Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences

School for Information Technology

Master of Statistics and Data Science

Master's thesis

Systematic Review and Comparative Analysis of Cell Segmentation Methods for Whole Slide Imaging

Isaiah M'camwata

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Data Science

SUPERVISOR :
Prof. dr. Dirk VALKENBORG

2024
2025

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Systematic Review and Comparative Analysis of Cell Segmentation Methods for Whole Slide Imaging

Isaiah M'camwata
Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Data Science

SUPERVISOR :
Prof. dr. Dirk VALKENBORG

Acknowledgements

In the midst of challenges, a captain navigates with courage and guidance, steering through both calm and storm.My academic journey at Hasselt University has been a similar voyage-marked by growth, discovery, and perseverance. Along this path, I have been fortunate to receive unwavering support and mentorship from individuals who provided stability in difficult times and guided me toward success.

First and foremost, I extend my deepest gratitude to my professors, led by Prof. Dr. Dirk Valkenborg. Your unwavering commitment to academic excellence and mentorship has been invaluable. You have challenged me, encouraged me, and provided the guidance necessary to navigate this journey. For that, I am truly grateful.

To Melvin and Joachim, I sincerely appreciate your insights and direction, which have been fundamental to this research. Your support has shaped the core of this work, and I cannot adequately express how much I value your contributions.

To my beloved wife, Jane, and my children, Juana and Karl-thank you for your unwavering support and patience. Despite my absence over the past two years, you stood by me with love and strength. I could not have reached this point without you.

Lastly, to my sister, Prof. Dr. Dorothy Amwata, your example has been a constant source of inspiration. Thank you for always encouraging me to dream big and pursue my goals with conviction.

Abstract

Cell and nucleus segmentation is critical for quantitative bioimage analysis, particularly in whole slide imaging (WSI), which has transformed digital pathology by enabling the study of entire tissue sections. Despite recent advancements, scalable and efficient segmentation methods are still needed to handle complex, real-world datasets. This thesis presents a systematic review and comparative analysis of state-of-the-art cell segmentation techniques, focusing on their application to Fluorescence, Hematoxylin, and Eosin (H&E) stained WSIs. Five prominent methods—Cellpose, StarDist, Mesmer, HoverNet, and InstaSeg—were evaluated using standardized, publicly available datasets: TissueNet and MoNuSeG. Performance was assessed using accuracy, precision, recall, F1-score, and Intersection over Union (IoU). Results indicate that StarDist consistently outperforms the other models regarding precision across both imaging modalities, while Cellpose tends to over-segment, often predicting a higher number of nuclei. These contrasting behaviors suggest that combining the strengths of different models could be a promising direction for improving segmentation accuracy. While not directly evaluated in this study, integrating a composite loss function would be a compelling area for future exploration. A sensitivity analysis employing StainGANs quantified the influence of stain variations on model robustness, providing significant insights into how deviations from standardized staining impact segmentation performance. The outcomes of this study are to guide researchers in selecting optimal segmentation approaches for WSI analysis and highlight potential directions for future improvements in digital pathology workflows.

Contents

1 Introduction **4**
1.1 Background 4
1.2 Research Objectives 5

2 Materials and Methods **5**
2.1 Data Description 7
2.1.1 TissueNet Dataset 7
2.1.2 MoNuSeg Dataset 8
2.2 Models 9
2.2.1 StarDist 9
2.2.2 Cellpose 10
2.2.3 Instaseg 12
2.2.4 Mesmer 14
2.2.5 HoverNet 15
2.3 Model Evaluation 16
2.3.1 Classical Segmentation Metrics 16
2.3.2 Multidimensional Scaling 18
2.3.3 Pair -Wise Bland-Altman plots 19
2.3.4 HiStauGAN- Sensitivity To Stain Variation Analysis 19

3 Results **21**
3.1 Fluorescence Image Analysis 21
3.1.1 Analysis with Ground Truth Masks 21
3.1.2 Multidimensional Scaling Analysis 22
3.1.3 Pair -Wise Bland-Altman plots Analysis 23
3.2 Analysis Of Hematoxylin And Eosin Images 26
3.2.1 Analysis With Ground Truth Masks 26
3.2.2 Multidimensional Scaling Analysis 27
3.2.3 Pair-Wise Bland-Altman Plots Analysis 28
3.3 Sensitivity Analysis 29
3.3.1 Pair-Wise Bland-Altman Plots Sensitivity Analysis (CWZ vs UMCU Domain) 32

4 Discussion **34**

5 Societal Relevance,Ethical Considerations and Key Stakeholders **37**

6 Conclusion **38**

Appendix **43**

1 Introduction

1.1 Background

Before the advent of computers and digital image analysis, cell segmentation—identifying and separating individual cells or nuclei in biological tissues—was performed manually using microscopes and physical tools, with results interpreted primarily by human observers. Pathologists examined stained tissue samples under microscopes, using visual indicators such as cell boundaries, coloration, morphology, and contrast to distinguish and delineate cellular structures. The process was highly labor-intensive and time-consuming, often requiring the manual counting of large numbers of cells, and was inherently susceptible to variability between observers. With the emergence of digital microscopy and early image processing techniques in the pre-AI era, software tools began to assist with basic segmentation tasks, offering limited levels of automation and reproducibility. Recent advances in computational imaging and artificial intelligence have significantly transformed the field, enabling highly scalable, fully automated, and more accurate approaches to cell segmentation—thereby reshaping modern workflows in digital pathology.

Detection and segmentation of nuclei are fundamental to pathology-based diagnoses, including carcinoma detection, grading, and quantitative analysis, all of which contribute to the reliability of clinical decisions [1]. These techniques are essential in both academic research and medical applications, serving as the foundation for developing advanced diagnostic tools [2].

In recent years, bioimage analysis has garnered significant attention in medical research, driving advancements in cellular analysis and understanding. A major factor behind this progress is the improvement in computational efficiency and power, as modern GPUs can scale up to process vast amounts of image data with ease [3].

Cell segmentation approaches can be broadly categorized into two groups: traditional methods and deep learning-based techniques. Traditional segmentation methods, such as thresholding, edge-based techniques, region-based approaches, clustering, and graph-based segmentation, are generally computationally efficient and interpretable. However, they often struggle in complex scenarios, particularly when handling noisy images or overlapping nuclei and cells [4].

In contrast, deep learning-based methods have demonstrated remarkable effectiveness in addressing these challenges by learning intricate, high-dimensional representations of cellular structures through convolutional frameworks. This paper focuses on comparing segmentation models developed using deep learning and convolutional neural networks, assessing their capabilities in improving accuracy and robustness in cellular image analysis.

The range of methods developed for nuclei detection and segmentation is extensive. However, this research focuses on a selection of state-of-the-art models that are well-documented and widely recognized in the field, including Mesmer, Cellpose, StarDist, HoverNet, and InstaSeg [5, 6, 7, 8, 9].

While reviewing existing literature, it became evident that although numerous studies compare deep learning-based segmentation models, truly comprehensive and independent evaluations are relatively scarce. Many of the available assessments appear to be conducted by the model developers themselves, which raises potential concerns regarding impartiality. Furthermore, many evaluations do not utilize standardized and diverse datasets like TissueNet, making it difficult to objectively compare model performance across various tissue types and imaging conditions. This project aims to bridge that gap by systematically comparing state-of-the-art deep learning segmentation models using the TissueNet benchmark dataset, providing researchers with clearer insights into the strengths and limitations of different approaches.
In this project, we deploy StainGANs to perform a sensitivity analysis aimed at evaluating the robustness of segmentation models under varying staining conditions. This approach helps identify models that demonstrate greater stability and generalizability across heterogeneous histopathological datasets.

1.2 Research Objectives

The primary aim of this research is to evaluate and compare the performance of state-of-the-art cell and nuclei segmentation models applied to histopathological images. Specifically:

- To conduct a comparative performance analysis of leading segmentation models—Cellpose, StarDist, Mesmer and InstaSeg—based on standard evaluation metrics including accuracy, precision, recall, F1-score, and Intersection over Union (IoU).
- To assess the robustness of these models when applied to datasets with varying staining characteristics, using stainGANs for stain normalization.

2 Materials and Methods

This section outlines both the material and methodological framework employed in this study, focusing on the application of five state-of-the-art deep learning-based cell segmentation methods built upon the U-Net architecture Figure 1. This section is structured into four main components. The first and second parts (Sections 2.1.1 and 2.1.2) provide an overview of the datasets used. The subsequent section presents a detailed overview of image analysis using fluorescence microscopy images from the TissueNet dataset, including descriptions of the deployed models: StarDist, Cellpose, InstaSeg, and Mesmer.To assess the generalizability of segmentation methods across different staining modalities, we apply the same segmentation techniques to H&E-stained images. We also perform a sensitivity analysis using stain translation via generative adversarial networks to evaluate the robustness of segmentation under varying stain appearances. Without ground truth masks, we use

Bland–Altman plots and multidimensional scaling as statistical tools to assess agreement and visualize differences between segmentation outputs [10].

Two widely used imaging techniques in histology are Hematoxylin and Eosin (H&E) staining and fluorescence-based imaging. Hematoxylin and Eosin staining uses two dyes: hematoxylin, which stains nuclei blue-purple, and eosin, which stains the cytoplasm and other structures pink. This method provides a general overview of tissue architecture and cell morphology and is the most commonly used technique in routine pathology for examining tissue structure and detecting abnormalities. Fluorescence imaging employs fluorescent dyes or tags that bind to specific molecules within tissues. When exposed to certain wavelengths of light, these dyes fluoresce at different wavelengths, allowing for multi-target labeling and precise molecular visualization. Fluorescence techniques work across diverse tissue types, providing high-contrast imaging of cytoplasmic, nuclear, and extracellular matrix structures [11, 12, 13]. This study focuses primarily on these two techniques.

A significant challenge in bioimage analysis is staining variability, particularly in histopathological images stained with Hematoxylin and Eosin.Variations in staining protocols, reagent concentrations, scanner types, and even tissue preparation practices can lead to substantial differences in color, contrast, and intensity across images. These inconsistencies directly affect the performance and generalizability of segmentation models, which are often sensitive to the visual characteristics of the training data. As a result, a model trained on one dataset may perform poorly when applied to another with different staining conditions, thereby limiting its robustness in real-world clinical settings[14].

In this study, we analyze both the Tissuenet dataset for fluorescence images and the Monuseg dataset for H&E-stained images, using four state-of-the-art models specifically designed for the segmentation of cell nuclei.

All deep learning models examined in this study, except for Mesmer, are based on the U-Net architecture, which combines encoder and decoder networks.

Figure 1 illustrates the architecture for U-Net. The framework is a symmetric encoder-decoder network designed for image segmentation. The encoder extracts spatial features through repeated blocks of two 3×3 convolutions followed by 2×2 max pooling, doubling the number of filters at each level. The decoder mirrors this structure, using 2×2 transposed convolutions to upsample and halve the number of filters, followed by two 3×3 convolutions. A final 1×1 convolution produces the segmentation map. ReLU is used in all layers except the last, which uses a Sigmoid activation. A key innovation is skip connections, where feature maps from the encoder are concatenated with decoder outputs at each level, preserving spatial information lost during downsampling.

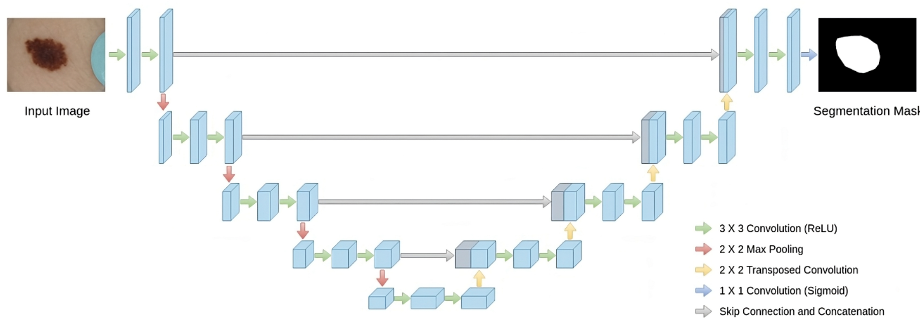

Figure 1: U-Net architecture with encoder-decoder structure for cellular image segmentation.

Some of the key software tools extensively used in this study include Deepnote notebooks, Google Colab and Drive, GitHub repositories, the TIA Toolbox[15], and reproducible codebases from StarDist[16, 7, 6], InstaSeg[17], Cellpose[8], and DeepCell[9]. The programming environment was based on Python 3.11, with deep learning models implemented using both TensorFlow and PyTorch frameworks.

2.1 Data Description

This study utilizes two publicly available GDPR-compliant datasets for nuclei and whole-cell segmentation. The first is the TissueNet dataset, which contains fluorescence-stained tissue images. The second is the MoNuSeg dataset, which focuses on H&E-stained histopathology images.

2.1.1 TissueNet Dataset

TissueNet dataset, contains fluorescence tissue images. It comprises approximately 2,600 training images (512×512 pixels), from which random 256×256 crops are extracted for data augmentation. The validation set includes around 300 images resized to 256×256 and expanded to approximately 3,000 images using resolution variants. The test set consists of about 300 images, also resized to 256×256, resulting in over 1,200 evaluation samples,the images were acquired with a resolution of 0.61 µm/pixel [9] .All the images have corresponding round truth masks with integer labels. Figure 2 presents a preview of the Tissunet dataset. The dataset includes two imaging channels: the nuclei and cell membrane channels, both stored in compressed array files. The nuclei data is stored in the first channel, while the cell membrane data occupies the second. Similarly, the corresponding labeled masks are provided in compressed arrays, with cell membrane annotations in the first channel and nuclei annotations in the second. This study primarily focused on the nuclei channel of the dataset.

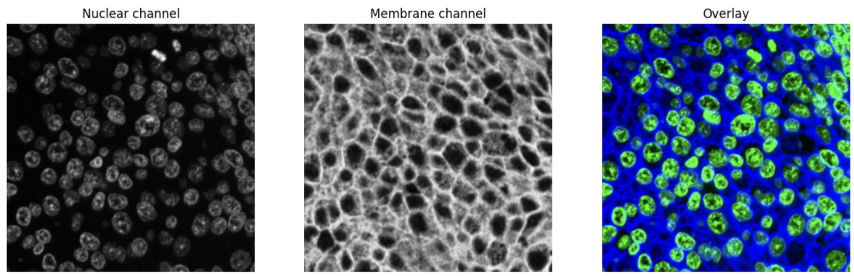

Figure 2: TissueNet Image sample with fluorescence staining and ground truth nuclei masks.

