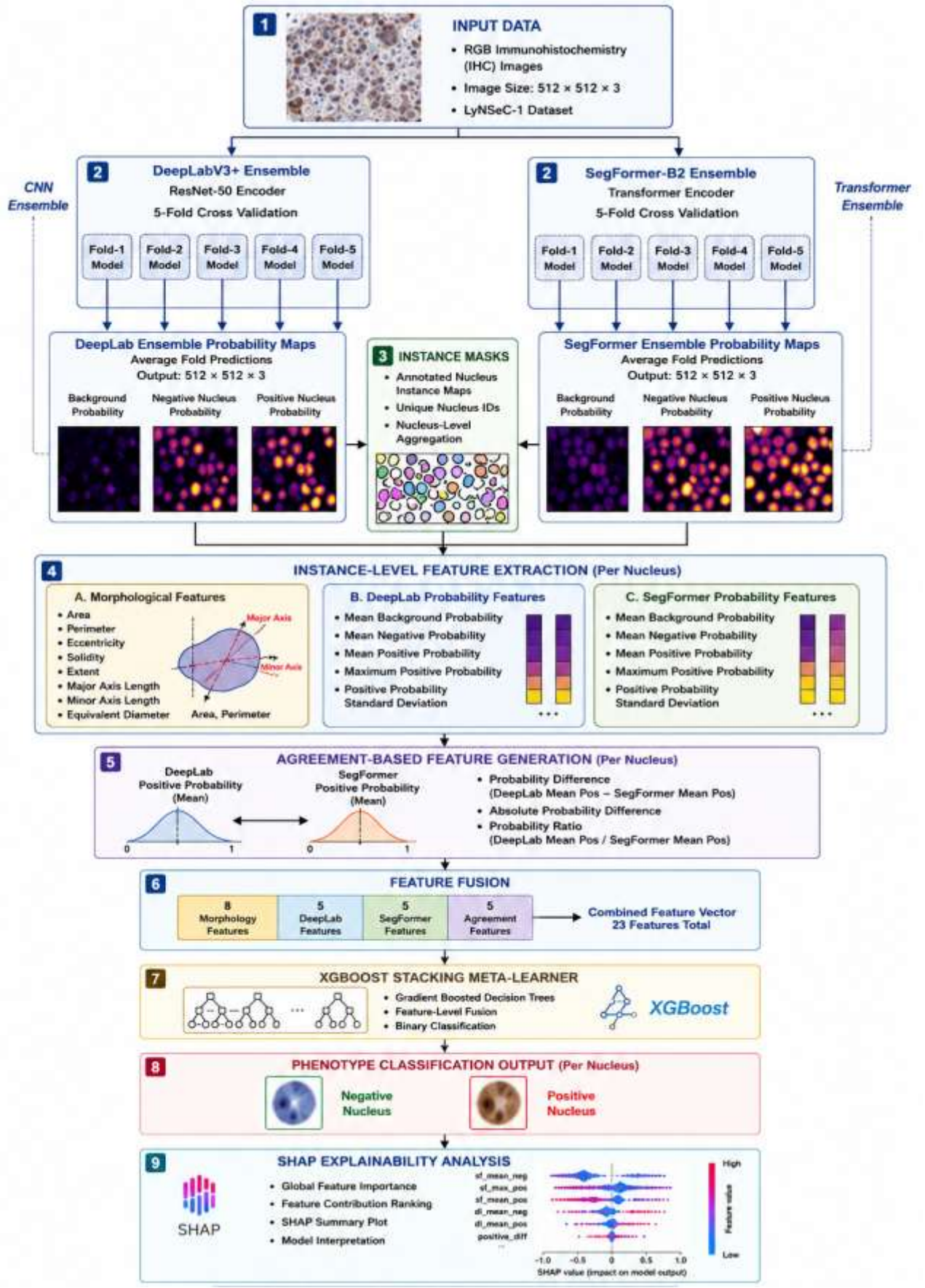

Figure 1. Overall architecture of the proposed explainable instance-level stacking framework. DeepLabV3+ and SegFormer-B2 five-fold ensembles generate probability maps that are aggregated using nucleus instance masks. Morphological descriptors, probability statistics, and agreement-based features are subsequently integrated through an XGBoost meta-learner, while SHAP analysis is used to interpret feature contributions.

4.2 DeepLabV3+ and SegFormer-B2 Ensembles

Two complementary segmentation branches were employed. The first branch uses DeepLabV3+ with a ResNet-50 encoder, which captures multi-scale contextual information through Atrous Spatial Pyramid Pooling. The second branch employs SegFormer-B2, a transformer-based architecture capable of modeling long-range spatial relationships. Both networks were trained independently using five-fold cross-validation, producing five models per branch. Ensemble probability maps were generated by averaging fold-specific predictions, thereby reducing prediction variance and providing more stable probability estimates.

