Silesian University of Technology
Faculty of Biomedical Engineering
Department of Medical Informatics and AI

Project Report

Nuclear Heterogeneity Analysis Using Deep Learning Models

A Comparative Evaluation of Cellpose and StarDist

Sania Dutta

Album No.: 308834

BSc Biomedical Engineering
Specialization: Medical Imaging Analysis

Supervisor: Dr. Arkadiusz Gertych

Zabrze, 2025

Nuclear Heterogeneity Analysis Using Deep Learning Models: A Comparative Evaluation of Cellpose and StarDist

1. Abstract

Quantifying nuclear heterogeneity in histopathology images is essential for assessing tissue abnormalities, particularly in cancer diagnosis. In this study, we evaluate two state-of-the-art deep learning models: Cellpose and StarDist, for their performance in segmenting nuclei from H&E-stained histological tiles. Both models were applied to the same dataset of tumour regions, and their outputs were analysed for shape, size, and staining intensity features. Statistical comparisons (t-test, Mann-Whitney U, Cohen's d) revealed significant differences across most metrics, while unsupervised clustering (Silhouette Score, DBI, CH Index) showed superior biological grouping in StarDist outputs. A weighted scorecard integrating these evaluations concluded that StarDist provides more consistent, interpretable, and biologically plausible segmentations.

This study underscores the importance of model selection in computational pathology, where subtle differences in segmentation can significantly impact downstream analyses such as phenotype assessment and disease characterization.

Contents

2. Introduction

Figure 1 Histology Slides in lab setting [1]

The appearance of nuclei in haematoxylin and eosin (H&E)-stained histopathology slides holds vital clues about tissue health, disease progression, and even cancer prognosis. Variability in nuclear size, shape, and staining intensity, collectively termed nuclear heterogeneity, is often one of the earliest indicators of pathological transformation [2]. Analysing this heterogeneity quantitatively has become an essential part of modern digital pathology.

With the rise of deep learning, models like Cellpose and StarDist have emerged as powerful tools for automated nuclei segmentation [3, 4]. However, each model uses fundamentally different assumptions and architectures. This raises a critical question for researchers and clinicians alike:

> "Do two state-of-the-art deep learning models (Cellpose and StarDist) produce the same analysis results on the same dataset, or do their inherent differences impact biological interpretation?"

This project explores that question in depth, evaluating how these models segment nuclei, quantify heterogeneity, and influence downstream analysis such as clustering or phenotype assessment.

3. Background & Theory

3.1 Overview of Nuclear Heterogeneity

In histopathology, nuclear heterogeneity refers to the variation in nuclear size, shape, staining, and spatial arrangement observed across cells within a tissue. These variations are often heightened in diseased tissues, particularly in cancers, where nuclear atypia is a hallmark of malignancy [2]. Quantifying this heterogeneity provides important diagnostic and prognostic insight, aiding in tumour grading, assessing aggressiveness, and even predicting treatment response [1].

Yet, manual analysis is subjective and time-consuming. That’s where automated deep learning models step in - promising consistency, scalability, and deeper phenotypic profiling.

3.2 Theoretical Background: Cellpose

Cellpose is a generalist, deep learning-based segmentation model designed to work across many cell types and image modalities [4]. Its core idea is to predict spatial vector fields that point from every pixel toward the centre of the nearest object (nucleus). This allows the model to robustly segment overlapping or elongated cells.

Key characteristics:

- ✓ Uses a U-Net architecture trained on diverse datasets.
- ✓ Excels at separating closely packed or touching nuclei.
- ✓ Outputs instance masks along with optional boundary confidence.

3.3 Theoretical Background: StarDist

StarDist approaches nuclei segmentation differently. It represents each nucleus as a star-convex polygon, estimating distances from the nucleus centre to its boundary along fixed radial directions [3].

Key characteristics:

- ✓ Tailored for well-separated nuclei with clear edges.

- ✓ Generates highly regular shapes, often resulting in cleaner masks.
- ✓ Strong performance in classical histopathological images.

StarDist is particularly powerful for tasks requiring morphological precision, such as calculating circularity or solidity [3].

3.4 Features Extracted & Their Relevance

The models extract features across three broad categories:

A. Spatial Features

- ➢ *Centroid X / Y:* Help evaluate nuclei distribution patterns within tissue.

B. Shape Features

- ➢ *Area, Length, Circularity, Solidity, Max/Min Diameter:* Capture geometric complexity; irregularity here often signals pathology.

C. Staining Features

- ➢ *Haematoxylin / Eosin (Mean, Median, Min, Max, Std):* Reflect intensity and consistency of staining; critical for nuclear-cytoplasmic contrast.

D. Density Metrics

- ➢ *Nuclei Count:* Indicates cellularity and contributes to understanding tissue organization.

Together, these features enable quantifying not just if nuclei are abnormal, but how and why forming the backbone of heterogeneity analysis [2].

4. Dataset & Preprocessing

4.1 Dataset Source & Description

The dataset used in this project originates from a real-world histopathology pipeline involving H&E-stained image tiles and their corresponding Cancer Area (CA) masks. These were pre-organized into 33 patient ID folders, each representing one patient or slide ID.