2.1.2 MoNuSeg Dataset

The second dataset used in this study is MoNuSeg (Multi-Organ Nuclei Segmentation). MoNuSeg provides annotated histopathological images focused on H&E-stained tissue. It consists of a training set of 30 H&E-stained images with approximately 22,000 manually annotated nuclear boundaries, originally published in IEEE Transactions on Medical Imaging (2017), and a test set with around 7,000 additional nuclear boundary annotations, released as part of the MoNuSeg 2018, Challenge , the images were obtained at a high resolution of 0.25 µm/pixel [18]. In this study, we used the dataset to evaluate nuclei segmentation performance across different types of staining. Figure 3 shows a preview of both the image and the corresponding annotation from the Monuseg dataset.

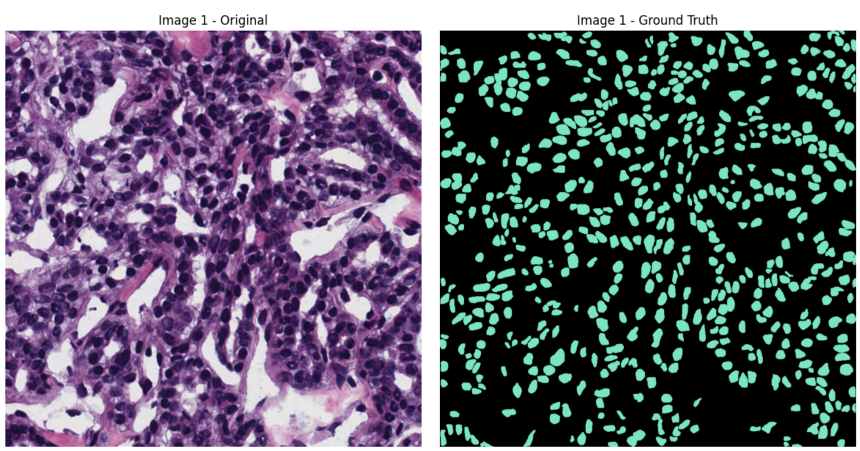

Figure 3: Sample image and corresponding segmentation mask from the Monuseg dataset

2.2 Models

2.2.1 StarDist

Instead of using bounding boxes, StarDist represents objects as star-convex polygons, making it particularly well-suited for round or elliptical structures such as cell nuclei. The method was originally developed with fluorescence microscopy images in mind[5]. For this particular analysis, we utilized the Versatile (fluorescent nuclei) model that was trained on a subset of the DSB 2018 nuclei segmentation challenge dataset.The training data consists of both images and masks, where each pixel is assigned either a unique object identifier or a background label (typically 0). The general approach for 2D image segmentation using StarDist is illustrated in Figure 4, which shows how the model processes images to predict radial distances and object probabilities. The model is trained to predict, for each pixel, the distances to the object boundary along a predefined set of radial directions, as well as the object probability. These predictions generate an overcomplete set of candidate polygons. The final segmented objects are then selected using non-maximum suppression (NMS) to eliminate redundant or overlapping candidates.

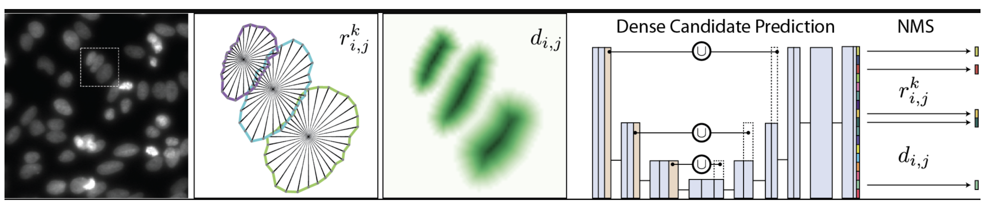

Figure 4: Stardist Framework (adapted from Schmidt et al., 2018).

Radial Distances (r): For each pixel, the model predicts distances from the pixel to the object's boundary along a fixed number of rays . This means that the model learns not directly the full contour of an object but a set of distances that define a star-convex polygon centered at that pixel[16, 6] .

Object Probability (d): Alongside the radial distances, the model predicts how likely it is that the given pixel is at the center of an object (nucleus), filtering out background pixels or non-object centers.

Non-Maximum Suppression (NMS): Since many pixels may predict overlapping or similar polygons, NMS is applied to retain only the most confident (highest probability) polygon predictions while discarding redundant ones, ensuring that each object is detected only, avoiding over-segmentation.

The '2D_versatile_fluo' model was trained using default values: prob_thresh = 0.479071 (Object probability threshold) and nms_thresh = 0.3 (Non-Maximum Suppression threshold) with radial rays of 32 evenly spaced directions.

The combined loss function minimized in the Stardist model is a weighted sum of the distance map loss and the object probability loss:

$$L_{\text{total}} = \lambda_1 \cdot L_{\text{dist}} + \lambda_2 \cdot L_{\text{prob}}$$

- L_{dist} is the loss for the radial distances (MSE).
- L_{prob} is the binary cross-entropy loss for the object probability map.
- λ_1 and λ_2 are hyperparameters that control the relative importance of each loss term.

The object probabilities are minimized using standard binary cross-entropy loss. For the polygon distances, a Mean Absolute Error (MAE) loss is used, where the pixel-wise errors are weighted by the ground truth object probabilities before averaging. This ensures that the model prioritizes the object regions during the optimization process, improving segmentation accuracy in areas where nuclei are present[16].

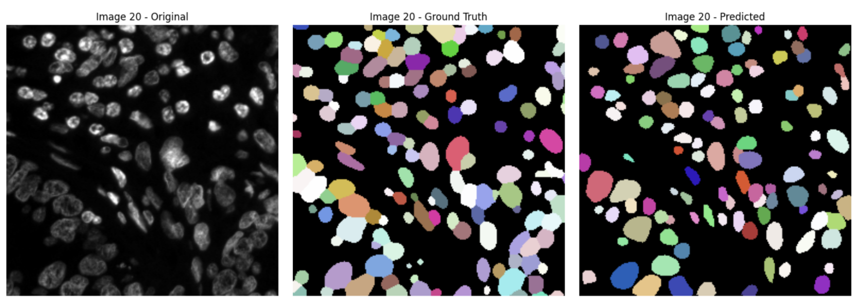

Figure 5: StarDist Input image alongside the ground truth and predicted segmentation masks

2.2.2 Cellpose

Cellpose is a deep learning-based segmentation method designed to address the limitations of traditional segmentation approaches, which often struggle with overlapping nuclei. Cellpose introduced a novel intermediate representation that enforces a smooth topological structure for each object, enabling more robust segmentation across a variety of cellular morphologies[8].

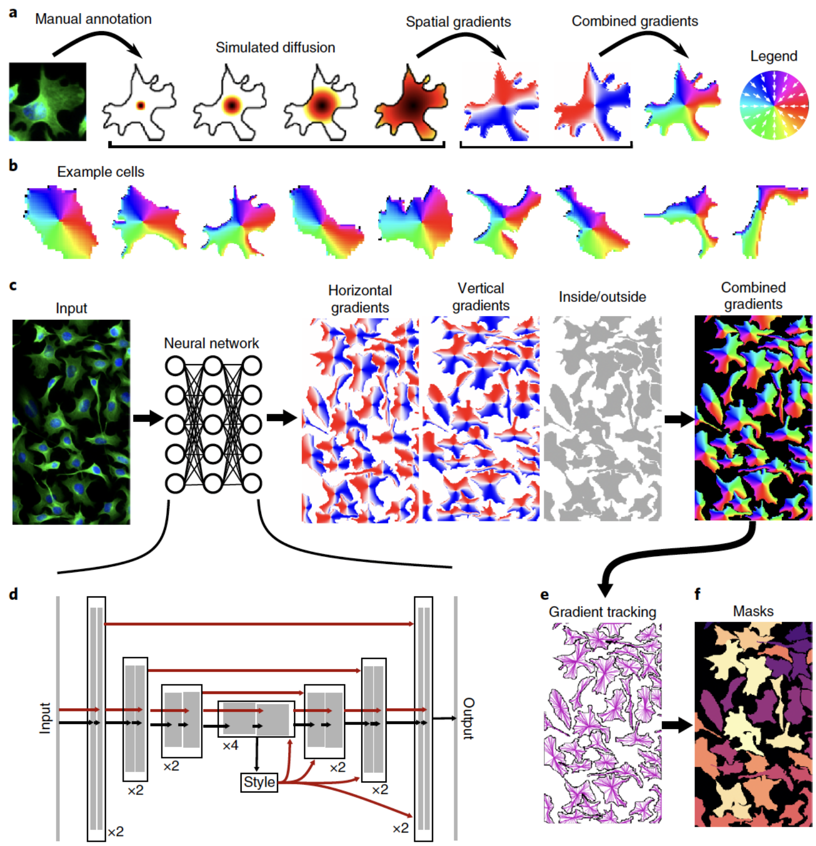

Figure 6: Cellpose framework.Neural network predicts horizontal and vertical vector flows and binary masks from topological maps. Pixels are grouped by gradient tracking. (adapted from [8]).

The method begins by simulating diffusion from human-annotated ground truth masks to generate topological maps where each object forms a single, smooth intensity basin.A neural network is trained to predict both the spatial gradients (horizontal and vertical) of these maps and a binary mask indicating object presence[8].During inference, the model predicts vector fields from which each pixel is traced via gradient tracking to its corresponding object center. Pixels that converge to the same location are grouped together to form a segmented cell. Predicted binary masks are used to refine the boundaries and eliminate false positives. The Cellpose model is anchored on a modified U-Net framework with residual blocks instead of standard convolutional units, direct summation rather than feature concatenation to reduce parameters, and increased depth for improved feature extraction. Additionally, the network incorporates global average pooling at the bottleneck to capture an image's overall "style." This style vector is injected into the upsampling pathway to adapt predictions based on image-specific characteristics[8].
To enhance segmentation quality during inference, Cellpose employs several test-time aug-

mentations, including model ensembling, test-time resizing, region-of-interest (ROI) quality estimation, and image tiling. These enhancements collectively contribute to improved robustness and accuracy across diverse datasets.
The model uses a composite loss function that combines both the accuracy of object segmentation and the correctness of spatial vector flows used for pixel routing. The total loss minimized during training is expressed as:

$$L_{\text{total}} = \lambda_1 \cdot L_{\text{mask}} + \lambda_2 \cdot L_{\text{flow}}$$

- L_{mask} is the binary cross-entropy loss applied for foreground/ background prediction.
- L_{flow} is the mean squared error (MSE) loss between the predicted vector flows (horizontal and vertical gradients) and the ground truth flows derived from simulated diffusion across cell masks.
- λ_1 and λ_2 are scalar weights used to balance the contributions of the two losses. In most training configurations, these are set to equal values.

This dual-objective approach enables the model to delineate cell boundaries while learning a vector field that routes pixels to object centers[8].

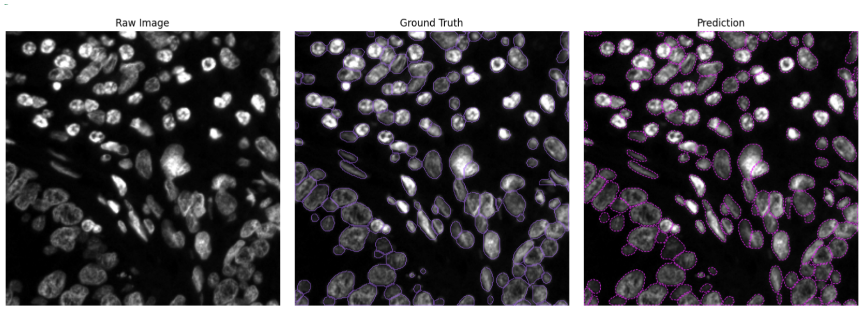

Figure 7: Cellpose input image with corresponding ground truth and predicted segmentation masks

2.2.3 Instaseg

Deep learning models such as CellPose, Mesmer, and StarDist have shown reasonable performance on specific datasets; however, they face limitations, especially when applied to multiplexed imaging [19]. While models like CellPose and Mesmer can technically be retrained with additional imaging channels, they often require users to merge or subset multiple biomarkers. This can result in the loss of biologically relevant information that might be useful for downstream analysis[20]. Furthermore, retraining these models typically demands careful tuning to specific biomarker compositions, which limits their generalizability across diverse datasets [21].

Other approaches, such as nuclear mask expansion or pixel classification using tools like Ilastik[22] and CellProfiler, introduce additional challenges, including user-dependent variability and reduced performance on complex or heterogeneous datasets.

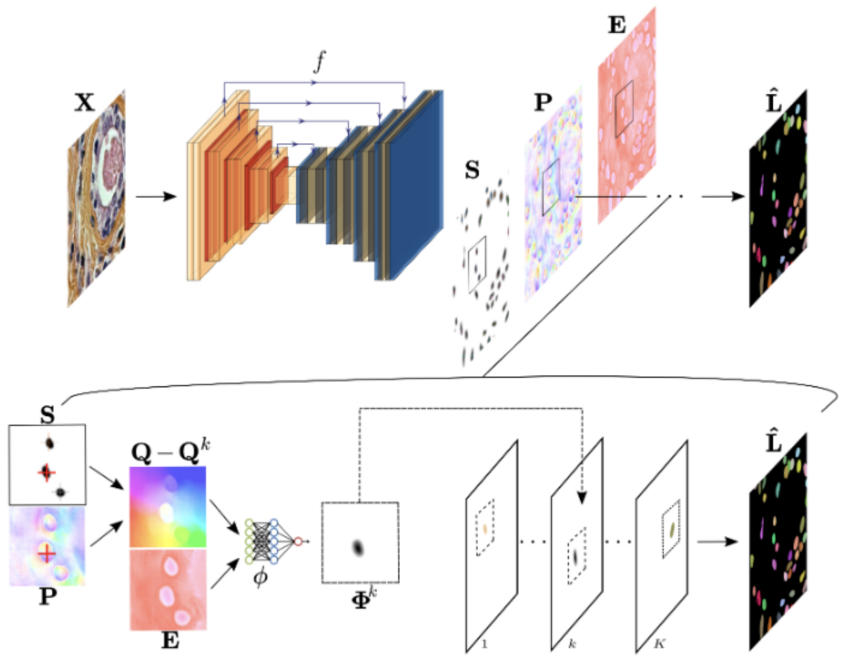

Figure 8: InstanSeg framework: segments both cell and nuclei in multiplexed microscopy and uses ChannelNet to learn informative three-channel representations without direct supervision (adapted from [17]).