4.3 Ensemble Probability Maps

For each segmentation branch, the outputs of the fivefold-specific models were averaged to obtain ensemble probability maps. If $P_k(x, y, c)$ denotes the probability predicted by fold k for class c at location (x, y), the ensemble probability is given by

$$P_{ens}(x, y, c) = \frac{1}{5} \sum_{k=1}^{5} P_k(x, y, c) \qquad (1)$$

The resulting probability maps have dimensions of 512 × 512 × 3 and represent pixel-wise probabilities for background, negative nuclei, and positive nuclei. Ensemble averaging was adopted to reduce sensitivity to variations among individual folds.

4.4 Feature Extraction and Agreement Modelling

Phenotype classification is performed at the nucleus level. Each annotated nucleus is treated as an independent sample and represented using three complementary feature groups. Eight morphological descriptors are extracted from the instance masks, including area, perimeter, eccentricity, solidity, extent, major axis length, minor axis length, and equivalent diameter. Probability-based statistics are then computed separately from DeepLabV3+ and SegFormer ensemble outputs. In addition, agreement-based features are derived from the positive-class probabilities of the two segmentation branches, including probability difference, absolute difference, and probability ratio. Unlike conventional fusion approaches that treat disagreement as noise, the proposed framework considers disagreement as potentially informative evidence. These descriptors collectively form a 23-dimensional feature vector representing each nucleus.

4.5 XGBoost Meta-Learner

All extracted descriptors are concatenated into a unified feature space and used to train an XGBoost classifier. XGBoost was selected because it effectively handles heterogeneous tabular features, models nonlinear interactions, and provides feature importance estimates. The classifier learns the relationship between nuclear morphology, segmentation confidence, and inter-model agreement to predict positive and negative nucleus phenotypes.

4.6 SHAP-Based Explainability

Model interpretability is incorporated using SHapley Additive exPlanations (SHAP). Feature contributions are estimated by measuring the marginal impact of each variable on the classifier output. Global SHAP analysis was performed to identify the most influential descriptors. The analysis revealed that SegFormer-derived probability features and agreement-related variables contributed more strongly than most morphology descriptors, providing insight into how the stacking classifier combines information from the two segmentation branches.

5. Experimental Setup

5.1 Hardware, Preprocessing, and Model Training

All experiments were conducted using Google Colaboratory with GPU acceleration. The framework was implemented in Python using PyTorch and Segmentation Models PyTorch for DeepLabV3+, Hugging Face Transformers for SegFormer-B2, XGBoost for stacking classification, and SHAP for explainability analysis. Supporting libraries included NumPy, Pandas, OpenCV, PIL, and Matplotlib. Each LyNSeC image is represented as a five-channel array of size 512 × 512 × 5. Only the RGB channels were used as inputs to the segmentation networks, while the instance and phenotype annotation channels were reserved for supervision and feature extraction. Images were resized to 384 × 384 prior to training and inference. PIL interpolation was adopted because preliminary experiments provided more consistent image quality and annotation alignment.

DeepLabV3+ with a ResNet-50 encoder and SegFormer-B2 were trained independently using five-fold cross-validation. Ensemble probability maps were subsequently obtained by averaging fold-specific predictions. Segmentation performance was evaluated using Dice similarity coefficient (DSC), intersection over union (IoU), and pixel-wise accuracy. Table 1 presents the average segmentation performance of DeepLabV3+ and SegFormer-B2 across five folds.

Table 1. Segmentation Performance of Individual Models

Model	Dice (DSC)	IoU	Accuracy
DeepLabV3+ (ResNet-50)	0.8284 ± 0.0039	0.7101 ± 0.0058	0.8554 ± 0.0055
SegFormer-B2	0.8449 ± 0.0027	0.7341 ± 0.0040	0.8700 ± 0.0029

5.2 Feature Construction and Evaluation Protocol

Predictions from the five models of each segmentation branch were averaged to generate probability maps of size 512 × 512 × 3. Using annotated instance masks, each nucleus was represented by 23 features comprising eight morphological

descriptors, five DeepLab-derived probability statistics, five SegFormer-derived statistics, and five agreement-based variables. After verification and cleaning, the dataset contained 86,997 nuclei, including 47,966 negative and 39,031 positive samples.

Phenotype classification was performed using XGBoost configured with 500 trees, maximum depth of six, learning rate of 0.05, and subsampling and column sampling ratios of 0.8. Evaluation employed stratified group cross-validation, where image identifiers served as grouping variables to prevent image-level leakage. Classification performance was assessed using accuracy, precision, recall, F1-score, and ROC-AUC.