Each patient ID folder contains:

- ✓ HEtiles/: H&E-stained image tiles (PNG format)
- ✓ CAtiles/: Cancer area masks identifying tumour regions (PNG format)

From this larger dataset, a subset of 5 image-mask pairs per patient ID folder was selected for testing and analysis, resulting in a total of 165 matched samples. This subset was chosen to balance computational feasibility with representativeness of the full dataset.

4.2 Image Sample Visualization

To help clarify how these samples were used, the following example images were generated:

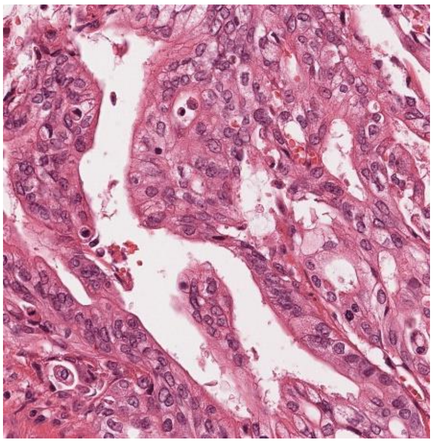


Figure 2 Original H&E Image - Tissue tile from HEtiles

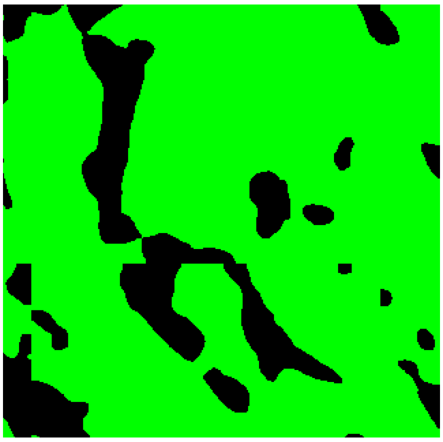


Figure 3 CA Region Mask - Tumour-specific region from CAtiles

The StarDist algorithm run by QuPath produced the following image segmentations:

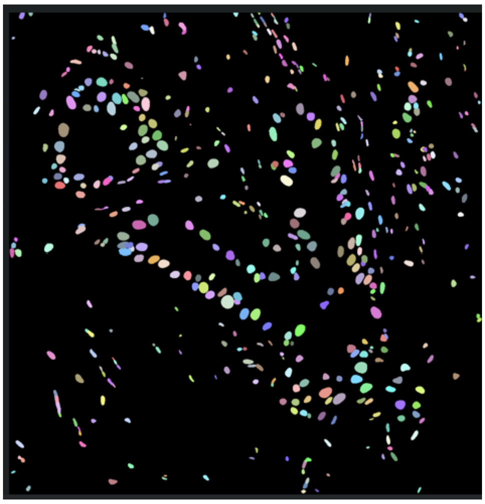


Figure 4 StarDist Mask - Instance segmentation output from StarDist

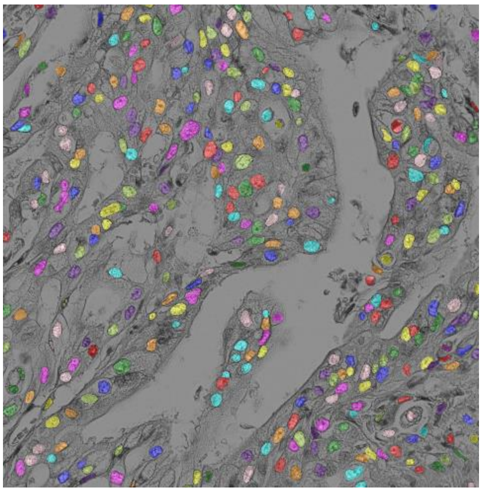


Figure 5 StarDist Overlay - Instance Segmentation of Nuclei on H&E Background

The Cellpose Algorithm produce the following image segmentation:

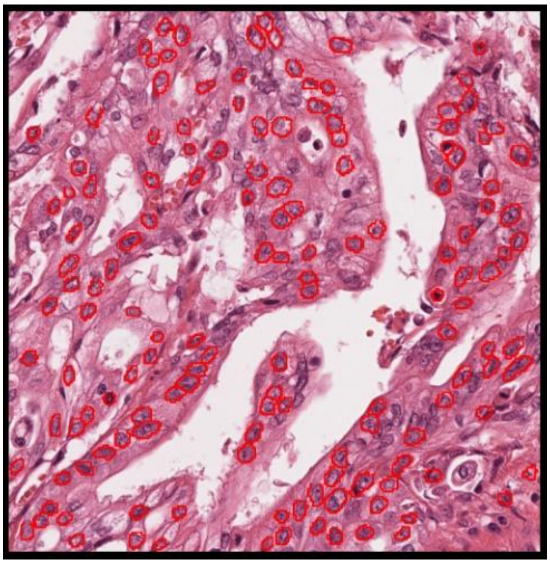


Figure 6 Cellpose - Predicted Outlines of each Individual Nucleus

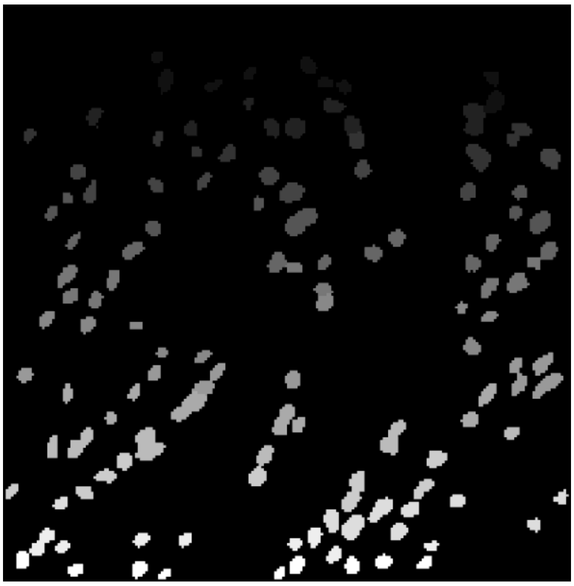


Figure 7 Predicted Mask of Cellpose

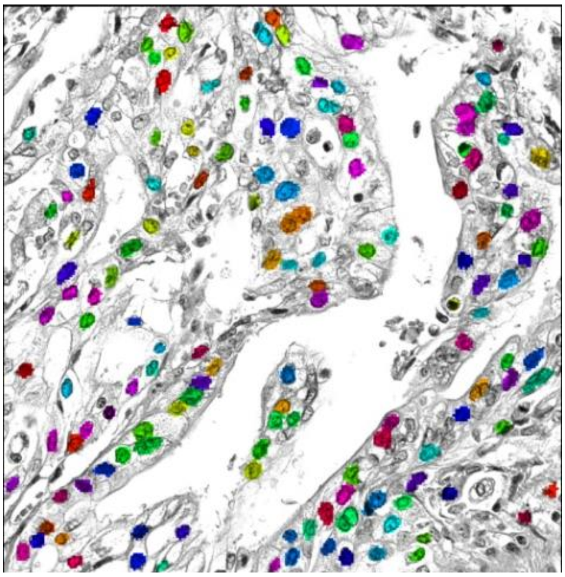


Figure 8 Cellpose Overlay with Predicted Mask

4.3 File Organization

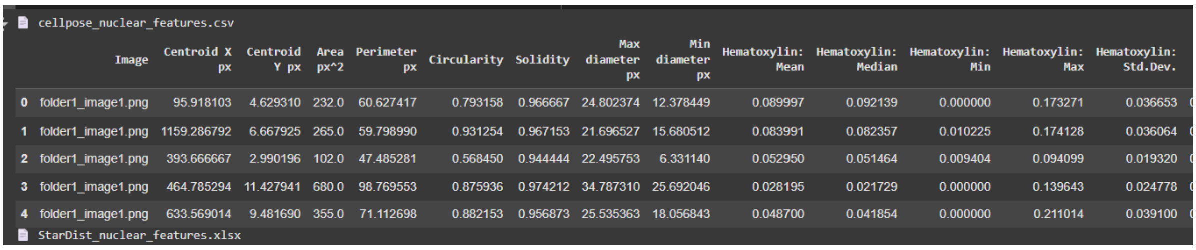
Folder and File Organization

Each of the 33 patient ID folders is named folder1, folder2, ..., folder33 and includes:

Subfolder/File	Purpose
HEtiles/	Contains 5 .png image tiles used for analysis
CAtiles/	Contains corresponding binary tumor masks
*_features.csv/.xlsx	Extracted nuclear features (Cellpose/StarDist)

Table 1 File Organisation structure

Summary feature files:

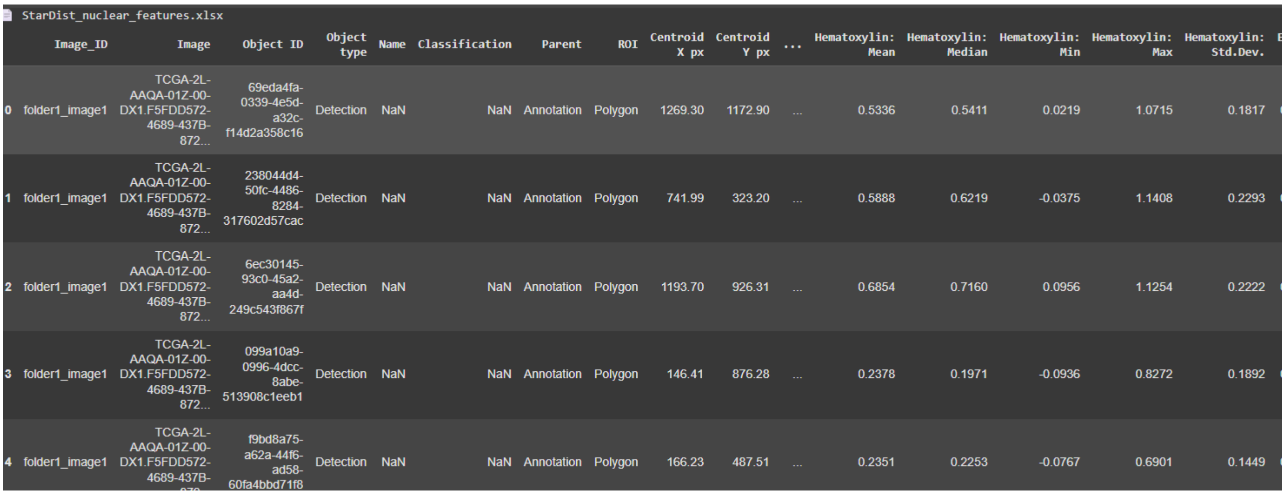


cellpose_nuclear_features.csv

	Image	Centroid X px	Centroid Y px	Area px^2	Perimeter px	Circularity	Solidity	Max diameter px	Min diameter px	Hematoxylin: Mean	Hematoxylin: Median	Hematoxylin: Min	Hematoxylin: Max	Hematoxylin: Std.Dev.
0	folder1_image1.png	95.918103	4.629310	232.0	60.627417	0.793158	0.966667	24.802374	12.378449	0.089997	0.092139	0.000000	0.173271	0.036653
1	folder1_image1.png	1159.286792	6.667925	265.0	59.798990	0.931254	0.967153	21.696527	15.680512	0.083991	0.082357	0.010225	0.174128	0.036064
2	folder1_image1.png	393.666667	2.990196	102.0	47.485281	0.568450	0.944444	22.495753	6.331140	0.052950	0.051464	0.009404	0.094099	0.019320
3	folder1_image1.png	464.785294	11.427941	680.0	98.769553	0.875936	0.974212	34.787310	25.692046	0.028195	0.021729	0.000000	0.139643	0.024778
4	folder1_image1.png	633.569014	9.481690	355.0	71.112698	0.882153	0.956873	25.535363	18.056843	0.048700	0.041854	0.000000	0.211014	0.039100

StarDist_nuclear_features.xlsx

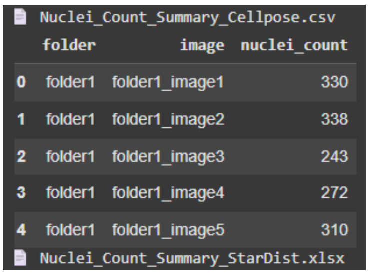
Figure 9 cellpose_nuclear_features.csv



StarDist_nuclear_features.xlsx

	Image_ID	Image	Object ID	Object type	Name	Classification	Parent	ROI	Centroid X px	Centroid Y px	...	Hematoxylin: Mean	Hematoxylin: Median	Hematoxylin: Min	Hematoxylin: Max	Hematoxylin: Std.Dev.
0	folder1_image1	TCGA-2L-AAQA-01Z-00-DX1.F5FDD572-4689-437B-872...	69eda4fa-0339-4e5d-a32c-f14d2a358c16	Detection	NaN	NaN	Annotation	Polygon	1269.30	1172.90	...	0.5336	0.5411	0.0219	1.0715	0.1817
1	folder1_image1	TCGA-2L-AAQA-01Z-00-DX1.F5FDD572-4689-437B-872...	238044d4-50fc-4486-8284-317602d57cac	Detection	NaN	NaN	Annotation	Polygon	741.99	323.20	...	0.5888	0.6219	-0.0375	1.1408	0.2293
2	folder1_image1	TCGA-2L-AAQA-01Z-00-DX1.F5FDD572-4689-437B-872...	6ec30145-93c0-45a2-aa4d-249c543f867f	Detection	NaN	NaN	Annotation	Polygon	1193.70	926.31	...	0.6854	0.7160	0.0956	1.1254	0.2222
3	folder1_image1	TCGA-2L-AAQA-01Z-00-DX1.F5FDD572-4689-437B-872...	099a10a9-0996-4dcc-8abe-513908c1eeb1	Detection	NaN	NaN	Annotation	Polygon	146.41	876.28	...	0.2378	0.1971	-0.0936	0.8272	0.1892
4	folder1_image1	TCGA-2L-AAQA-01Z-00-DX1.F5FDD572-4689-437B-	f9bd8a75-a62a-44f6-ad58-60fa4bbd71f8	Detection	NaN	NaN	Annotation	Polygon	166.23	487.51	...	0.2351	0.2253	-0.0767	0.6901	0.1449

Figure 10 StarDist_nuclear_features.xlsx

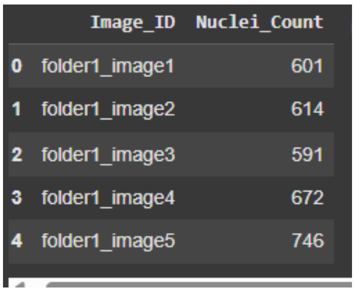


Nuclei_Count_Summary_Cellpose.csv

	folder	image	nuclei_count
0	folder1	folder1_image1	330
1	folder1	folder1_image2	338
2	folder1	folder1_image3	243
3	folder1	folder1_image4	272
4	folder1	folder1_image5	310

Nuclei_Count_Summary_StarDist.xlsx

Figure 11 Nuclei_Count_Summary_Cellpose.csv



	Image_ID	Nuclei_Count
0	folder1_image1	601
1	folder1_image2	614
2	folder1_image3	591
3	folder1_image4	672
4	folder1_image5	746

Figure 12 Nuclei_Count_Summary_StarDist.xlsx

Only these paired and filtered samples were used for all downstream processing and comparison to ensure consistency and biological relevance.