InstanSeg overcomes these challenges by offering a fast, relatively accurate, and flexible deep learning-based pipeline for both cell and nuclear segmentation in fluorescence and brightfield microscopy[17]. Built in PyTorch, InstanSeg is optimized to handle highly multiplexed images (more than three channels) without requiring retraining or manual preprocessing, enabling researchers to analyze novel biomarker panels with minimal effort[17].

Its speed advantage stems from an efficient model architecture, integrated postprocessing via TorchScript, and full GPU acceleration. By compiling both the segmentation and postprocessing pipelines into TorchScript, InstanSeg supports seamless use in Python and can also be deployed independently via LibTorch, facilitating integration with tools like QuPath[17].

During training, segmentation losses are computed only for the labels available in the ground truth. Let $\hat{y}_{nucleus}$ and $\hat{y}_{cell}$ denote the predicted masks, and $y_{nucleus}$ and y_{cell} the corresponding ground truth masks. Binary Cross-Entropy (BCE) loss for each prediction, is applied conditionally as follows:

$$\mathcal{L}_{total} = \begin{cases} \text{BCE}(\hat{y}_{nucleus}, y_{nucleus}), & \text{if only nucleus labels exist} \\ \text{BCE}(\hat{y}_{cell}, y_{cell}), & \text{if only cell labels exist} \\ \text{BCE}(\hat{y}_{nucleus}, y_{nucleus}) + \text{BCE}(\hat{y}_{cell}, y_{cell}), & \text{if both labels exist} \end{cases}$$

This conditional training enables the model to learn effectively from partially labeled datasets.

This approach works with partially labeled datasets, without needing fully paired annotations. ChannelNet, trained jointly with InstanSeg, converts high-dimensional multiplexed inputs into a three-channel representation. It receives no separate loss; instead, it learns through the segmentation loss alone. This allows it to discover the most informative channel combinations for accurate segmentation, even if they don't correspond to specific biological markers.

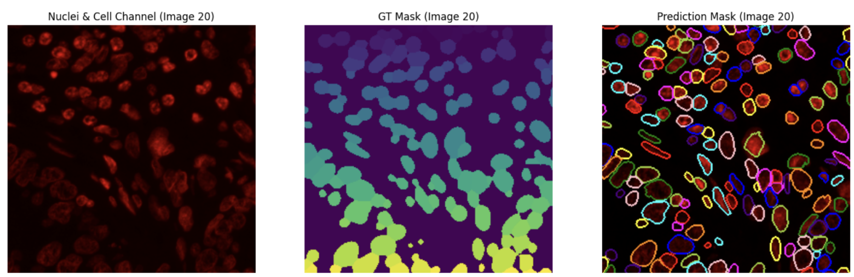

Figure 9: Instaseg input image with ground truth and predicted segmentation masks

2.2.4 Mesmer

Mesmer is a deep learning algorithm designed to segment cell nuclei and entire cells in tissue images. It utilizes a ResNet50 backbone integrated with a Feature Pyramid Network (FPN) to extract and process image features efficiently[9]. While FPNs are traditionally used in object detection [23], in this context, the FPN enhances segmentation performance by combining high-resolution, low-level features with low-resolution, high-level semantic features. This multi-scale representation enables the model to accurately delineate cells and nuclei of varying sizes and shapes, while preserving fine structural details such as boundaries and contours—crucial for precise instance segmentation in tissue images. The model has four prediction heads: two dedicated to nuclear segmentation and two for whole-cell segmentation.

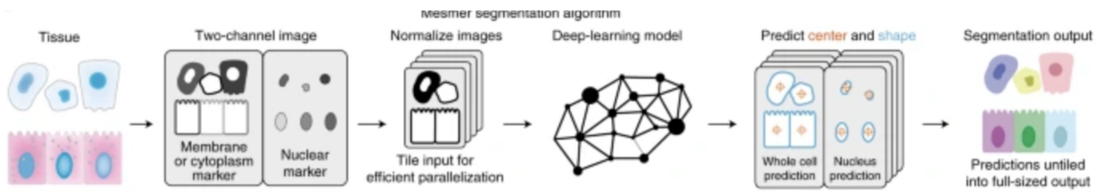

Figure 10: Mesmer Framework (Adopted from [9])

The model takes in a pair of images: a nuclear stain to identify nuclei and a membrane or cytoplasmic marker to outline whole cells. The images are then normalized and tiled for efficient processing [9, 24].The model then predicts the centers and boundaries of each nucleus and cell,the reconstructs the full-image predictions by untangling the tiled outputs[9]. Finally, a watershed algorithm is applied to the center and boundary maps to refine the segmentation masks for individual nuclei and whole cells [9].

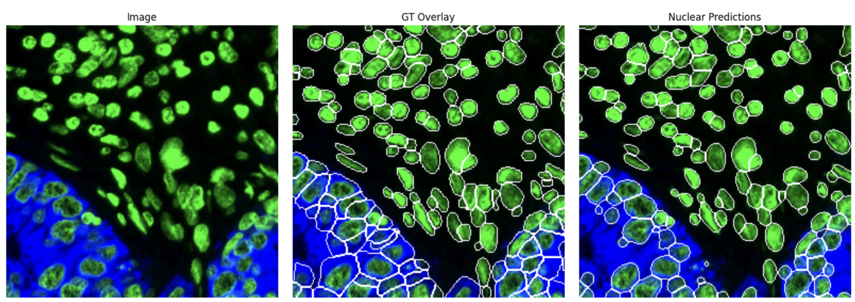

Figure 11: Mesmer input image with ground truth and predicted segmentation results

2.2.5 HoverNet

Hover-net is a deep-learning method that performs nuclear instance segmentation and classification simultaneously. It distinguishes clustered nuclei by using the horizontal and vertical distances of nuclear pixels to their centers of mass. Each segmented nucleus is then assigned a type through a dedicated classification process. HoverNet is trained on different datasets, each with its unique advantage: CoNSeP, PanNuke, MoNuSAC, Kumar, and CPM17. ConSeP, PanNuke, and MonuSAC are designed to handle both segmentation and classification, while Kumar and CPM17 are designed for segmentation only[25]. HoverNet model was only trained on Hematoxylin and Eosin(HE) images.In this study, we utilized the PanNuKe checkpoint weights trained on the PanNuke dataset and only used the segmentation head, PanNuke was preferred since it contains images from multiple organs (19 tissue types) and covers a broader range of pathological conditions [26]. In figure 12, the Hovernet model is shown,which comprises both encoder and decoder components and is designed to perform nuclear instance segmentation and classification through multi-task learning [26].

The Nuclear Pixel Segmentation (NP) branch distinguishes nuclear regions from the background using binary classification. In parallel, the Horizontal and Vertical Distance Maps (H, V) branch predicts the displacement of each nuclear pixel from the center of mass of its respective nucleus—this spatial encoding is critical for accurately separating clustered or overlapping nuclei.
To train these outputs jointly, HoVer-Net minimizes a composite loss function that combines pixel-wise Binary Cross-Entropy (BCE) for the NP output and Mean Squared Error (MSE) for the H and V maps. The total loss is defined by:

$$\mathcal{L}_{\text{total}} = \lambda_{np} \cdot \mathcal{L}_{\text{BCE}}^{\text{NP}} + \lambda_{hv} \cdot \mathcal{L}_{\text{MSE}}^{\text{H}} + \lambda_{hv} \cdot \mathcal{L}_{\text{MSE}}^{\text{V}} + \mathcal{L}_{\text{cls}} \tag{1}$$

where λ_{np} and λ_{hv} are weighting coefficients for the segmentation and displacement map losses, respectively, and $\mathcal{L}_{\text{cls}}$ is an optional classification loss applied when nuclear type classification is included. This multi-task loss encourages the network to learn spatial, morphological, and categorical features of nuclei simultaneously.

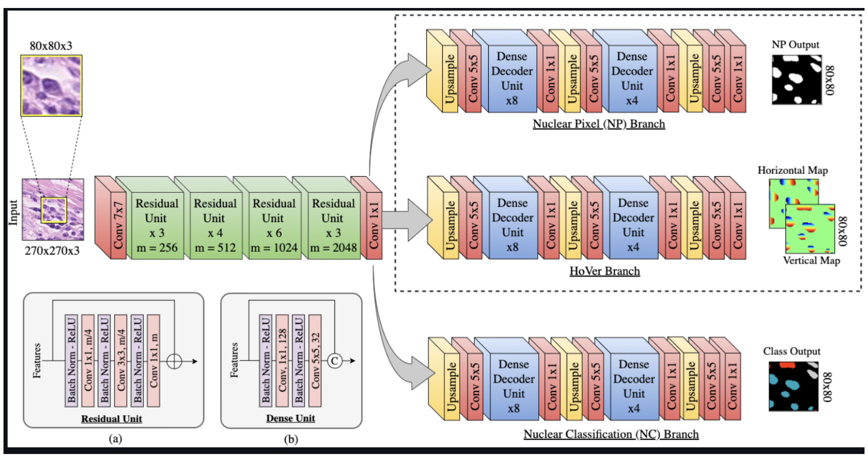

Figure 12: Hovernet framework (adapted from [25])

2.3 Model Evaluation

This subsection outlines the key metrics utilized in assessing the performance of the models presented in sections 2.2.1, 2.2.3,2.2.2,2.2.4 and 2.2.5. Initially, all models were benchmarked against annotated ground truth objects, as described in section 2. Subsequently, their performance was evaluated independently of ground truth objects.

2.3.1 Classical Segmentation Metrics

In segmentation tasks such as this, the loss function minimized is often binary cross-entropy (BCE), especially in binary segmentation where the objective is to distinguish foreground

from background pixels. BCE, however, does not measure the spatial quality of predictions; as such, during evaluation, region-based performance metrics are used to capture how well the predicted segmentation masks align with the ground truth.
In this study we concentrated on specific evaluation methods such as:

- **True Positives (TP)**: Foreground pixels correctly classified as foreground.
- **False Positives (FP)**: Background pixels incorrectly predicted as foreground.
- **False Negatives (FN)**: Foreground pixels incorrectly predicted as background.
- **Precision:**
$$\text{Precision} = \frac{TP}{TP + FP}$$
Precision measures the proportion of predicted foreground pixels that are indeed correct, and is especially important when false positives is costly[27].
- **Recall :**
$$\text{Recall} = \frac{TP}{TP + FN}$$
Recall assesses the proportion of actual foreground pixels that were successfully predicted[27]. High recall is critical when missing objects (false negatives) is more detrimental than false alarms.
- **F1 Score:**
$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$
The F1 score balances precision and recall, providing a single metric to evaluate models, particularly when dealing with class imbalance.
- **Intersection over Union (IoU)**:
$$\text{IoU} = \frac{TP}{TP + FP + FN}$$

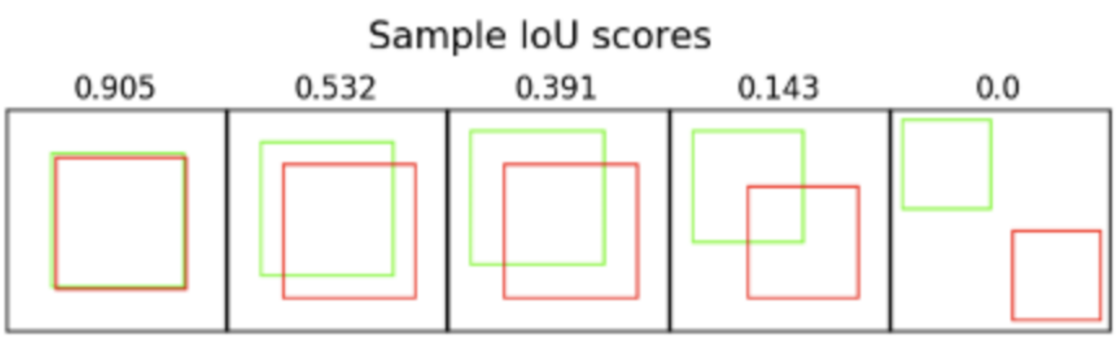

Figure 13: A score of 1 means that the predicted segments precisely matches the ground truth. A score of 0 implies that the predicted and true segments do not overlap at all.

Intersection over Union (IoU) measures the overlap between predicted and ground truth masks, penalizing false positives and false negatives [28]. We selected IoU over metrics such as the Dice score because it provides a more stringent evaluation of segmentation quality. We aimed to identify an optimal model with strong performance across varying overlap thresholds. To this end, we examined the relationship between F1-score, Precision, Recall, False Positives (FP), False Negatives (FN), and True Positives (TP) across IoU thresholds ranging from 0.1 to 0.9.

2.3.2 Multidimensional Scaling

Multidimensional Scaling(MDS) is a dimensionality reduction technique used to visualize the similarity or dissimilarity between high-dimensional data points.
Given a squared distance matrix $D_X \in \mathbb{R}^{n\times n}$, which represents the dissimilarities among n observations, MDS aims to find a configuration of points in a k-dimensional space (where $k \ll n$) such that the distances between the points in this space closely reflect the structure of D_X.
MDS first centers the data using the centering matrix

$$H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T,$$

I is $n \times n$ identity matrix, n is number of data points, and $\mathbf{1}$ is an n-dimensional vector of ones. The centered Gram matrix G_X, which contains the inner products between points, is computed from the squared distance matrix $D_X \in \mathbb{R}^{n\times n}$ as

$$G_X = -\frac{1}{2}HD_XH.$$

D_X consists of squared pairwise distances d_{ij}^2 between data points i and j. The eigendecomposition of G_X .

$$G_X = U\Lambda U^T,$$

$U \in \mathbb{R}^{n\times n}$ is a matrix whose columns are the eigenvectors of G_X, and $\Lambda \in \mathbb{R}^{n\times n}$ is a diagonal matrix of corresponding eigenvalues.
The low-dimensional embedding $Z \in \mathbb{R}^{n\times k}$ is obtained by selecting the top k eigenvectors $U_k \in \mathbb{R}^{n\times k}$ and their eigenvalues $\Lambda_k \in \mathbb{R}^{k\times k}$, where k is the target embedding dimension:

$$Z = U_k\Lambda_k^{1/2}.$$

Each row of Z represents the coordinates of a data point in the reduced k-dimensional space, preserving the original pairwise distances as closely as possible.
In this study, MDS was applied to compare segmentation models by analyzing the dissimilarities in predicted cell areas and cell counts. This enabled a qualitative assessment of how closely related different models are in terms of their segmentation behavior. The method is

used to reduce the number of model-pair comparisons by focusing on models that exhibit similar segmentation characteristics, as well as on those that show marked dissimilarities. We calculated per-image foreground segmentation areas and cell counts for each of the four segmentation models— 2.2.1,2.2.2,2.2.3, and 2.2.4. Using these metrics, a dissimilarity matrix was constructed employing the correlation distance metric, emphasizing agreement in trends and variation between models rather than absolute difference

2.3.3 Pair -Wise Bland-Altman plots

As a successor to the MDS method introduced in Section 2.3.2, the Bland–Altman plot was used to assess the magnitude of agreement or disagreement between pairs of segmentation methods [10]. This method refines the analysis by focusing not on association or predictive correlation but on how closely the two methods agree in their outputs.
Given paired measurements A_i and B_i from two segmentation models on the same instance, the Bland–Altman method begins by computing the mean and difference for each pair:

$$\text{Mean}_i = \frac{A_i + B_i}{2}, \qquad \text{Difference}_i = A_i - B_i.$$

The average of the differences across all n samples yields the *bias* between the two methods:

$$\text{Bias} = \frac{1}{n}\sum_{i=1}^{n}(A_i - B_i).$$

To understand the variability, the standard deviation (SD) of these differences is calculated as:

$$\text{SD} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left((A_i - B_i) - \text{Bias}\right)^2}.$$

Using this, the 95% limits of agreement are defined as:

$$\text{Lower Limit} = \text{Bias} - 1.96 \times \text{SD}, \qquad \text{Upper Limit} = \text{Bias} + 1.96 \times \text{SD}.$$

These limits and the bias line are plotted to produce the Bland–Altman diagram, which visualizes the differences on the vertical axis against the average values on the horizontal axis. This visual and statistical interpretation helps detect systematic biases and the extent of agreement or disagreement between segmentation models [10, 29].