6. RESULTS AND DISCUSSION

6.1 Segmentation Performance Evaluation

Both segmentation branches were trained under identical five-fold cross-validation settings and evaluated using Dice similarity coefficient (DSC), intersection over union (IoU), and pixel-wise accuracy . SegFormer-B2 consistently outperformed DeepLabV3+, achieving a Dice score of 0.8449 compared with 0.8284 for DeepLabV3+. Similar trends were observed for IoU and accuracy, indicating that transformer-based contextual modelling provides advantages in heterogeneous IHC regions.

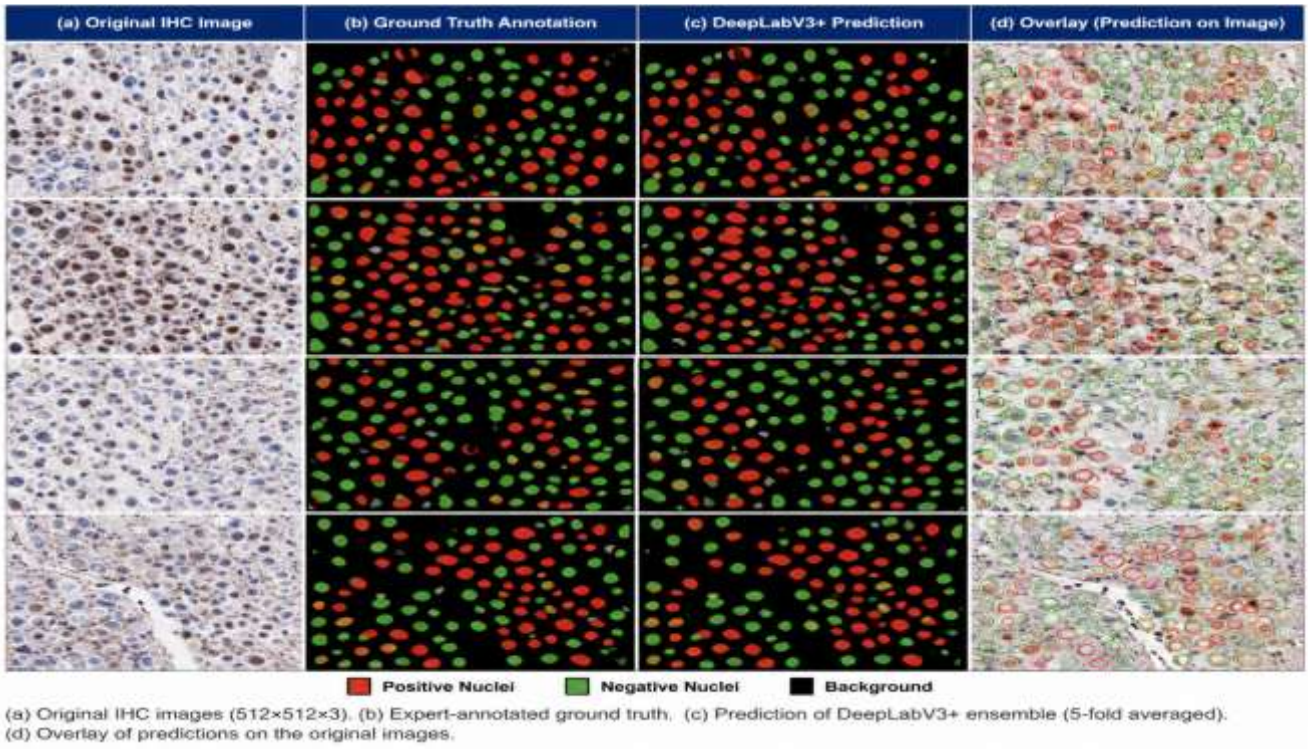

Figure 2. Representative segmentation results produced by the DeepLabV3+ ensemble.

As shown in Figure 2, DeepLabV3+ preserves nuclear boundaries effectively and provides reliable segmentation for isolated nuclei. Visual examination showed that DeepLabV3+ preserved local boundaries effectively, especially for isolated nuclei. Difficulties mainly appeared in weakly stained regions and densely packed structures.