5. Segmentation Workflow

5.1 Cellpose Segmentation Pipeline

The Cellpose segmentation was performed directly in Google Colab on RGB .png H&E image tiles.

Pipeline Steps:

1. Input: H&E .png tiles from each folder were passed into Cellpose.
2. Model: The pre-trained Cellpose nuclei model was used with an approximate diameter parameter.
3. Segmentation Output: Cellpose returned:
    - Instance masks (each nucleus as a unique label).
    - Overlay previews for visual inspection.
4. Feature Extraction: Custom Python scripts were used to extract shape and stain-related features per nucleus. These included:
    - Area, circularity, solidity, max/min diameter, and centroid coordinates.
    - Haematoxylin and eosin mean, median, min, max, and standard deviation.
5. Output Format: Results were saved in cellpose_nuclear_features.csv, and summary counts were saved in Nuclei_Count_Summary_Cellpose.csv.

The model used was the pre-trained *Cellpose nuclei model* which predicts vector flows for segmentation [4].

5.2 StarDist Segmentation Pipeline

StarDist segmentation was executed using the QuPath + StarDist integration [5]. The workflow was performed on original .png H&E images.

Pipeline Steps:

1. Input: H&E-stained .png images were loaded directly into QuPath.

2. StarDist Execution: The StarDist model was run within QuPath via its extension interface (Bioimage.io).

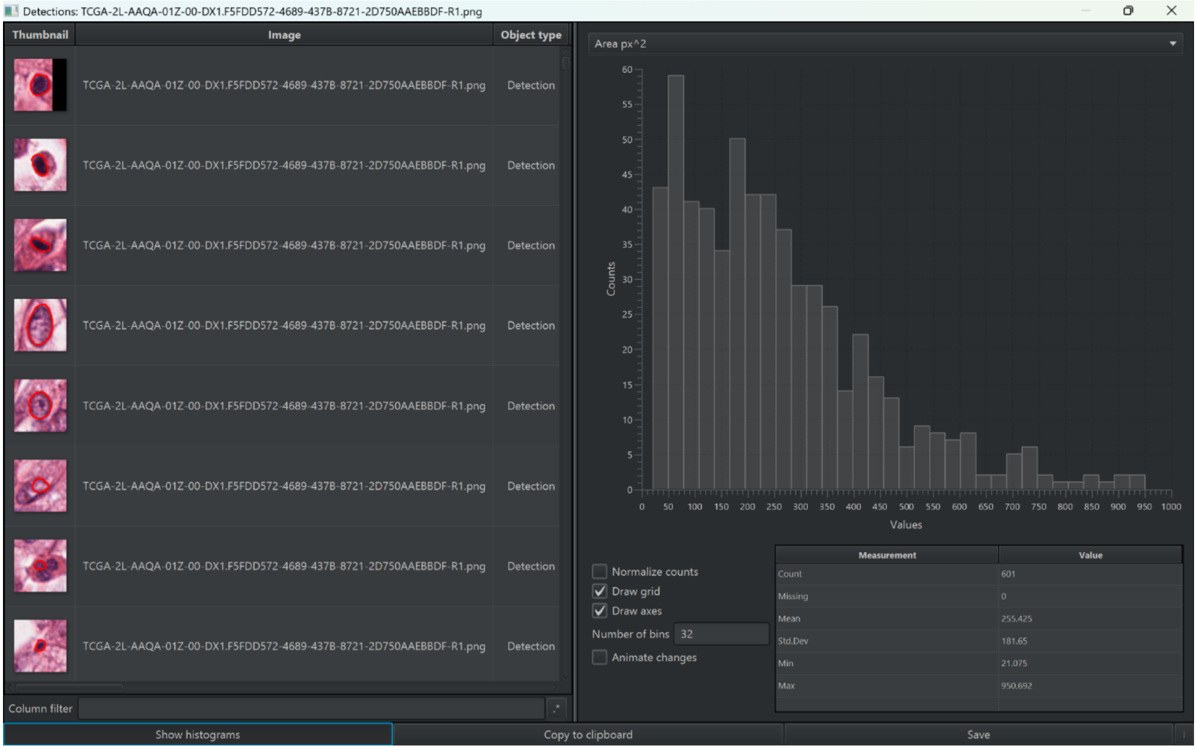
3. Detection and Annotation:

    - Each nucleus was annotated as a "Detection" object.

    - Cancer region (CA) masks were used to filter out irrelevant nuclei, ensuring feature extraction occurred only in biologically meaningful regions.

4. Feature Export: Geometric and stain-related metrics were computed for each nucleus and exported using Qu Path’s built-in measurement tools.

5. Output Format: The feature matrix was saved as StarDist_nuclear_features.xlsx, with per-image nucleus counts in Nuclei_Count_Summary_StarDist.xlsx.