2.3.4 HiStauGAN- Sensitivity To Stain Variation Analysis

Histopathology, especially H&E images, predominantly suffers from variations in stain intensity and application resulting from laboratory procedures, different scanners, and staining variations[30, 31]. One way of combating this is standard normalization; however, it may oversimplify this variation or distort structural information[32]. While deep learning models are generally expected to be robust to variations in H&E staining due to their ability to learn complex data distributions, their performance may still depend on the training data's diversity and the model's architectural focus. Architectural designs that emphasize structural features over color information may improve robustness to stain variation but could slightly compromise overall generalizability across diverse staining styles. It is for this reason that we implemented HiStauGAN—generative adversarial networks designed to disentangle style (stain/color) from content (structure) using two separate encoders while maintaining the tissue's morphological features[33].The model was trained on the CAMELYON17 dataset, which consists of 1,000 whole-slide images (WSIs) of sentinel lymph node biopsies collected from five distinct domains (medical centers). The dataset is split into train and test sets, each containing data from 100 patients—20 from each center—with five WSIs per patient. This study generated domain-specific synthetic images from each original image, resulting in five variations per input, each corresponding to a different domain. These synthetic images were subsequently evaluated using four instance segmentation models:2.2.1, 2.2.3,2.2.2 and 2.2.5. Figure 14 shows synthetic variations of the original image generated by HiStauGAN from the MoNuSeg dataset.

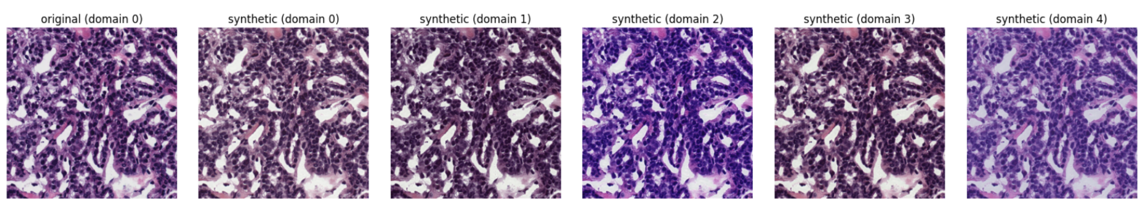

Figure 14: Original and domain-specific images generated by HiStauGAN.

In order to assess if there are differences between the domains, we quantitatively assessed the similarity of stain styles across different histopathological image domains using Macenko's method for stain matrix extraction combined with cosine similarity analysis[34, 35]. For each image within a domain, the Macenko algorithm was applied to extract a 2×3 matrix representing the hematoxylin and eosin (H&E) stain basis vectors. These matrices were then flattened into 6-dimensional stain vectors, and all vectors within a domain were averaged to produce a mean stain representation per domain. Cosine similarity was computed between all pairs of mean vectors to compare these domain-level stain profiles. We chose Cosine similarity because it measures vectors' orientation (relative composition) independently of their magnitude, making it robust to variations in image brightness and staining [35]. The method helps directly compare stain composition across domains, providing insight into whether different institutions or synthetic augmentation processes introduce significant staining variations. We thus evaluated IoU metrics on the most dissimilar domains (UMCU and CWZ) to check if there are differences in model performance.

3 Results

This section is divided into three key subsections: Section 3.1 presents results on fluorescence images using the TissueNet dataset; Section 3.2 focuses on Hematoxylin and Eosin stained images from the MoNuSeg dataset; and Section 3.3 explores sensitivity analysis using StainGANs.

3.1 Fluorescence Image Analysis

3.1.1 Analysis with Ground Truth Masks

Figure 15 shows how each segmentation model—, StarDist 2.2.1, Cellpose 2.2.2,InstaSeg 2.2.3, and Mesmer 2.2.4—compares to the ground truth labeled masks. The evaluation was performed using various metrics across a range of Intersection over Union (IoU) thresholds, from 0.1 to 0.9.
StarDist demonstrated the most robust performance overall, keeping the highest precision and F1 scores across nearly all thresholds, along with strong recall and the fewest false positives. Cellpose closely follows, performing well in both recall and F1 score, though it slightly trails StarDist in precision and panoptic quality. In contrast, Mesmer and InstaSeg exhibit weaker performance, with noticeably lower precision and recall that diminish further for stricter IoU thresholds. Furthermore, these two models reported a higher number of false positives, particularly Mesmer, as IoU thresholds increase.

It was also noted that across all these models the number of true positives decreased with increasing IoU thresholds, reflecting the greater stringency in match criteria. In contrast, false positives tend to increase under these conditions, illustrating the challenge of maintaining specificity at higher overlap requirements. In general, StarDist and Cellpose show superior balance in precision and recall, making them more reliable for nuclei segmentation tasks under varying overlap tolerances, whereas Mesmer and InstaSeg may require further tuning or refinement for comparable effectiveness.

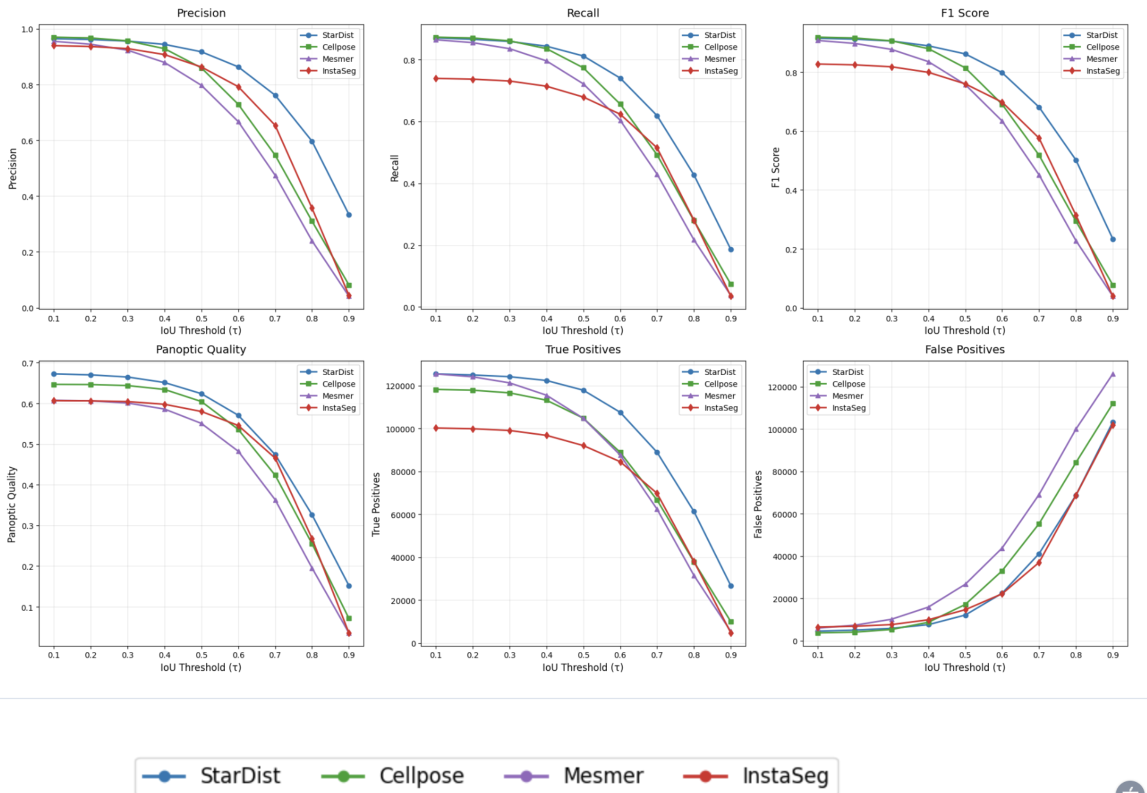

Figure 15: Performance of segmentation models on fluoresence images across varying IoU thresholds based nn Ground truth.

3.1.2 Multidimensional Scaling Analysis

Figure 16 presents the results of multidimensional scaling, applied to assess similarity between segmentation methods based on their per-image nucleus count and segmentation area distributions. The MDS plots summarize the pairwise correlation distances between methods, providing insight into how similarly each model behaves across the dataset in terms of area and count variability.

In the left panel (MDS based on the segmentation area), Cellpose is positioned farthest from the other methods. It indicates that its segmentation area patterns across the images are least correlated, likely reflecting a distinct or inconsistent trend. StarDist and InstaSeg appear close together, suggesting that they follow a similar pattern in area variation across images (both tend to detect larger segmented regions in the same images). Mesmer lies at an intermediate distance, implying that its area variation pattern is partially similar but not strongly correlated with the others.

In the right panel (MDS based on nucleus count), StarDist and Mesmer are closely positioned, showing strong agreement in how their nucleus counts vary across images; both tend to detect more nuclei in denser images and fewer in sparser ones. In contrast, InstaSeg and Cellpose are situated further apart, reflecting distinct count variation patterns that differ from each other and those of StarDist and Mesmer. Notably, Cellpose again exhibits a unique variation profile, highlighting its divergence in consistency relative to the different segmentation methods.

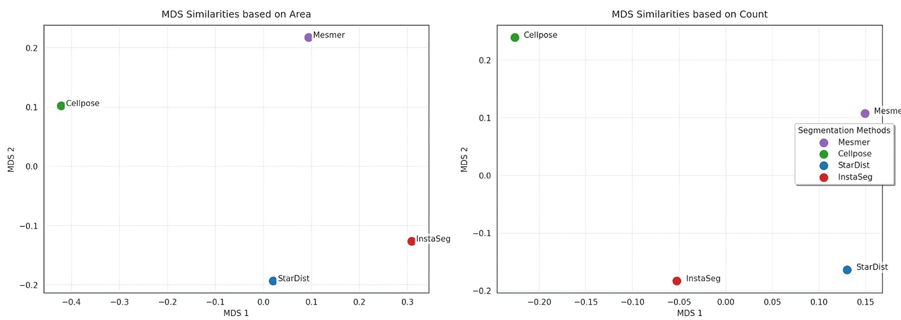

Figure 16: Multidimensional Scaling

3.1.3 Pair -Wise Bland-Altman plots Analysis

Figure 17 and 18 presents pairwise Bland–Altman comparisons for model combinations described in Section 2.3.2, focusing on total segmentation area (in pixels) and nuclei count per image.

Table 1 summarises results from the Bland-Altman plots. Across all comparisons, the presence of proportional bias—where differences between methods grow larger as the average segmented area or nuclei count increases—indicates that segmentation discrepancies are not constant but scale with image density and area, suggesting that models may behave similarly on simpler or smaller images but diverge significantly on more complex or dense samples. Additionally, funnel-shaped heteroscedasticity in the Bland–Altman plots reflects a pattern where variance increases with image complexity. In practical terms, this means that segmentation performance becomes less reliable as the structural density of tissue increases—likely due to challenges such as overlapping nuclei, irregular shapes, and poor contrast. These patterns highlight that segmentation consistency deteriorates under more demanding conditions, with each algorithm reacting differently based on its design. For example, StarDist, which assumes star-convex shapes, may underperform in highly clustered or irregular morphologies. At the same time, Cellpose, which uses a vector flow approach, may over-segment to capture all spatial gradients. Models like Mesmer and InstaSeg may also err on conservative delineation, missing finer structures or densely packed nuclei.

Table 1: Bland–Altman analysis comparing segmentation methods on total area (pixels) and nuclei count for fluorescence images

Comparison	**Metric**	**Mean Difference**	**SD (Variability)**	**Interpretation**
Cellpose vs StarDist	Total Area	−10,101.10 px	15,874.85	Cellpose segments larger regions; strong heteroscedasticity in complex images
	Nuclei Count	−8.64 nuclei	57.57	StarDist undercounts in dense regions images
InstaSeg vs StarDist	Total Area	−4,124.37 px	8,114.22	InstaSeg under-segments; variance increases with object size
	Nuclei Count	−2.90 nuclei	∼ 75−100 range	Small average bias; inconsistent in dense areas
StarDist vs Mesmer	Total Area	−6,765.30 px	10,625.13	Mesmer segments more conservatively; increasing disagreement in large areas
	Nuclei Count	−15.96 nuclei	46.41	Mesmer detects fewer nuclei; heteroscedasticity in crowded images

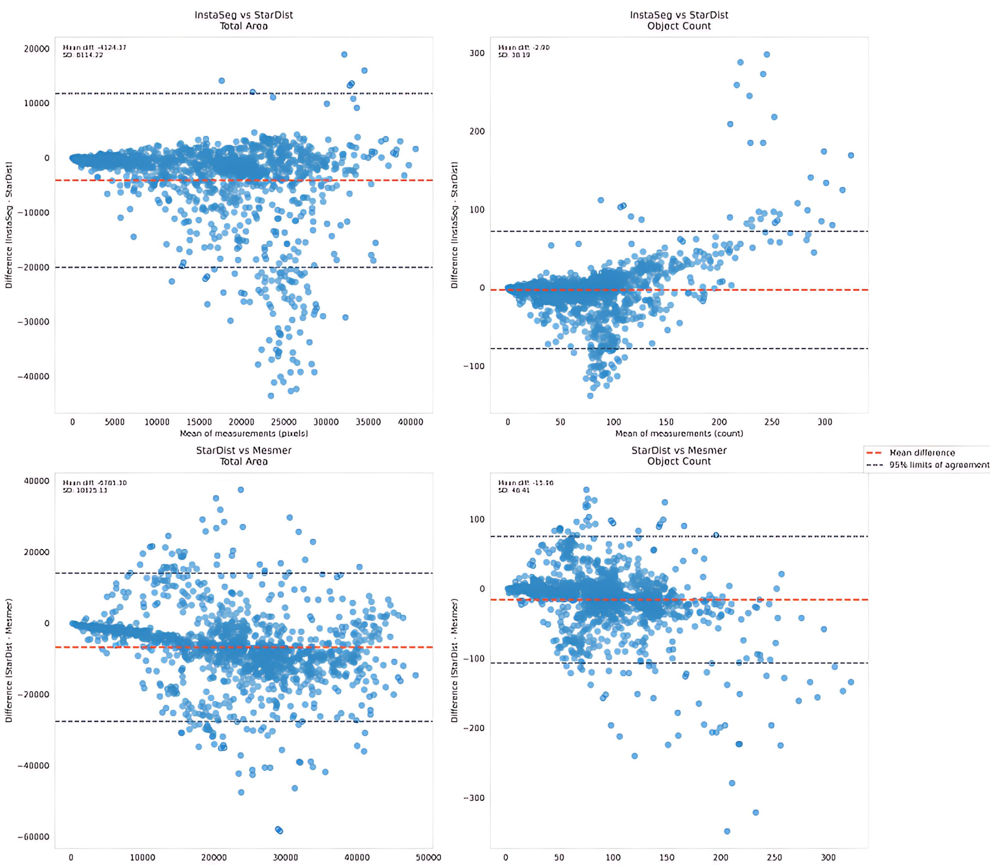

Figure 17: Bland–Altman plots comparing InstaSeg,StarDist and Mesmer for segmented area and nuclei count.