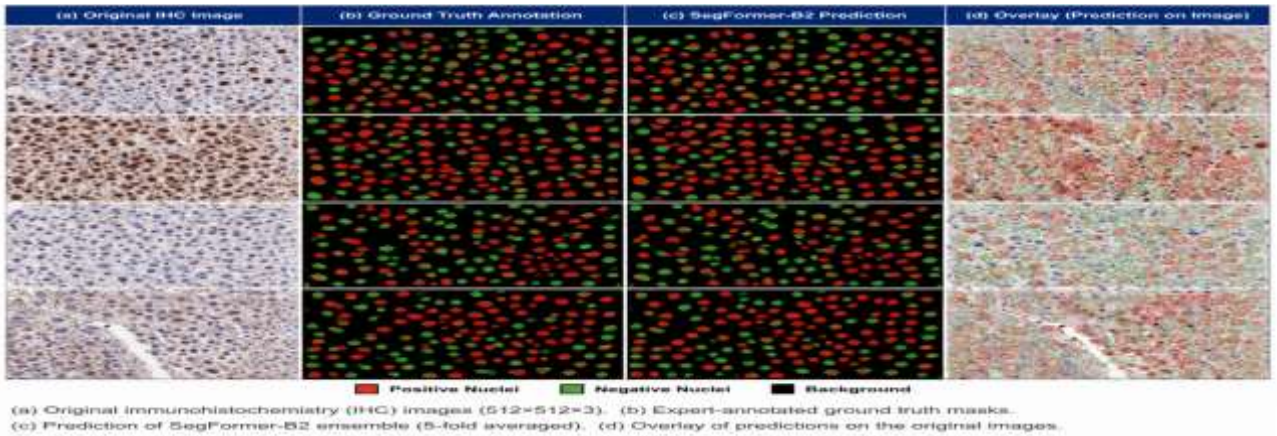

Figure 3. Representative segmentation results obtained from the SegFormer-B2 ensemble.

Figure 3 demonstrates that SegFormer-B2 produces smoother boundaries and better contextual consistency. SegFormer-B2 generated smoother boundaries and fewer fragmented regions. Its attention mechanism appeared particularly useful in

clustered regions, whereas DeepLabV3+ retained complementary local information. These observations motivated their integration rather than selecting a single architecture.

6.2 Nucleus Phenotype Classification Performance

Following nucleus-level aggregation and feature fusion, the centroid-independent model achieved 95.94% accuracy, 95.04% precision, 95.97% recall, 95.50% F1-score, and a ROC-AUC of 99.24% . Precision and recall remained closely balanced, indicating that performance was not achieved by favouring one phenotype class over the other. Table 2 summarizes the overall phenotype classification performance of the proposed model.

Table 2. Classification Performance of the Final Model

Metric	Value
Accuracy	95.94%
Precision	95.04%
Recall	95.97%
F1-score	95.50%
ROC-AUC	99.24%

6.3 Confusion Matrix Analysis

Among negative nuclei, 46,089 were correctly identified, while 37,468 positive nuclei were classified correctly. False positives and false negatives numbered 1,877 and 1,563, respectively. The corresponding confusion matrix statistics are summarized in Table 3.

Table 3. Confusion Matrix Statistics

Actual Class	Predicted Negative	Predicted Positive
Negative	46,089	1,877
Positive	1,563	37,468

Lower false-negative counts are encouraging because missing positive nuclei directly affects biomarker estimation. Most errors occurred in nuclei exhibiting weak staining or ambiguous boundaries.