Detections: TCGA-2L-AAQA-01Z-00-DX1.F5FDD572-4689-437B-8721-2D750AAEBBDF-R1.png

Thumbnail	Image	Object type
	TCGA-2L-AAQA-01Z-00-DX1.F5FDD572-4689-437B-8721-2D750AAEBBDF-R1.png	Detection
	TCGA-2L-AAQA-01Z-00-DX1.F5FDD572-4689-437B-8721-2D750AAEBBDF-R1.png	Detection
	TCGA-2L-AAQA-01Z-00-DX1.F5FDD572-4689-437B-8721-2D750AAEBBDF-R1.png	Detection
	TCGA-2L-AAQA-01Z-00-DX1.F5FDD572-4689-437B-8721-2D750AAEBBDF-R1.png	Detection
	TCGA-2L-AAQA-01Z-00-DX1.F5FDD572-4689-437B-8721-2D750AAEBBDF-R1.png	Detection
	TCGA-2L-AAQA-01Z-00-DX1.F5FDD572-4689-437B-8721-2D750AAEBBDF-R1.png	Detection
	TCGA-2L-AAQA-01Z-00-DX1.F5FDD572-4689-437B-8721-2D750AAEBBDF-R1.png	Detection
	TCGA-2L-AAQA-01Z-00-DX1.F5FDD572-4689-437B-8721-2D750AAEBBDF-R1.png	Detection

Area px^2

□ Normalize counts
☑ Draw grid
☑ Draw axes
Number of bins 32
□ Animate changes

Measurement	Value
Count	601
Missing	0
Mean	255.425
Std.Dev	181.65
Min	21.075
Max	950.692

Figure 13 Screenshot from QuPath showing the Area distribution histogram of nuclei detected by StarDist, along with thumbnails of annotated nuclei. The feature statistics (mean, std dev, min, max) are also automatically computed by QuPath.

StarDist's polygon-based instance segmentation approach was used via Qu Path's integration [3].

6. Feature Extraction

After segmentation, each nucleus was analyzed for shape and stain-related metrics, and the extracted data were saved into structured CSV and Excel files. These features form the quantitative basis for evaluating nuclear heterogeneity between Cellpose and StarDist.

For each segmented nucleus, the following features were extracted:

1. Centroid X px: X-coordinate of the nucleus centre
2. Centroid Y px: Y-coordinate of the nucleus centre
3. Area px^2: Surface area of the nucleus
4. Length px: Object length (available only for StarDist)
5. Eosin: Mean: Average pink channel intensity
6. Eosin Median
7. Eosin Min
8. Eosin Max
9. Eosin Std.Dev.
10. Image ID / Folder: Source of the tile
11. Circularity: Degree of roundness
12. Solidity: Ratio of area to convex hull area; measures concavity
13. Max diameter px: Longest straight line across the nucleus
14. Min diameter px: Shortest straight line across the nucleus
15. Haematoxylin Mean: Average blue channel intensity
16. Haematoxylin Median
17. Haematoxylin Min
18. Haematoxylin Max
19. Haematoxylin Std.Dev.

7. Features Analysis

7.1 Nuclei Count Comparison

A bar chart was generated to compare the total nuclei count across 33 images processed using both Cellpose and StarDist. Each image is labelled as folderX_imageX, and for each, we plotted two bars: one for Cellpose (in black) and one for StarDist (in grey). The bar heights represent the number of segmented nuclei in that image.