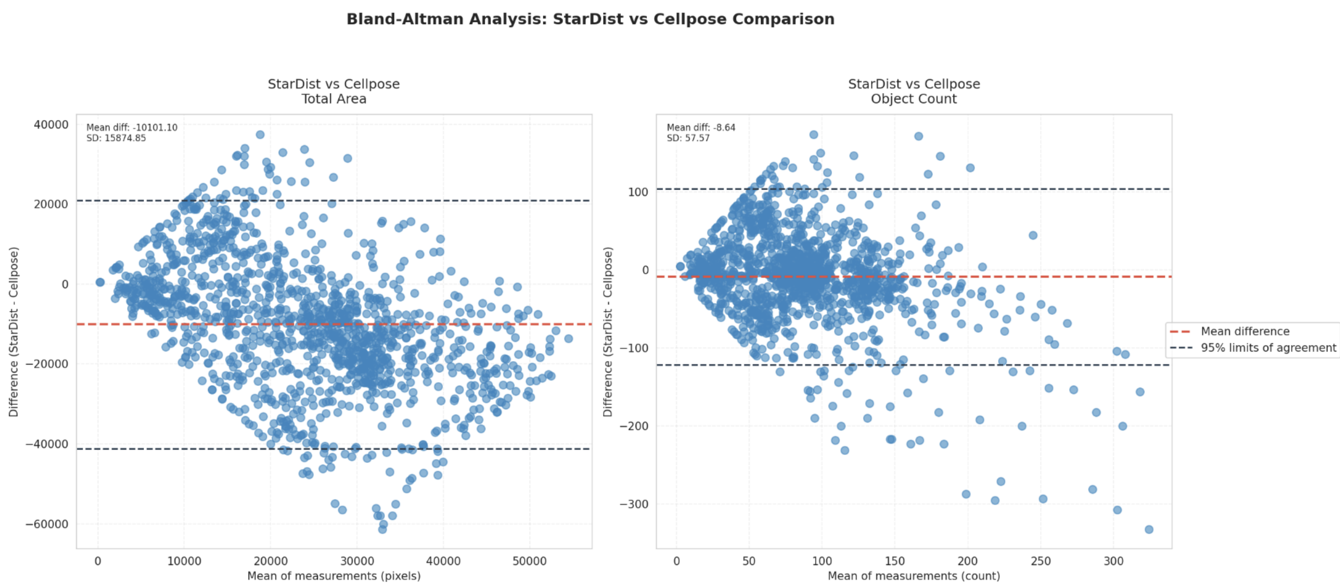

Figure 18: Bland–Altman plots comparing Stardist vs Cellpose Comparison for segmented area and nuclei count.

3.2 Analysis Of Hematoxylin And Eosin Images

In this section, the focus of the analysis shifts to H&E-stained images, with primary benchmarking conducted on the MoNuSeg dataset. Among the five methods, only four—StarDist, Cellpose, InstaSeg, and HoverNet—provide pretrained models specifically tailored for H&E-stained images. Consequently, the analysis in this section primarily concentrates on these three methods. We used all the methods in Sections 2.3.1, 2.3.2 and 2.3.3 to analyze H&E-stained images.

3.2.1 Analysis With Ground Truth Masks

The models were evaluated using key metrics: Recall, Precision, True Positives, True Negatives, False Positives, and IoU. A summary of the results is presented in Figure 19. The recall analysis indicated higher sensitivity for Cellpose, maintaining values above 0.9 up to $\tau = 0.5$ and gradually decreasing thereafter. HoverNet closely follows this trend, though slightly lower in magnitude. Both models slightly outperform StarDist, but all three perform better than InstaSeg, which shows a rapid decline in recall, indicating a higher tendency to miss ground truth instances as τ increases. StarDist, Cellpose, and HoverNet consistently achieved the highest F1 scores across IoU thresholds, reflecting a strong balance between precision and recall. In contrast, InstaSeg demonstrated limited effectiveness, with its F1 score dropping sharply at stricter thresholds ($\tau > 0.5$). In terms of accuracy, StarDist, Cellpose, and HoverNet performed comparably well, reaching values up to 80% at lower IoU thresholds and maintaining accuracy above 65% around $\tau = 0.5$. InstaSeg, however, exhibited significantly lower accuracy across the entire range, indicating less reliable segmentation performance.

True positive counts were highest for Cellpose ,StarDist and HoverNet, with both models achieving peak performance at $\tau = 0.1$ and maintaining gradual declines as τ for stricter IoU.InstaSeg showed significantly lower TP counts, especially beyond $\tau = 0.5$.

False positive analysis confirmed that InstaSeg consistently produces the highest number of spurious detections, which aligns with its poor precision. In contrast, StarDist, Cellpose, and HoverNet maintain low FP counts, indicating strong discriminatory capability.

False negatives are lowest for Cellpose. Both StarDist and Hovernet produce almost similar results, followed by StarDist, corroborating their superior recall. InstaSeg recorded notably higher FN counts, with increasing severity at higher IoU thresholds.

Overall, each model showed strengths in different areas: StarDist excels in precision and minimizing false detections, Cellpose offers high recall , while HoverNet strikes a strong balance between the two, making it a well-rounded performer across varying IoU thresholds. In contrast, InstaSeg consistently underperforms across all evaluated metrics—precision, recall, F1 score, accuracy, and error rates—indicating limited reliability for accurate instance segmentation under stricter matching criteria.

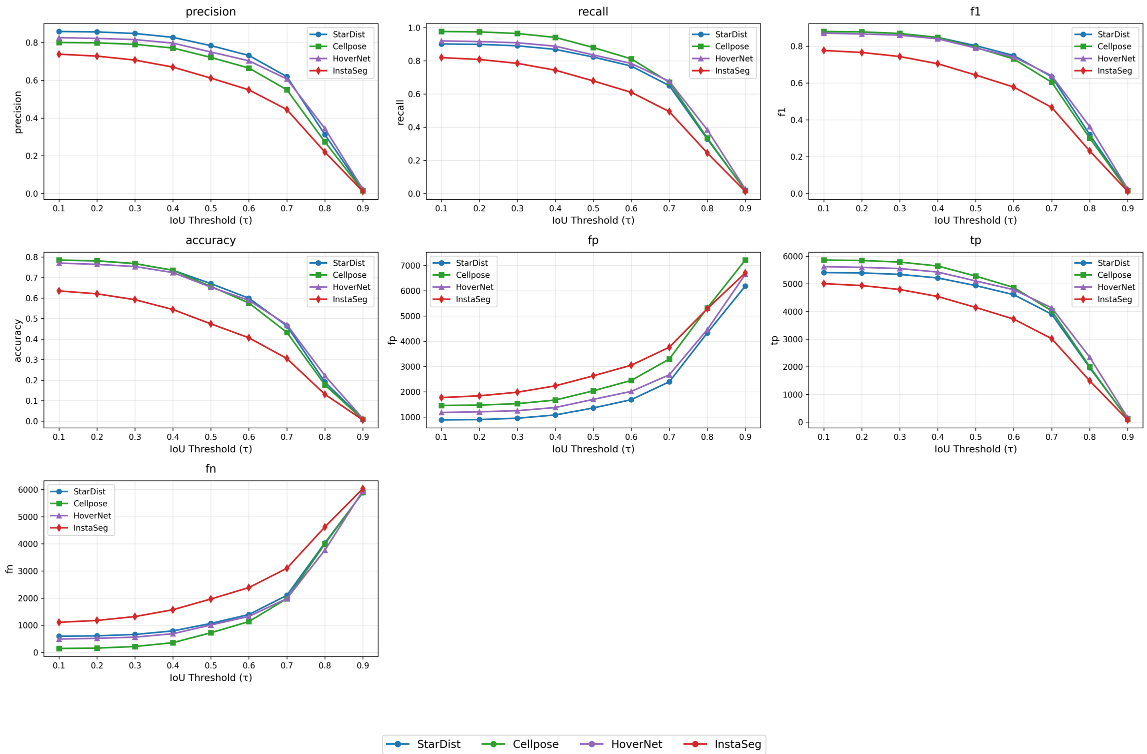

Figure 19: Performance of segmentation models on H&E-stained images across varying IoU thresholds based on ground truth.

3.2.2 Multidimensional Scaling Analysis

Multidimensional analysis was conducted based on the total segmented area and nuclei count to gain deeper insights into the dissimilarities between segmentation methods. The resulting 2D embeddings are illustrated in Figure 20.

The left panel of Figure 20 indicates that HoverNet exhibits the most remarkable dissimilarity in terms of total segmented area, being spatially distant from the cluster formed by Cellpose, StarDist, and InstaSeg—an indication that HoverNet's segmentation behavior differs significantly from the rest. Cellpose and StarDist are positioned closely together, implying close similarity in their area measurements, whereas InstaSeg lies slightly apart but still within reasonable proximity.

The right panel shows MDS based on the number of segmented nuclei.HoverNet is again an outlier, suggesting it segments a drastically different number of nuclei compared to the other methods. Cellpose, StarDist, and InstaSeg cluster tightly, indicating high agreement in nuclei count estimation, with minimal variation between them.

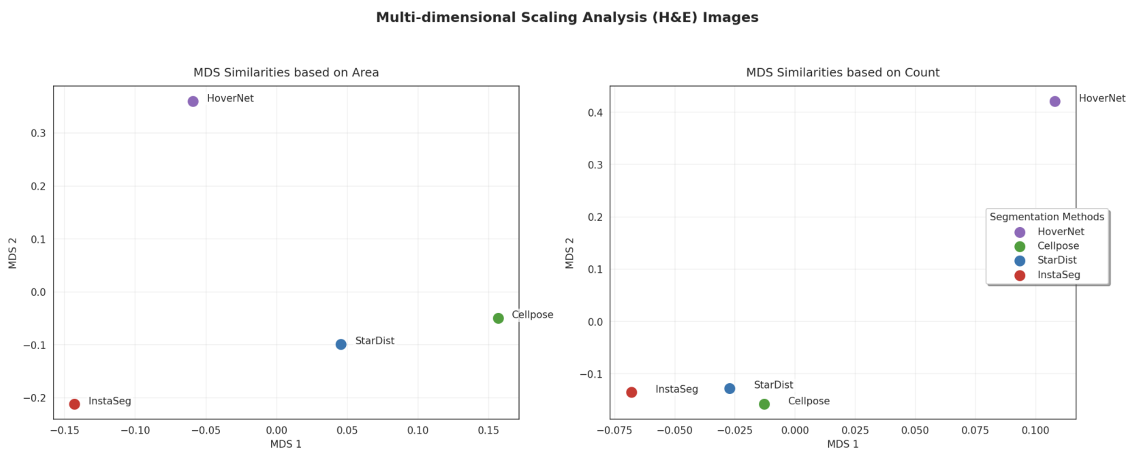

Figure 20: Model comparison with MDS

3.2.3 Pair-Wise Bland-Altman Plots Analysis

Pair-Wise Bland– Altman analyses further compared Total segmentation area and nuclei count to evaluate the magnitude of agreement or disagreement between models. Figure 21 summarizes the analysis results. Table 2 provides a summary of mean difference, limits of agreements, nuclei count and standard deviation between model pairs.

Table 2: Summary of Bland–Altman analysis comparing segmentation methods on total area (pixels) and nuclei count.

Comparison	**Metric**	**Mean Difference**	**Agreement / SD**	**Observation**
Cellpose vs StarDist	Total Area	+13,230.50 px	±45,582.18 px	Cellpose predicts larger areas
	Nuclei Count	+73.00 nuclei	SD = 31.27	Higher counts, tight agreement
StarDist vs HoverNet	Total Area	−25,756.29 px	±90,454.62 px	High variability in area
	Nuclei Count	−36.07 nuclei	SD = 149.82	Substantial inconsistency
InstaSeg vs Cellpose	Total Area	−63,941.64 px	SD = 43,429.47	Large underestimation by InstaSeg
	Nuclei Count	−38.64 nuclei	SD = 79.85	Moderate variability

Cellpose consistently produced the largest segmentation areas and nuclei counts, while InstaSeg yielded the smallest, indicating differing segmentation behavior. Cellpose and StarDist showed the closest agreement in nuclei count, suggesting strong consistency. Comparisons involving HoverNet and InstaSeg showed the highest variability, reflecting differences in segmentation precision across models.This observation synchronizes with results from 20.