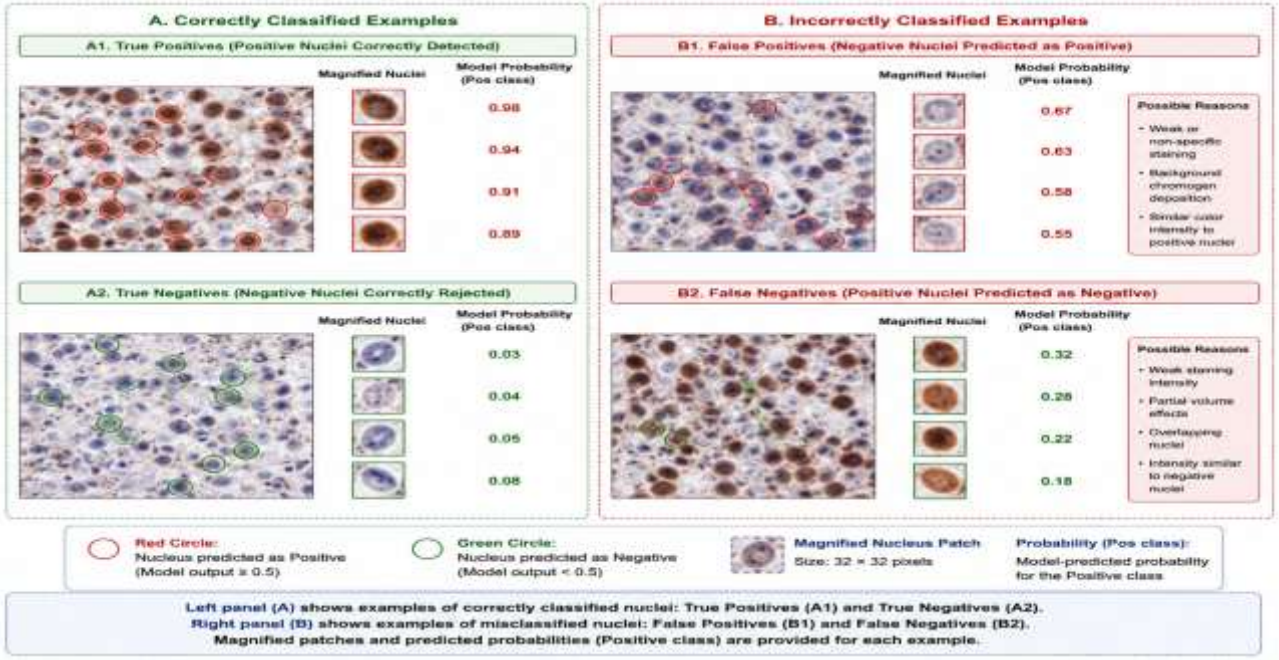

Figure 4. Representative examples of true positives, true negatives, false positives, and false negatives.

Figure 4 reveals that most misclassified nuclei correspond to weakly stained regions and ambiguous boundaries. Inspection of failure cases suggested that ambiguity in staining intensity contributed more strongly to misclassification than segmentation errors themselves.

6.4 SHAP Explainability Analysis

SHAP analysis was used to examine how different feature groups influenced the XGBoost classifier .

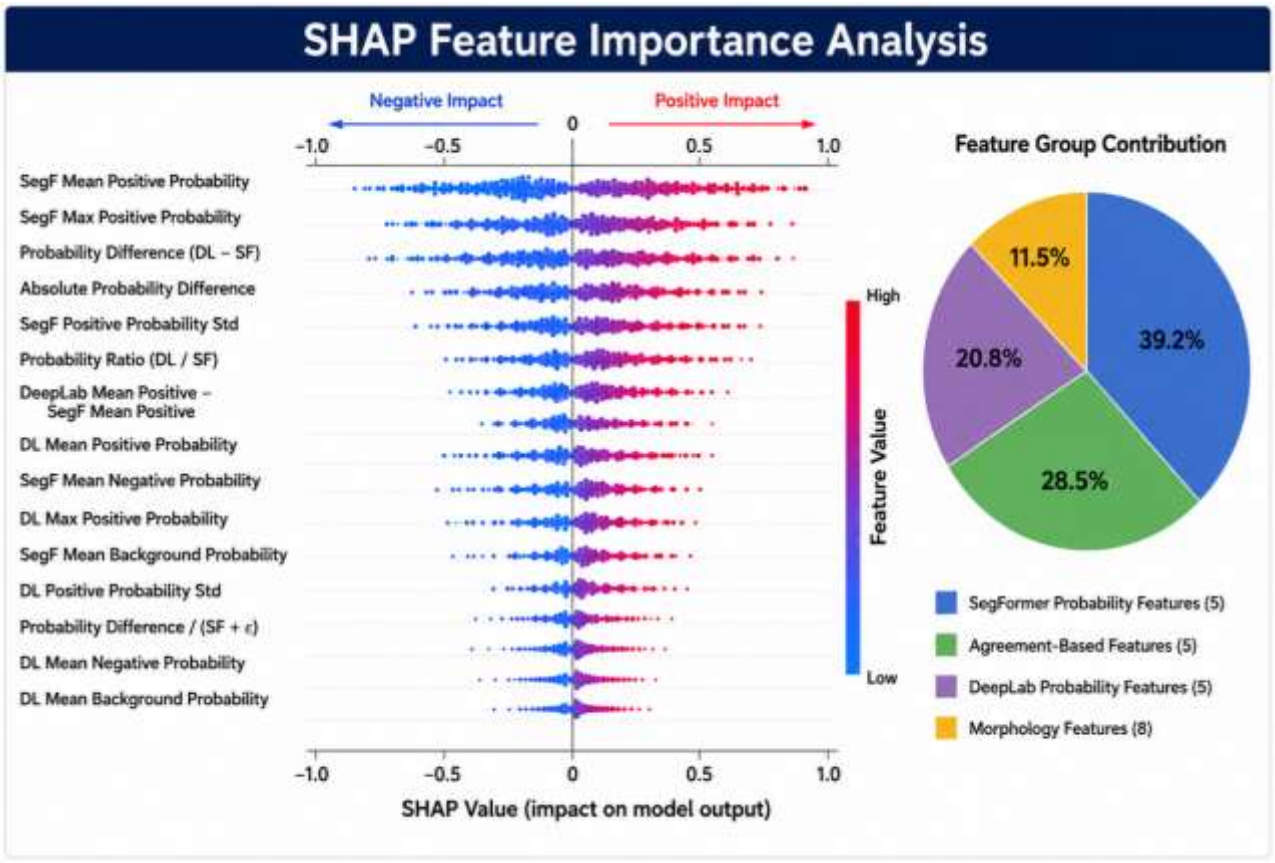

Figure 5. SHAP summary plot illustrating feature contributions.