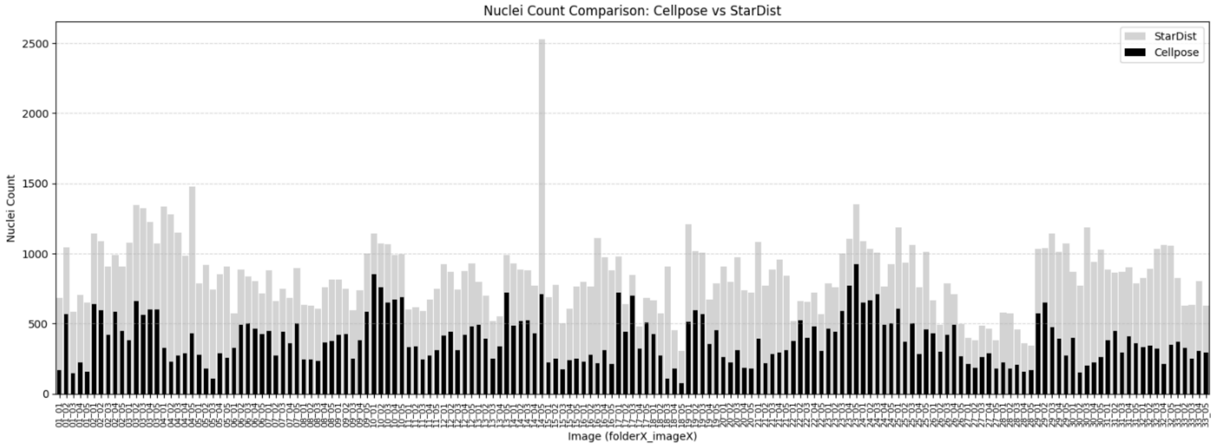


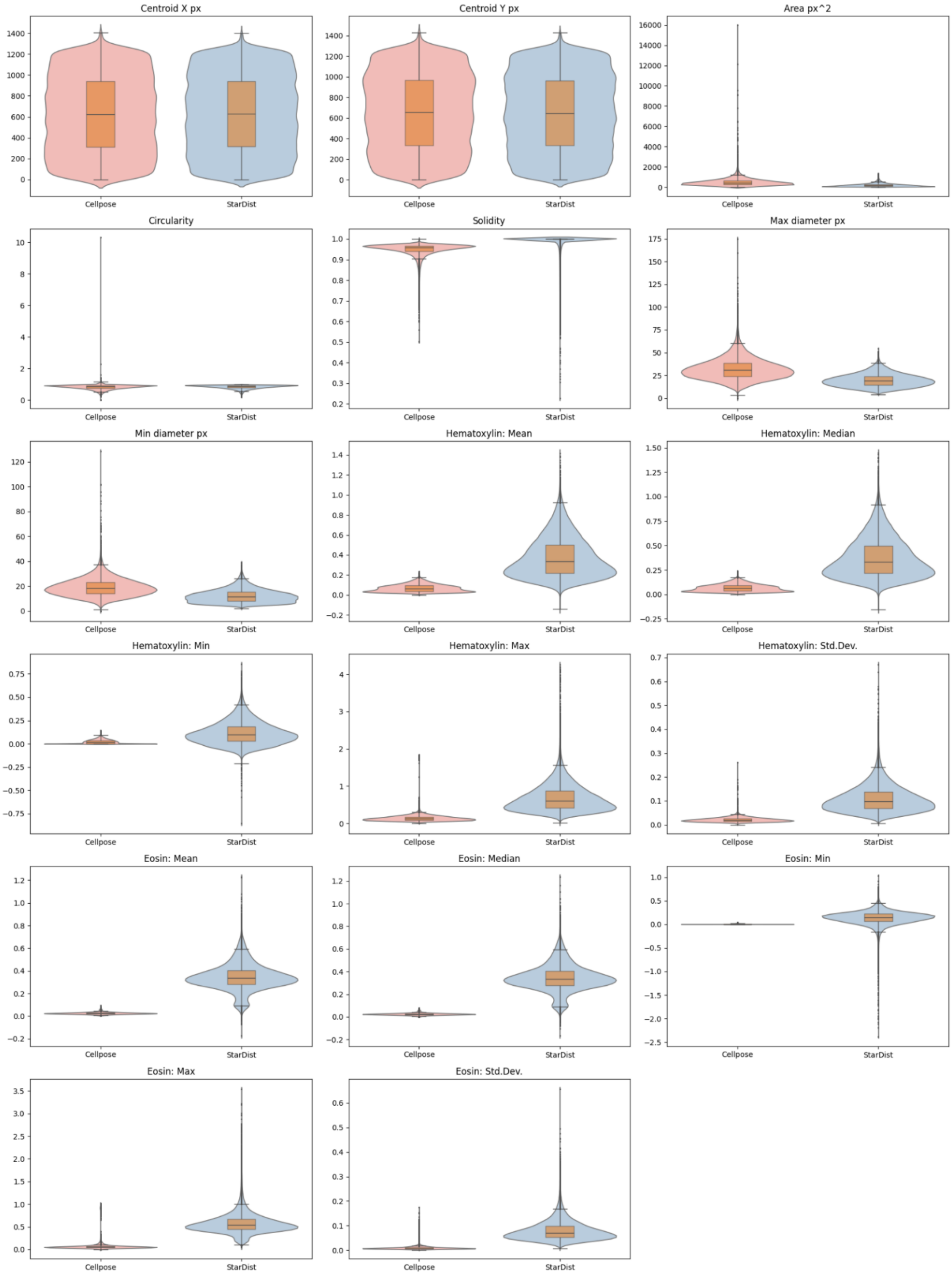
Figure 14 Comparison Graph of Nuclei Count of Stardist and Cellpose

Key Observations:

- ✓ StarDist generally detected more nuclei per image compared to Cellpose.
- ✓ In some images, Cellpose missed significant regions, possibly due to its less adaptive boundary detection in dense regions.
- ✓ The variation in detection suggests that model architecture affects sensitivity to nuclear presence, which directly impacts downstream analysis of heterogeneity [2].

7.2 Violin-Box Plot Analysis

To compare the distribution and consistency of extracted features, violin plots with embedded box plots were created for 17 shared nuclear metrics across Cellpose and StarDist.


Centroid X px

Centroid Y px

Area px^2

Circularity

Solidity

Max diameter px

Min diameter px

Hematoxylin: Mean

Hematoxylin: Median

Hematoxylin: Min

Hematoxylin: Max

Hematoxylin: Std.Dev.

Eosin: Mean

Eosin: Median

Eosin: Min

Eosin: Max

Eosin: Std.Dev.