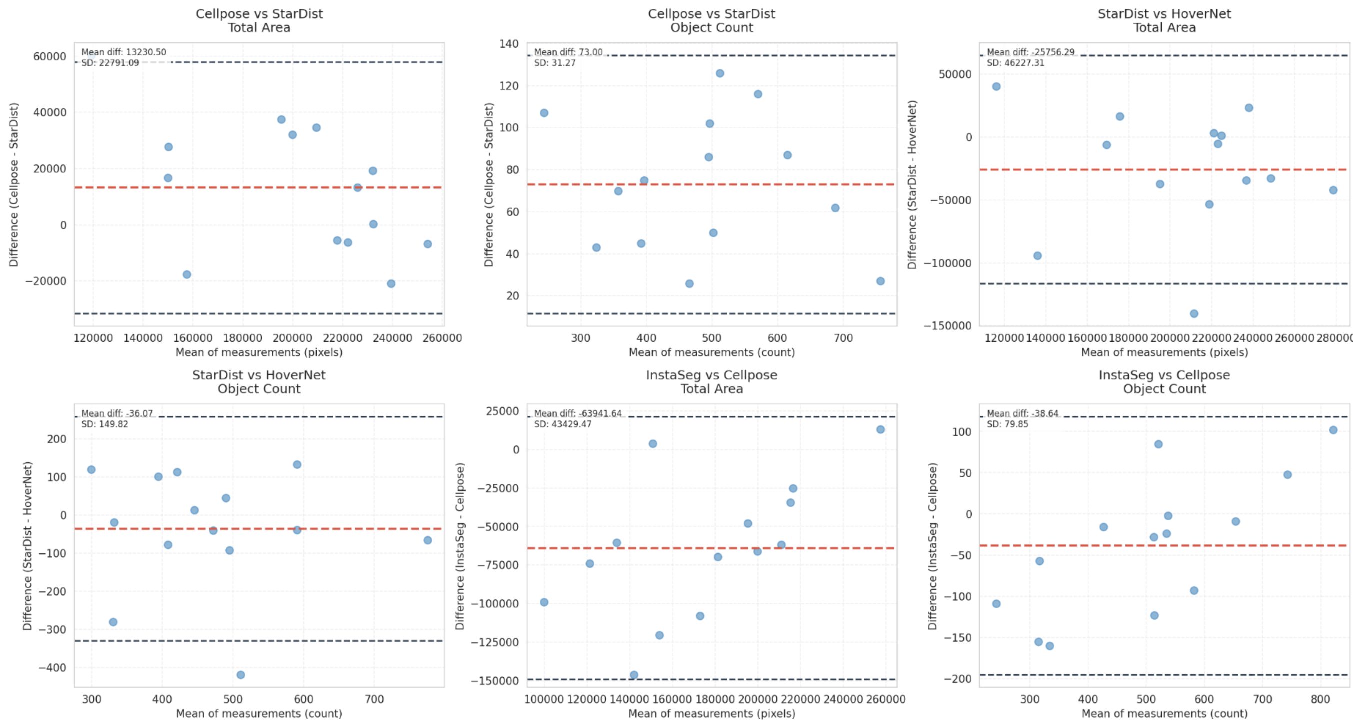

Figure 21: Bland-Altman plots

3.3 Sensitivity Analysis

Figure 22 presents a heatmap of cosine similarities between stain vectors of HistauGAN-generated images from five domains introduced in section3.3: Radboud, CWZ, UMCU, Rijnstate, and Oost-Nederland. Cosine similarity values close to 1.0 indicate very high similarity in stain composition. Most domain pairs have similarities above 0.98, suggesting that HistauGAN-generated outputs are largely consistent across domains regarding staining. However, UMCU exhibits slightly lower similarity scores (0.981 with CWZ, 0.983 with Radboud, and 0.987 with Rijnstate), indicating it may have more distinct staining characteristics. In contrast, Rijnstate and Radboud share one of the highest off-diagonal similarities (0.999), suggesting nearly identical stain styles. While there is high inter-domain consistency in generated stain appearances, subtle differences remain, particularly involving UMCU.

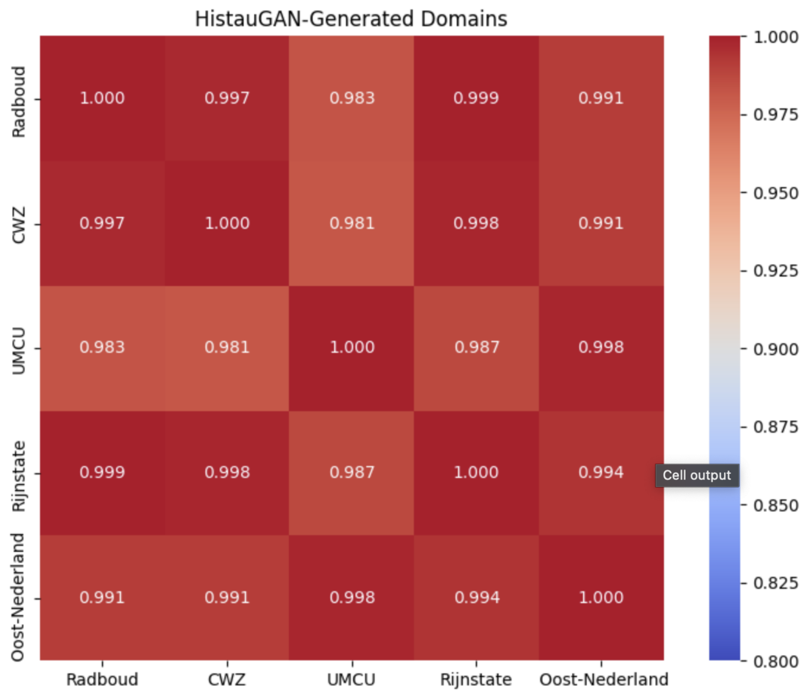

Figure 22: Cosine similarity heatmap of HistauGAN stain vectors across five domains

Figures 24 ,23 and Table 3show results for sensitivity analysis conducted to determine how variability in stains affects four segmentation models—StarDist, Cellpose, HoverNet, and InstaSeg—across CWZ and UMCU domains, selected based on their relative divergence in the HistauGAN-generated domain similarity heatmap Figure 22. At an IoU threshold of 0.5, StarDist achieved the best overall performance with high precision (0.74), recall (0.96), and F1-score (0.83), maintaining low false positives (2400) and false negatives (1200), suggesting strong robustness to domain shifts. Cellpose and HoverNet followed closely with slightly lower precision (0.72 and 0.70) and F1-scores (0.82 and 0.80), exhibiting moderate domain sensitivity. InstaSeg, however, performed poorly across both domains with a precision of 0.58, F1-score of 0.70, and the highest FP (4700) and FN (2300), indicating substantial degradation in heterogeneous domains like UMCU. These indicate StarDist's suitability for cross-domain generalization, while InstaSeg may require domain-specific fine-tuning to improve reliability across clinical sites.

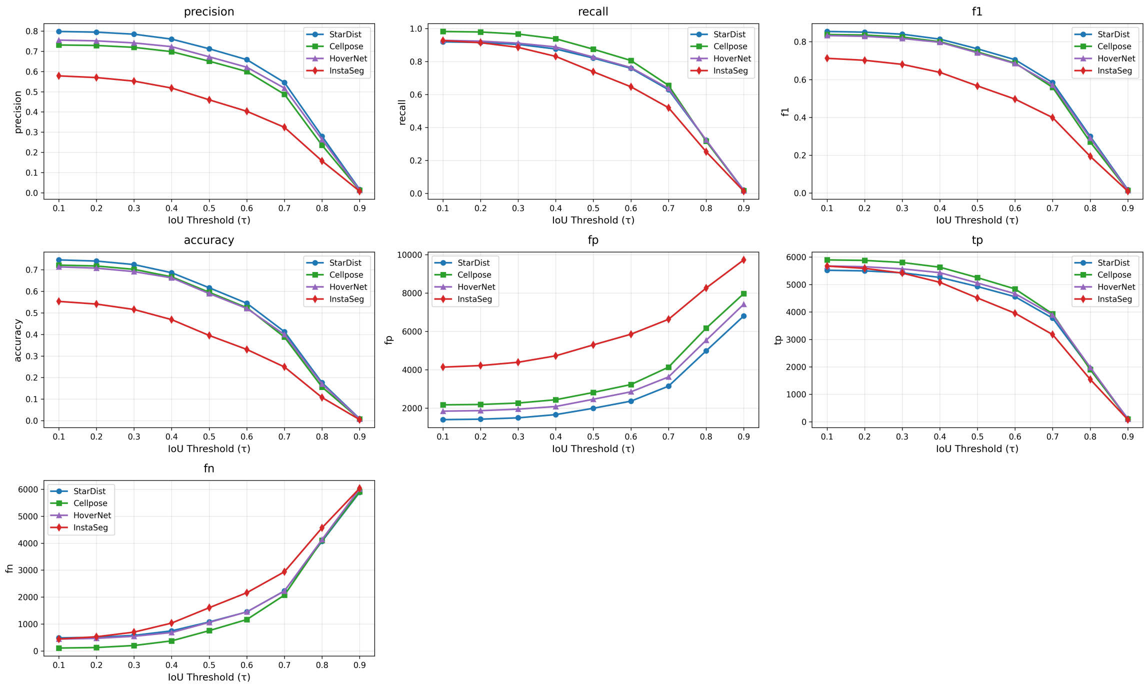

Figure 23: Performance of segmentation models on the CWZ domain across varying IoU thresholds

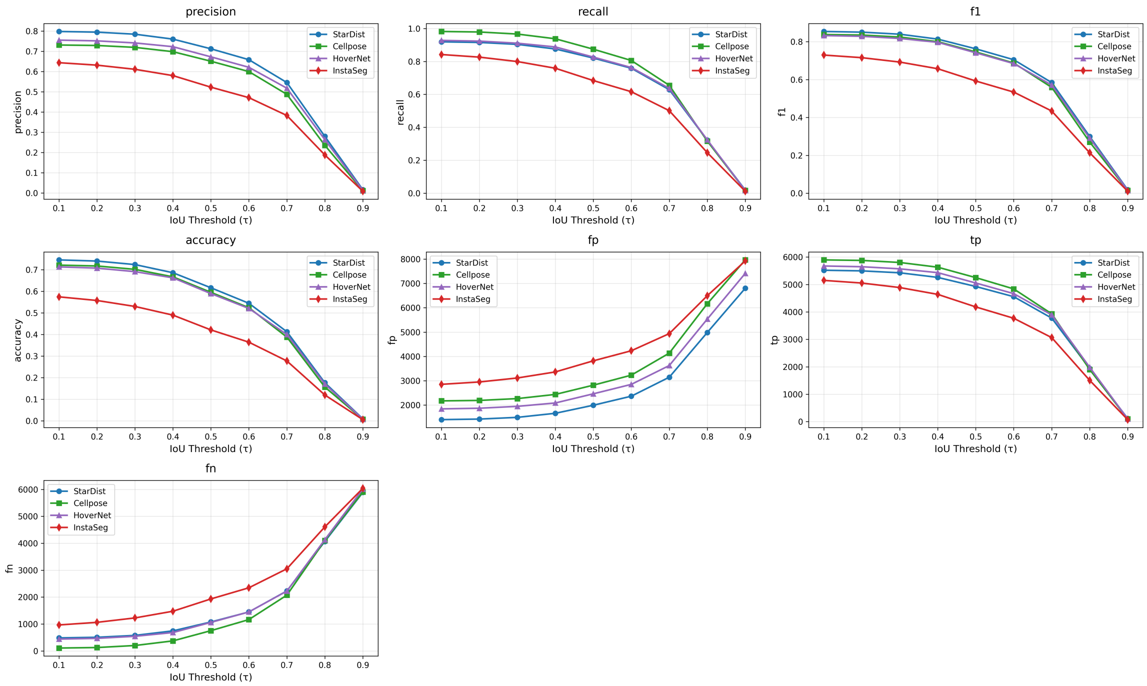

Figure 24: Performance of segmentation models on the UMCU domain across varying IoU thresholds

3.3.1 Pair-Wise Bland-Altman Plots Sensitivity Analysis (CWZ vs UMCU Domain)

Table 3: Summary of Bland–Altman analysis comparing segmentation models across CWZ and UMCU domains.

Comparison	Metric	Mean Difference	Agreement / SD	Observation
Cellpose vs StarDist	Total Area	−1660 px	±7692 px	Small diff in CWZ; positive bias in UMCU
	Nuclei Count	+81 nuclei	SD = 57	Cellpose detects more Nuclei in both
StarDist vs HoverNet	Total Area	−29854 px	±61945 px	StarDist underperforms more in UMCU
	Nuclei Count	−43 nuclei	SD = 78	StarDist detects fewer Nuclei, worse in UMCU
InstaSeg vs Cellpose	Total Area	−1695 px	SD = 11048 px	Smaller disagreement in CWZ
	Nuclei Count	+126 nuclei	SD = 149	InstaSeg detects more Nuclei, especially in UMCU

From figures 25,26, and Table 3, across both domains, a consistent trend emerged: Cellpose detects more nuclei than StarDist, though the magnitude of this difference is slightly smaller in UMCU compared to CWZ. Similarly, InstaSeg consistently identifies more nuclei than Cellpose, with a slightly greater difference in UMCU. Key disparities appear in total area estimations—while Cellpose and StarDist show relatively minor differences in CWZ, Cellpose predicts significantly larger areas in UMCU. StarDist notably underestimates the segmentation area in both domains, but the extent is more pronounced in UMCU. HoverNet keeps a more stable detection pattern relative to StarDist, but StarDist's performance deteriorates significantly in UMCU, suggesting domain-specific variability. Nuclei count consistently ranks across domains, while total area exhibits greater domain-dependent fluctuations.