Figure 5 shows that SegFormer-derived features dominate feature importance rankings. SegFormer-derived variables, particularly sf_mean_neg, sf_max_pos, and sf_mean_pos, exhibited the highest SHAP values. DeepLab probability features also ranked highly, confirming that both architectures provided complementary information. Agreement-based descriptors, especially positive_diff, emerged among the influential variables, indicating that disagreement between CNN and transformer predictions contains useful phenotype-related information rather than merely reflecting uncertainty. Morphological descriptors contributed less strongly, suggesting that staining-related probability distributions carry greater discriminative value than nuclear geometry alone.

6.5 Ablation Study and Discussion

Ablation experiments were conducted to examine how each feature group contributed to the final phenotype classification performance. Rather than evaluating only the complete stacking framework, progressively simpler configurations were constructed by removing or isolating individual feature categories. This analysis was intended to determine whether the observed performance resulted from a particular segmentation model or from the interaction between complementary sources of information.

The results presented in Table 4 indicate that morphological descriptors alone provided limited discriminative capability, achieving an accuracy of 57.71%. Although features such as area, perimeter, eccentricity, and solidity describe nuclear geometry, they do not adequately capture variations in staining intensity or protein expression that distinguish positive and negative nuclei in immunohistochemistry images. In contrast, probability features derived from the segmentation models preserved confidence information learned directly from image appearance, resulting in substantially higher classification performance. SegFormer-B2 outperformed DeepLabV3+ when evaluated independently, which is consistent with its stronger segmentation accuracy reported in Section 6.1, yet neither branch alone achieved the performance obtained through feature fusion.

A notable observation is the contribution of agreement-based descriptors. These variables encode the relationship between CNN and transformer predictions instead of relying solely on their individual confidence scores. Their inclusion consistently improved or maintained classification performance, suggesting that disagreement between heterogeneous models often reflects informative uncertainty rather than prediction error. The marginal reduction observed after removing centroid coordinates also indicates that the classifier relied primarily on morphology, probability statistics, and inter-model agreement instead of positional information, reducing the likelihood of learning dataset-specific spatial biases. Taken together, the ablation results support the design choices introduced in Section 4. The performance gains arise from combining complementary representations at the nucleus level rather than from any single model or feature category, demonstrating that the proposed stacking strategy benefits from integrating local structural information, contextual probability estimates, and explicit measures of model agreement.

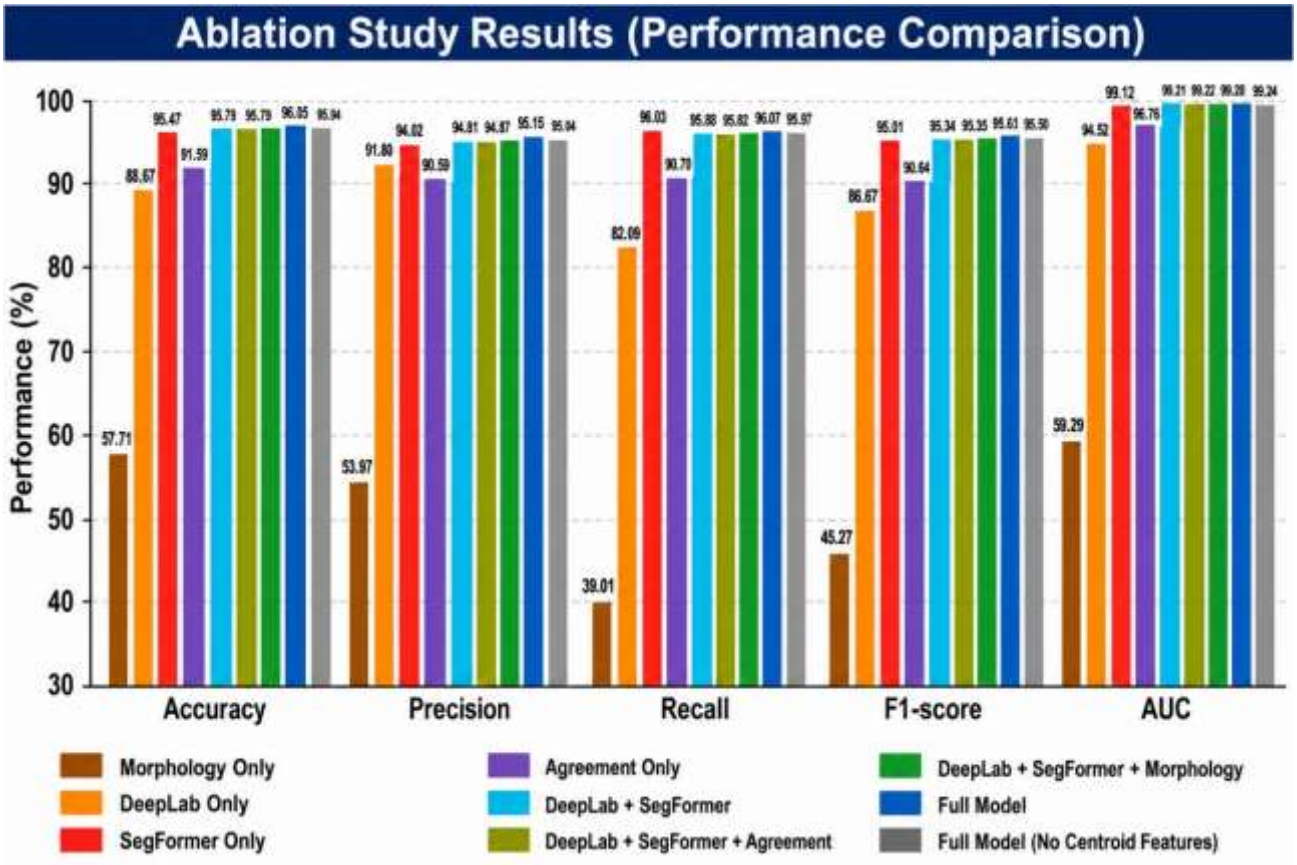

Figure 6. Ablation study illustrating the influence of different feature combinations.

Figure 6 clearly demonstrates the progressive improvement obtained when multiple feature groups are combined.

Table 4. Ablation Study Results

Experiment	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
Morphology Only	57.71	53.97	39.01	45.27	0.5929
DeepLab Only	88.67	91.80	82.09	86.67	0.9452
SegFormer Only	95.47	94.02	96.03	95.01	0.9912
Agreement Only	91.59	90.59	90.70	90.64	0.9676
DeepLab + SegFormer	95.79	94.81	95.88	95.34	0.9921
DeepLab + SegFormer + Agreement	95.79	94.87	95.82	95.35	0.9922
DeepLab + SegFormer + Morphology	96.03	95.15	96.07	95.61	0.9928
Full Model	96.05	95.25	96.01	95.63	0.9928
Full Model (No Centroid Features)	95.94	95.04	95.97	95.50	0.9924

Morphology alone provided poor discrimination, indicating that nuclear geometry is insufficient for reliable phenotype classification. SegFormer-derived probabilities proved more informative than DeepLab-derived probabilities, yet neither model alone achieved the performance obtained through feature fusion. Agreement features also demonstrated substantial predictive value.

The complete model produced the highest accuracy. Removing centroid coordinates caused only a marginal decline, suggesting that classification decisions relied primarily on biologically meaningful descriptors rather than positional information. Taken together, these findings support the central premise of this work: combining CNN-derived probabilities, transformer-derived probabilities, and inter-model agreement at the nucleus level provides more reliable phenotype classification than any individual source of information.

7. CONCLUSION AND FUTURE WORK

This study presented an explainable instance-level stacking framework for nucleus phenotype classification in immunohistochemistry images. Instead of relying on a single segmentation architecture or direct pixel-level fusion, the method integrates DeepLabV3+ and SegFormer-B2 through nucleus-level feature aggregation and XGBoost-based meta-learning. Ensemble probability maps generated from five-fold cross-validation were converted into biologically meaningful descriptors describing nuclear morphology, prediction confidence, and inter-model agreement. This design

was motivated by the observation that convolutional and transformer architectures often capture different aspects of tissue organization and staining patterns.

Experiments on the LyNSeC dataset showed that segmentation-derived probability features provide substantially greater discriminative power than morphology alone. Although SegFormer-B2 achieved higher segmentation accuracy, DeepLabV3+ supplied complementary information that improved phenotype classification when both branches were combined. The final centroid-independent model achieved 95.94% accuracy and a ROC-AUC of 99.24% under a stratified group cross-validation protocol designed to prevent image-level leakage. SHAP analysis indicated that SegFormer-derived probabilities and agreement-based descriptors were among the most influential variables, while ablation experiments demonstrated that no individual feature group matched the performance obtained through feature fusion.

Several challenges remain before broader clinical adoption can be considered. Extending the framework with foundation models graph-based cellular representations, and uncertainty-aware learning may improve generalization across different staining protocols and tissue types. External validation on larger multi-center cohorts and whole-slide images will also be necessary to determine how well the proposed methodology translates from controlled experimental settings to routine digital pathology workflows. Future work may also benefit from incorporating benchmark segmentation frameworks and larger multicenter nuclei datasets to improve generalization across staining protocols and tissue types.