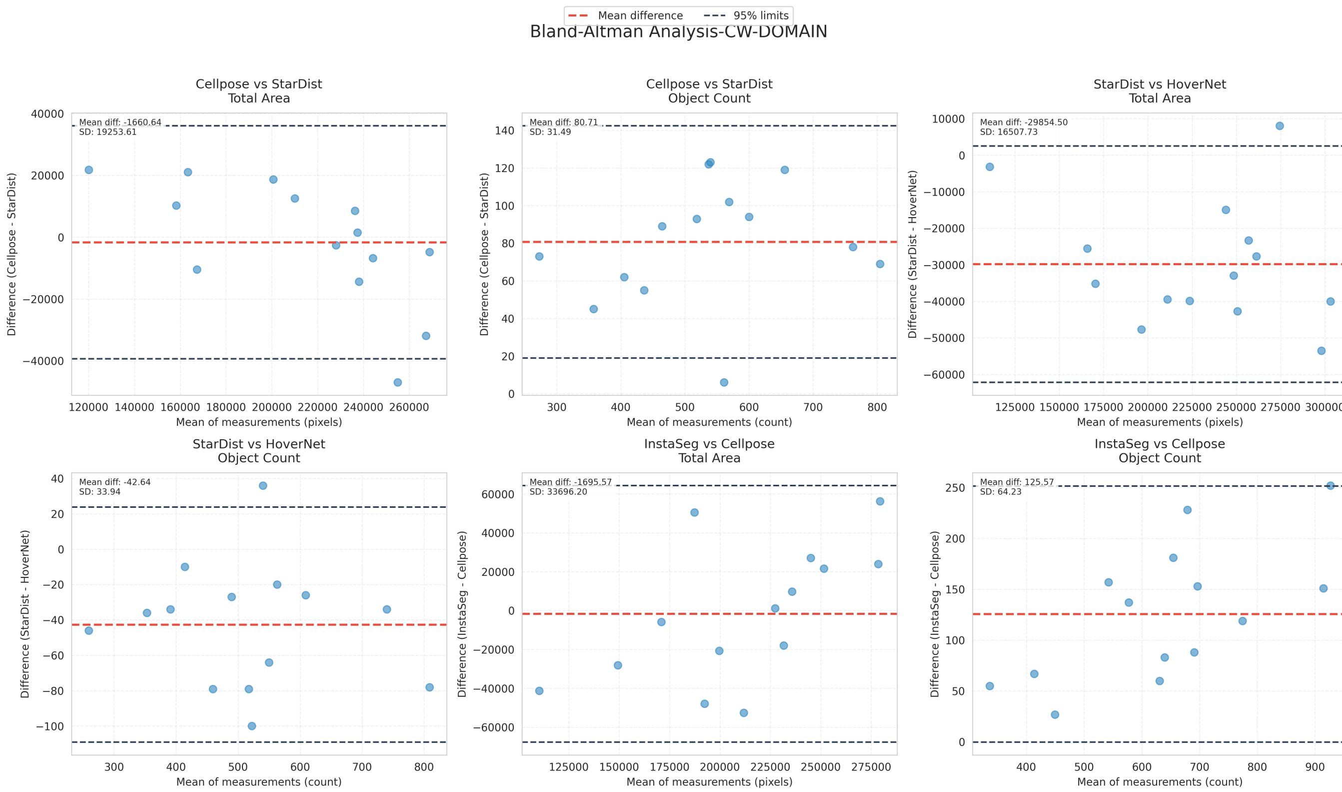

Figure 25: Comparative analysis of segmentation area and detected Nuclei count in the CWZ domain across multiple models

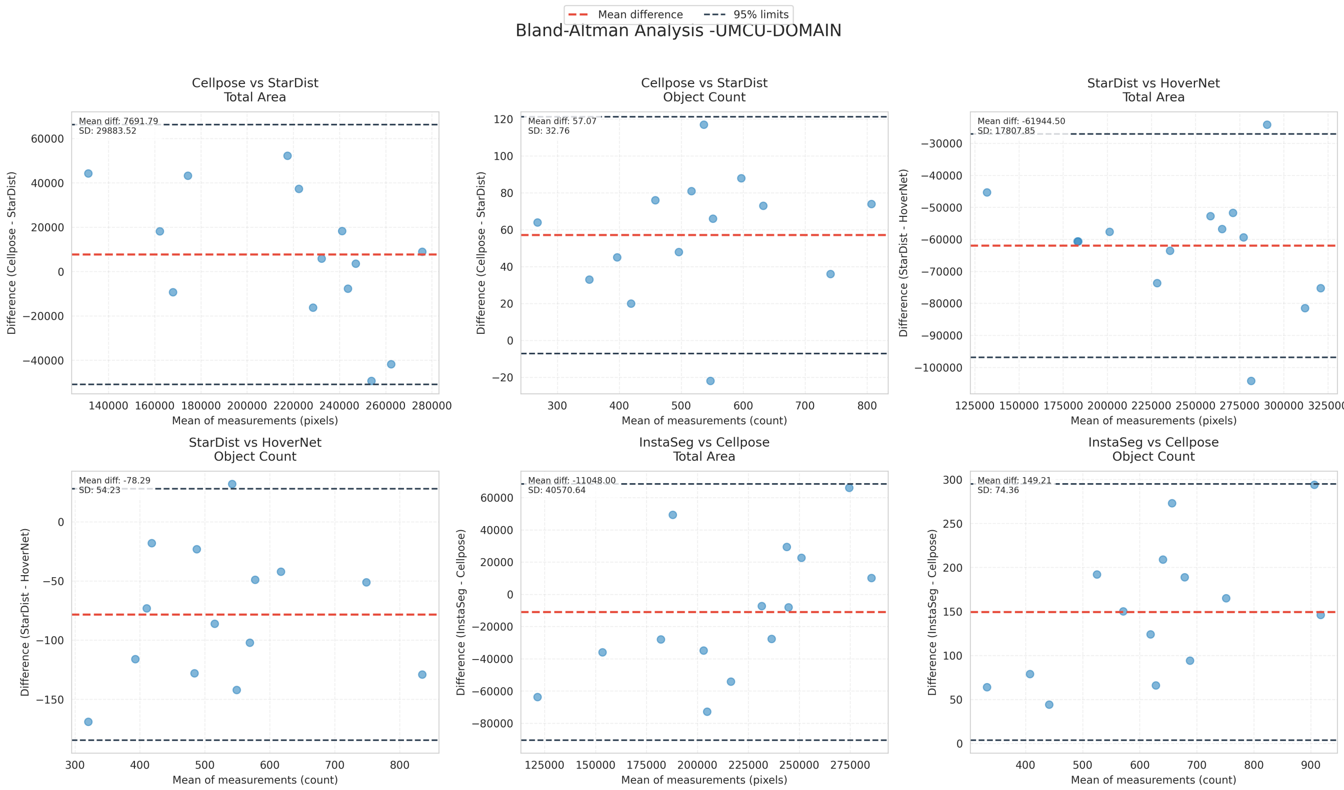

Figure 26: Comparative analysis of segmentation area and detected cell count in the UMCU domain across multiple models.

4 Discussion

Previous comparative studies of segmentation models have been mainly conducted by the model developers, a recipe for potential bias in evaluation[5, 6, 7, 8, 9]. Moreover, these studies often emphasize performance against ground truth annotations, overlooking broader, real-world variability. In contrast, our study extended beyond conventional evaluation by assessing model performance through segmentation area, cell count, and a targeted sensitivity analysis under varying stain conditions. While many existing models are based on proven architectures like U-Net and Mask R-CNN—renowned for their robustness—they may fall short in leveraging the advancements offered by newer models trained on larger, more diverse datasets that better represent the full spectrum of cellular and nuclear variability.

To robustly assess segmentation performance, we evaluated five models—StarDist, Cellpose, HoverNet, Mesmer, and InstaSeg—against ground truth annotations using a range of Intersection-over-Union (IoU) thresholds ($\tau = 0.1$ to 0.9). This analysis spanned both

H&E and fluorescence microscopy images, providing a comparative view of model behavior across modalities and how metrics like accuracy, false positives, precision, and false negatives change with different thresholds.

In H&E-stained images, HoverNet, Cellpose, and StarDist showed higher performance across multiple metrics. Cellpose maintained high recall values (> 0.9) up to $\tau = 0.5$, reflecting strong sensitivity, with HoverNet closely following. StarDist stood out for its high precision and lowest false positive rate, indicating reliable instance discrimination. InstaSeg, however, consistently underperformed, exhibiting sharp declines in F1 score and recall at stricter IoU thresholds, along with high false positive and false negative counts. True positives were most consistently identified by Cellpose, HoverNet, and StarDist, with InstaSeg missing many ground truth instances, especially as overlap requirements increased.

For fluorescence images, comparing Mesmer, StarDist, InstaSeg, and Cellpose revealed parallel trends. StarDist again emerged as the strongest overall performer, achieving the highest precision and F1 scores across most IoU thresholds and the fewest false positives. Cellpose followed closely, excelling particularly in recall and balancing precision adequately. Mesmer and InstaSeg underperformed, particularly at higher IoU thresholds, where both models produced increasing false positives and decreasing recall. The sharp drop in true positives and rise in false positives with increasing τ across all models underscores the growing difficulty of accurate segmentation under stricter overlap constraints.

Based on ground truth comparison, StarDist consistently offers the best trade-off between precision and recall across both imaging modalities, making it highly reliable for instance segmentation. Cellpose is particularly strong in recall and sensitivity, suitable for applications requiring maximal detection. HoverNet, evaluated only in the H&E context, strikes a balanced performance profile, while InstaSeg shows limitations in both precision and consistency across modalities.

Analysis of segmentation behavior without ground truth annotations, based on total segmented area and nuclei count, indicated distinct model-specific patterns across both fluorescence and H&E images. Multidimensional scaling analyses highlighted major differences in how models segment nuclei and estimate areas. For fluorescence images, Cellpose stood out with markedly divergent area and count patterns compared to other models, reflecting its tendency toward larger, more variable segmentation regions. StarDist and InstaSeg showed close similarity in area variation, while StarDist and Mesmer aligned strongly in nucleus count estimates, underscoring methodological affinities in handling dense or sparse regions. Bland–Altman analyses further confirmed proportional biases and heteroscedasticity across model pairs, with Cellpose generally predicting larger areas and higher nucleus counts than StarDist, and Mesmer showing more conservative segmentation than StarDist. In H&E image segmentation, a parallel pattern emerged. HoverNet consistently deviated most strongly from the other models in both segmentation area and nucleus count, indicating a fundamentally different approach or criteria for instance delineation. Cellpose and StarDist clustered tightly in terms of nucleus count, with Cellpose producing consistently larger segmented areas and higher object counts than StarDist, echoing the fluorescence findings. InstaSeg tended toward smaller segmentation areas and fewer nuclei than Cellpose, aligning with its fluorescence image behavior. The wide limits of agreement in Bland–Altman comparisons involving HoverNet and InstaSeg suggest higher variability and less consistency in their segmentations, possibly due to distinct strategies for boundary definition or object separation. Across both modalities, Cellpose typically generated the most extensive segmentations, while InstaSeg was more conservative, and StarDist exhibited intermediate behavior but struggled somewhat in complex, dense regions—likely a consequence of its star-convex shape assumptions. These results emphasize that segmentation model performance and behavior are highly context-dependent, varying by imaging modality and dataset complexity. The observed heteroscedasticity and biases highlight that segmentation disagreements tend to amplify with increasing image complexity and density.

Sensitivity analysis (Figures 23 and 24) highlights insights into how well segmentation models generalize under domain shifts induced by stain variability. These experiments focused on the CWZ and UMCU domains, selected due to their relatively divergent stain characteristics as identified in the HistauGAN-generated stain similarity heatmap (Figure 22). Although all domains demonstrated high inter-domain stain consistency (cosine similarity 0.98), UMCU consistently exhibited slightly lower similarity scores with other domains.

At an IoU threshold of 0.5, StarDist demonstrated the highest robustness and domain invariance, outperforming other models across all key metrics—achieving a precision of 0.74, recall of 0.96, and an F1-score of 0.83—while also recording the lowest false positive (2400) and false negative (1200) counts. Its consistent performance, even in a challenging and heterogeneous domain such as UMCU, highlights StarDist's resilience to stain variability and its suitability for cross-domain generalization.
Cellpose and HoverNet also performed well, though slightly below StarDist, with F1 scores of 0.82 and 0.80, respectively. Their moderate drops in precision and slightly elevated error rates indicate some sensitivity to domain-specific stain variations but still within acceptable generalization margins.
InstaSeg showed the weakest performance under stain variability, with low precision (0.58), lower F1-score (0.70), and the highest error rates (FP = 4700, FN = 2300). The results suggest that InstaSeg is more vulnerable to domain shifts and may benefit from additional domain-specific fine-tuning or augmentation strategies to improve its robustness.

Single-modality or single-metric evaluations can obscure critical differences in model behavior, particularly regarding generalization and robustness. Our findings emphasize the importance of ground-truth-free evaluations for real-world applicability, especially when annotations are limited or unavailable. Among the models assessed, StarDist consistently demonstrated the most reliable performance across diverse settings, positioning it as a strong default choice, for instance, segmentation—though it may require fine-tuning in densely clustered environments. Cellpose showed high sensitivity and excelled in Recall, making it viable for applications requiring robust detection. These findings align with results from the Bland-Altman analysis and sensitivity testing using HistoGAN, which confirmed Cellpose's tendency to identify more nuclei across modalities. However, its tendency toward over-segmentation necessitates careful interpretation. InstaSeg showed limited domain generalization and would benefit from domain-specific retraining or augmentation strategies. HoverNet performed well under consistent imaging conditions, but its performance varied with changes in image type, indicating it is best suited for uniform, controlled datasets.

5 Societal Relevance,Ethical Considerations and Key Stakeholders

This study utilized fully anonymized datasets, particularly publicly available MoNuSeg [18] and Tissuenet datasets[9]. Both datasets are standardized and compliant with the General Data Protection Regulation (GDPR). By relying on these ethically sourced datasets, we uphold data privacy and security standards, particularly critical in medical and biomedical research involving patient-related data. This study's societal contribution is anchored in its value proposition to contribute to digital pathology and biomedical image analysis developments. Precise and accurate cell segmentation is critical in various medical diagnostics and research, including carcinoma grading, disease progression monitoring, and tissue analysis. Highly generalizable models can support clinicians and researchers in delivering faster, more accurate diagnoses and insights.

Deep learning models perform remarkably in biomedical image analysis and generally exhibit good generalizability. However, their level of precision may still vary when applied to unseen data, which could pose limitations in clinical or diagnostic contexts. Therefore, it is ethical that any model integrated into real-world applications be subject to continuous performance monitoring, regular retraining with updated and representative data, and thorough cross-validation. Moreover, to mitigate the risk of misdiagnosis, outputs from these models should be reviewed by qualified pathologists. Human oversight remains essential to ensure diagnostic reliability and to uphold patient safety and ethical standards in medical decision-making.

Key stakeholders in this research include academic and research institutions, which can benefit from validating robust and reproducible methodologies. Medical institutions, too, stand to gain as improved segmentation tools streamline diagnostic workflows and enhance patient care. Finally, data scientists and developers in biomedical imaging and artificial intelligence will find value in these findings as a benchmark for building and refining domain-adaptive, high-performing models. This study thus stresses the importance of ethically sound practices and interdisciplinary collaboration in driving meaningful technological progress in healthcare and life sciences.

6 Conclusion

While this study focused primarily on segmentation quality—a critical first step in cell classification and other downstream analyses—there are several promising directions for future research. Our findings demonstrated that StarDist consistently outperforms other models in precision, demonstrating robust performance in accurately delineating cell boundaries across diverse imaging modalities, including Hematoxylin and Eosin,fluorescence and even images generated synthetically using stain GANs. However,it is worth noting that even though HoverNet lacks an off-the-shelf version trained for fluorescence images, it has shown significant performance in H&E data. Furthermore, HoverNet showed robustness during sensitivity analysis under varying staining conditions, suggesting potential for broader applicability with domain-specific adaptation. Cellpose, on the other hand, tended to over-segment nuclei in both fluorescence and H&E images. Despite this, its architecture shows strong potential for handling complex and crowded cellular environments, where StarDist may underperform by undersegmenting overlapping cells. A promising avenue for future work would be developing a hybrid model that combines the strengths of StarDist and Cellpose. Specifically, such a model could employ a composite loss function that integrates the StarDist loss—which includes star-convex polygon representations, distance maps, and centroid probability maps—with the vector flow-based loss functions used in Cellpose. This fusion could leverage the precision of StarDist in accurately outlining nuclei with the adaptability of Cellpose in managing densely packed or irregular cellular arrangements.

References

[1] K. Chen, N. Zhang, L. Powers, and J. Roveda, "Cell nuclei detection and segmentation for computational pathology using deep learning," in *2019 Spring Simulation Conference (SpringSim)*, (Tucson, AZ, USA), pp. 1–6, 2019.

[2] S. Xia, Q. Sun, Y. Zhou, Z. Wang, C. You, K. Ma, and M. Liu, "A lightweight neural network for cell segmentation based on attention enhancement," *Information*, vol. 16, no. 4, 2025.

[3] A. Eklund, P. Dufort, D. Forsberg, and S. M. LaConte, "Medical image processing on the gpu – past, present and future," *Medical Image Analysis*, vol. 17, no. 8, pp. 1073–1094, 2013.

[4] Y. Xu, R. Quan, W. Xu, Y. Huang, X. Chen, and F. Liu, "Advances in medical image segmentation: A comprehensive review of traditional, deep learning and hybrid approaches," *Bioengineering*, vol. 11, no. 10, 2024.

[5] M. Weigert and U. Schmidt, "Nuclei instance segmentation and classification in histopathology images with stardist," in *2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC)*, pp. 1–4, IEEE, 2022.

[6] M. Weigert, U. Schmidt, R. Haase, K. Sugawara, and G. Myers, "Star-convex polyhedra for 3d object detection and segmentation in microscopy," in *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[7] M. Weigert and U. Schmidt, "Nuclei instance segmentation and classification in histopathology images with stardist," in *The IEEE International Symposium on Biomedical Imaging Challenges (ISBIC)*, 2022.

[8] C. Stringer, T. Wang, M. Michaelos, and M. Pachitariu, "Cellpose: a generalist algorithm for cellular segmentation," *Nature Methods*, vol. 18, no. 1, pp. 100–106, 2021.

[9] N. F. Greenwald, G. Miller, E. Moen, A. Kong, A. Kagel, T. Dougherty, C. C. Fullaway, B. J. McIntosh, K. X. Leow, M. S. Schwartz, *et al.*, "Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning," *Nature biotechnology*, vol. 40, no. 4, pp. 555–565, 2022.

[10] P. Kaur and J. C. Stoltzfus, "Bland–altman plot: A brief overview," *International Journal of Academic Medicine*, vol. 3, no. 1, pp. 110–111, 2017.

[11] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE Reviews in Biomedical Engineering*, vol. 2, pp. 147–171, 2009.

[12] K. Al-Refu, "General methods in preparation of skin biopsies for haematoxylin and eosin stain and immunohistochemistry," *Skin Biopsy–Perspectives. InTech*, pp. 19–30, 2011.

[13] L. Bu, B. Shen, and Z. Cheng, "Fluorescent imaging of cancerous tissues for targeted surgery," *Advanced Drug Delivery Reviews*, vol. 76, pp. 21–38, 2014. Targeted imaging.

[14] N. Bouteldja, D. L. Hölscher, R. D. Bülow, I. S. Roberts, R. Coppo, and P. Boor, "Tackling stain variability using cyclegan-based stain augmentation," *Journal of Pathology Informatics*, vol. 13, p. 100140, 2022.

[15] J. Pocock, S. Graham, Q. D. Vu, M. Jahanifar, S. Deshpande, G. Hadjigeorghiou, A. Shephard, R. M. S. Bashir, M. Bilal, W. Lu, D. Epstein, F. Minhas, N. M. Rajpoot, and S. E. A. Raza, "TIAToolbox as an end-to-end library for advanced tissue image analytics," *Communications Medicine*, vol. 2, p. 120, sep 2022.

[16] U. Schmidt, M. Weigert, C. Broaddus, and G. Myers, "Cell detection with star-convex polygons," in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II*, pp. 265–273, 2018.

[17] T. Goldsborough, A. O'Callaghan, F. Inglis, L. Leplat, A. Filby, H. Bilen, and P. Bankhead, "A novel channel invariant architecture for the segmentation of cells and nuclei in multiplexed images using instanseg," *bioRxiv*, 2024.

[18] N. Kumar, R. Verma, D. Anand, Y. Zhou, O. F. Onder, E. Tsougenis, H. Chen, P.-A. Heng, J. Li, Z. Hu, Y. Wang, N. A. Koohbanani, M. Jahanifar, N. Z. Tajeddin, A. Gooya, N. Rajpoot, X. Ren, S. Zhou, Q. Wang, D. Shen, C.-K. Yang, C.-H. Weng, W.-H. Yu, C.-Y. Yeh, S. Yang, S. Xu, P. H. Yeung, P. Sun, A. Mahbod, G. Schaefer, I. Ellinger, R. Ecker, O. Smedby, C. Wang, B. Chidester, T.-V. Ton, M.-T. Tran, J. Ma, M. N. Do, S. Graham, Q. D. Vu, J. T. Kwak, A. Gunda, R. Chunduri, C. Hu, X. Zhou, D. Lotfi, R. Safdari, A. Kascenas, A. O'Neil, D. Eschweiler, J. Stegmaier, Y. Cui, B. Yin, K. Chen, X. Tian, P. Gruening, E. Barth, E. Arbel, I. Remer, A. Ben-Dor, E. Sirazitdinova, M. Kohl, S. Braunewell, Y. Li, X. Xie, L. Shen, J. Ma, K. D. Baksi, M. A. Khan, J. Choo, A. Colomer, V. Naranjo, L. Pei, K. M. Iftekharuddin, K. Roy, D. Bhattacharjee, A. Pedraza, M. G. Bueno, S. Devanathan, S. Radhakrishnan, P. Koduganty, Z. Wu, G. Cai, X. Liu, Y. Wang, and A. Sethi, "A multi-organ nucleus segmentation challenge," *IEEE Transactions on Medical Imaging*, vol. 39, no. 5, pp. 1380–1391, 2020.

[19] R. Wang, Y. Qiu, X. Hao, S. Jin, J. Gao, H. Qi, Q. Xu, Y. Zhang, and H. Xu, "Simultaneously segmenting and classifying cell nuclei by using multi-task learning in

multiplex immunohistochemical tissue microarray sections," *Biomedical Signal Processing and Control*, vol. 93, p. 106143, 2024.

[20] W. C. C. Tan, S. N. Nerurkar, H. Y. Cai, H. H. M. Ng, D. Wu, Y. T. F. Wee, J. C. T. Lim, J. Yeong, and T. K. H. Lim, "Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy," *Cancer Communications*, vol. 40, no. 4, pp. 135–153, 2020.

[21] S. Jiang, C. N. Chan, X. Rovira-Clavé, H. Chen, Y. Bai, B. Zhu, E. McCaffrey, N. F. Greenwald, C. Liu, G. L. Barlow, J. L. Weirather, J. P. Oliveria, T. Nakayama, I. T. Lee, M. S. Matter, A. E. Carlisle, D. Philips, G. Vazquez, N. Mukherjee, K. Busman-Sahay, M. Nekorchuk, M. Terry, S. Younger, M. Bosse, J. Demeter, S. J. Rodig, A. Tzankov, Y. Goltsev, D. R. McIlwain, M. Angelo, J. D. Estes, and G. P. Nolan, "Combined protein and nucleic acid imaging reveals virus-dependent B cell and macrophage immunosuppression of tissue microenvironments," *Immunity*, vol. 55, pp. 1118–1134.e8, June 2022.

[22] C. Sommer, C. Straehle, U. Köthe, and F. A. Hamprecht, "Ilastik: Interactive learning and segmentation toolkit," in *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 230–233, 2011.

[23] L. Zhu, F. Lee, J. Cai, H. Yu, and Q. Chen, "An improved feature pyramid network for object detection," *Neurocomputing*, vol. 483, pp. 127–139, 2022.

[24] F. Ozge Unel, B. O. Ozkalayci, and C. Cigla, "The power of tiling for small object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[25] S. Graham, Q. D. Vu, S. E. A. Raza, A. Azam, Y. W. Tsang, J. T. Kwak, and N. Rajpoot, "Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images," *Medical Image Analysis*, p. 101563, 2019.

[26] J. Gamper, N. A. Koohbanani, K. Benes, S. Graham, M. Jahanifar, S. A. Khurram, A. Azam, K. Hewitt, and N. Rajpoot, "Pannuke dataset extension, insights and baselines," 2020.

[27] Kurnianingsih, K. H. S. Allehaibi, L. E. Nugroho, Widyawan, L. Lazuardi, A. S. Prabuwono, and T. Mantoro, "Segmentation and classification of cervical cells using deep learning," *IEEE Access*, vol. 7, pp. 116925–116941, 2019.

[28] A. N. Gajjar and J. Jethva, "Intersection over union based analysis of image detection/segmentation using cnn model," in *2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T)*, pp. 1–6, 2022.

[29] P. S. Myles and J. Cui, "I. using the bland–altman method to measure agreement with repeated measures," *BJA: British Journal of Anaesthesia*, vol. 99, pp. 309–311, 09 2007.

[30] J. C. Gutiérrez Pérez, D. Otero Baguer, and P. Maass, "Staincut: Stain normalization with contrastive learning," *Journal of Imaging*, vol. 8, no. 7, 2022.

[31] F. G. Zanjani, S. Zinger, B. E. Bejnordi, J. A. W. M. van der Laak, and P. H. N. de With, "Stain normalization of histopathology images using generative adversarial networks," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 573–577, 2018.

[32] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, and N. Navab, "Structure-preserving color normalization and sparse stain separation for histological images," *IEEE transactions on medical imaging*, vol. 35, no. 8, pp. 1962–1971, 2016.

[33] K. N. S. R. B. M. M. C. d. B. W. P. T. Wagner, S. J., "Structure-preserving multi-domain stain color augmentation using style-transfer with disentangled representations," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, 2021.

[34] A. Anghel, M. Stanisavljevic, S. Andani, N. Papandreou, J. H. Rüschoff, P. Wild, M. Gabrani, and H. Pozidis, "A high-performance system for robust stain normalization of whole-slide images in histopathology," *Frontiers in Medicine*, vol. Volume 6 - 2019, 2019.

[35] S. Mookiah, K. Parasuraman, and S. Kumar Chandar, "Color image segmentation based on improved sine cosine optimization algorithm," *Soft Computing*, vol. 26, pp. 13193–13203, Dec. 2022.

Appendix:

Appendix

The code used for this project is publicly available on GitHub:

- **Repository:** https://github.com/isamwata/Cell-Segmentation_project
- The repository includes all Jupyter notebooks (`.ipynb`), scripts, and evaluation tools used in this thesis.

Table 4: F1 values across IoU thresholds

Model	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
StarDist	0.88	0.88	0.87	0.85	0.80	0.75	0.63	0.32	0.02
Cellpose	0.88	0.88	0.87	0.85	0.79	0.73	0.60	0.30	0.02
HoverNet	0.87	0.87	0.86	0.84	0.79	0.74	0.64	0.36	0.03
InstaSeg	0.78	0.77	0.74	0.70	0.64	0.58	0.47	0.23	0.01

Table 5: ACCURACY values across IoU thresholds

Model	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
StarDist	0.78	0.78	0.77	0.74	0.67	0.60	0.46	0.19	0.01
Cellpose	0.78	0.78	0.77	0.74	0.66	0.58	0.43	0.18	0.01
HoverNet	0.77	0.76	0.75	0.72	0.65	0.59	0.47	0.22	0.01
InstaSeg	0.63	0.62	0.59	0.54	0.47	0.41	0.31	0.13	0.01

Table 6: PRECISION values across IoU thresholds

Model	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
StarDist	0.86	0.86	0.85	0.83	0.78	0.73	0.62	0.31	0.02
Cellpose	0.80	0.80	0.79	0.77	0.72	0.66	0.55	0.27	0.02
HoverNet	0.83	0.82	0.82	0.80	0.75	0.70	0.61	0.34	0.02
InstaSeg	0.74	0.73	0.71	0.67	0.61	0.55	0.44	0.22	0.01

Table 7: RECALL values across IoU thresholds

Model	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
StarDist	0.90	0.90	0.89	0.87	0.82	0.77	0.65	0.33	0.02
Cellpose	0.98	0.97	0.96	0.94	0.88	0.81	0.67	0.33	0.02
HoverNet	0.92	0.92	0.91	0.89	0.83	0.78	0.68	0.38	0.03
InstaSeg	0.82	0.81	0.78	0.74	0.68	0.61	0.49	0.24	0.01

Table 8: FP values across IoU thresholds

Model	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
StarDist	891.00	905.00	957.00	1087.00	1363.00	1691.00	2399.00	4330.00	6183.00
Cellpose	1462.00	1476.00	1534.00	1676.00	2041.00	2453.00	3298.00	5316.00	7212.00
HoverNet	1185.00	1211.00	1255.00	1381.00	1703.00	2019.00	2675.00	4461.00	6646.00
InstaSeg	1774.00	1843.00	1986.00	2238.00	2634.00	3055.00	3765.00	5289.00	6702.00

Table 9: FN values across IoU thresholds

Model	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
StarDist	594.00	608.00	660.00	790.00	1066.00	1394.00	2102.00	4033.00	5886.00
Cellpose	143.00	157.00	215.00	357.00	722.00	1134.00	1979.00	3997.00	5893.00
HoverNet	491.00	517.00	561.00	687.00	1009.00	1325.00	1981.00	3767.00	5952.00
InstaSeg	1107.00	1176.00	1319.00	1571.00	1967.00	2388.00	3098.00	4622.00	6035.00

Table 10: TP values across IoU thresholds

Model	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
StarDist	5409.00	5395.00	5343.00	5213.00	4937.00	4609.00	3901.00	1970.00	117.00
Cellpose	5860.00	5846.00	5788.00	5646.00	5281.00	4869.00	4024.00	2006.00	110.00
HoverNet	5623.00	5597.00	5553.00	5427.00	5105.00	4789.00	4133.00	2347.00	162.00
InstaSeg	5007.00	4938.00	4795.00	4543.00	4147.00	3726.00	3016.00	1492.00	79.00