REFERENCES

1. G. Bera et al., “Artificial Intelligence in Digital Pathology,” Nature Reviews Clinical Oncology, 2019.
2. A. Janowczyk and A. Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," Journal of Pathology Informatics, vol. 7, 2016.
3. D. Komura and S. Ishikawa, "Machine Learning Methods for Histopathological Image Analysis," Computational and Structural Biotechnology Journal, vol. 16, pp. 34-42, 2018.
4. F. Mahmood et al., “Deep Learning for Computational Pathology,” Nature Reviews Clinical Oncology, 2020.
5. W. Bulten et al., "Artificial Intelligence for Diagnosis and Gleason Grading of Prostate Cancer," The Lancet Oncology, vol. 21, no. 2, pp. 233-241, 2020.
6. G. Campanella et al., "Clinical-grade computational pathology using weakly supervised deep learning on whole-slide images," Nature Medicine, vol. 25, pp. 1301-1309, 2019.
7. A. Dosovitskiy et al., “An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale,” ICLR, 2021.
8. M. Weigert and U. Schmidt, “Nuclei Instance Segmentation and Classification in Histopathology Images with StarDist,” 2022.
9. C. Stringer et al.,"Cellpose: A Generalist Algorithm for Cellular Segmentation," Nature Methods, vol. 18, pp. 100–106, 2021.
10. S. Lundberg and S. Lee, “A Unified Approach to Interpreting Model Predictions,” NeurIPS, 2017.
11. G. Litjens et al., "A Survey on Deep Learning in Medical Image Analysis," Medical Image Analysis, vol. 42, pp. 60-88, 2017.
12. O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” MICCAI, 2015.
13. L. C. Chen et al., “Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation,” ECCV, 2018.
14. R. J. Chen et al., “Towards a General-Purpose Foundation Model for Computational Pathology,” Nature Medicine, vol. 30, pp. 850–862, 2024.
15. M. Y. Lu et al., “A Visual-Language Foundation Model for Computational Pathology,” Nature Medicine, vol. 30, pp. 863–874, 2024.
16. E. Vorontsov et al., “A Foundation Model for Clinical-Grade Computational Pathology,” Nature Medicine, 2024.
17. H. Xu et al., “A Whole-Slide Foundation Model for Digital Pathology from Real-World Data,” Nature, 2024.
18. C. Xiong et al., “A Survey of Pathology Foundation Models: Progress and Future Directions,” IJCAI, 2025.
19. T. Ding et al., “A Multimodal Whole-Slide Foundation Model for Pathology,” Nature Medicine, 2025.
20. A. Kirillov et al., "Segment Anything," ICCV, 2023.
21. J. Ma et al., "Segment Anything in Medical Images (MedSAM)," Nature Communications, vol. 15, 2024.
22. Z. Liu et al., “Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows,” *ICCV*, 2021.
23. J. Valanarasu et al.,"Medical Transformer: Gated Axial-Attention for Medical Image Segmentation," MICCAI, 2021.
24. J. Chen et al.,"TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," arXiv, 2021.
25. F. Isensee et al., "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," Nature Methods, vol. 18, pp. 203-211, 2021.
26. S. Graham et al., “Hover-Net: Simultaneous Segmentation and Classification of Nuclei,” Medical Image Analysis, vol. 58, 2019.
27. H. Naji et al., “Segmentation of Histological Images of Diffuse Large B-Cell Lymphoma,” Computers in Biology and Medicine, vol. 168, 2024.
28. H. Naji et al., “HoLy-Net: Segmentation of Histological Images of Diffuse Large B-Cell Lymphoma,” 2024.
29. S. Graham et al., “Lizard: A Large-Scale Dataset for Colonic Nuclear Instance Segmentation and Classification,” ICCV, 2021.
30. A. Mahbod et al., “NuInsSeg: A Fully Annotated Dataset for Nuclei Instance Segmentation,” 2023.
31. J. Gamper et al.,"PanNuke Dataset Extension, Insights and Baselines," arXiv, 2020.
32. S. Graham et al.,"CoNIC: Colon Nuclei Identification and Counting Challenge 2022," Medical Image Analysis, vol. 87, 2023.

33. R. Guidotti et al., “A Survey of Methods for Explaining Black Box Models,” ACM Computing Surveys, 2018.
34. H. Naji et al., “LyNSeC: Lymphoma Nuclear Segmentation and Classification Dataset,” Zenodo, 2023.
35. E. Xie et al., “SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers,” NeurIPS, 2021.
36. G. Jaume et al., "Modeling Dense Nuclei Interactions for Mitosis Detection and Classification," Medical Image Analysis, vol. 84, 2023.
37. T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” KDD, 2016.
38. S. Lundberg et al., “Explainable AI for Trees,” Nature Machine Intelligence, 2020.
39. K. He et al., “Deep Residual Learning for Image Recognition,” CVPR, 2016.