Figure 15 Violin-Box Plots of all features

Key Observations from Plots:

- ✓ Shape Metrics:
    - ➢ *Circularity and Solidity*: StarDist outputs were tightly centred around higher values, suggesting more consistent and biologically plausible nuclear shapes.
    - ➢ Cellpose showed wider variability, especially in circularity, indicating less regular nuclear boundaries.
- ✓ Size Metrics:
    - ➢ *Area, Max/Min Diameter*: Cellpose detected nuclei with larger areas and diameters, but with greater variance.
    - ➢ StarDist had more compact distributions, indicating more uniform segmentation.
- ✓ Stain Intensity Metrics:
    - ➢ *Haematoxylin and Eosin*: StarDist exhibited higher and more varied intensity values, especially in max and mean, indicating better sensitivity to nuclear chromatin.
    - ➢ Cellpose often compressed stain values, which may suggest under-segmentation or poor staining region coverage.

Interpretation:

These feature distributions suggest that StarDist produces more uniform and biologically interpretable outputs, especially in shape regularity and staining metrics. On the other hand, Cellpose introduces more spread, which might be useful for detecting irregular cells but could also reflect noise or segmentation inconsistency.

8. Statistical Comparison

8.1 Hypothesis Testing

To compare the values of each extracted feature between the two models, we used:

- ➢ Two-sample t-test (assumes normality): Checks if the means are significantly different [6].
- ➢ Mann-Whitney U test (non-parametric): Checks if one distribution is stochastically larger than the other [6].

Feature	t-test p-value	Mann-Whitney U p-value
Centroid X px	3.4416e-01	3.0600e-01
Centroid Y px	9.9603e-04	1.1060e-03

Feature	t-test p-value	Mann-Whitney U p-value
Area px^2	0.0000e+00	0.0000e+00
Circularity	1.4455e-97	8.0453e-113
Solidity	0.0000e+00	0.0000e+00
Max diameter px	0.0000e+00	0.0000e+00
Min diameter px	0.0000e+00	0.0000e+00
Haematoxylin: Mean	0.0000e+00	0.0000e+00
Haematoxylin: Median	0.0000e+00	0.0000e+00
Haematoxylin: Min	0.0000e+00	0.0000e+00
Haematoxylin: Max	0.0000e+00	0.0000e+00
Haematoxylin: Std.Dev.	0.0000e+00	0.0000e+00
Eosin: Mean	0.0000e+00	0.0000e+00
Eosin: Median	0.0000e+00	0.0000e+00
Eosin: Min	0.0000e+00	0.0000e+00
Eosin: Max	0.0000e+00	0.0000e+00
Eosin: Std.Dev.	0.0000e+00	0.0000e+00

Table 2 Results of Stardist and Cellpose Hypothesis Testing

Interpretation:

- ✓ Most features show extremely significant differences between Cellpose and StarDist, with p-values effectively zero.
- ✓ Only Centroid X and Centroid Y show p-values above typical significance thresholds (> 0.05 for Centroid X), indicating spatial positioning of nuclei is more similar between models than other features.

8.1.1 Investigating Extremely Low p-values

To understand why the p-values were so small, I visualized the distributions of select features using kernel density estimation (KDE) plots. The goal was to assess whether the distributions were actually different in shape and spread, which would justify the low p-values, or whether the significance was inflated.

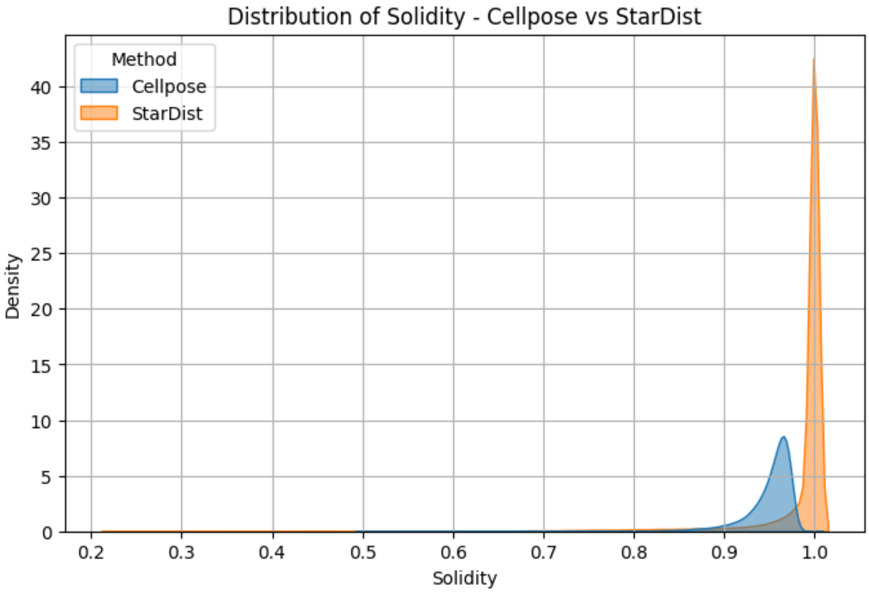


Figure 16 Solidity

StarDist exhibited a narrow peak near 1.0, whereas Cellpose had a wider, lower peak. This indicates that StarDist nuclei are more consistently compact, while Cellpose outputs are more variable in solidity.

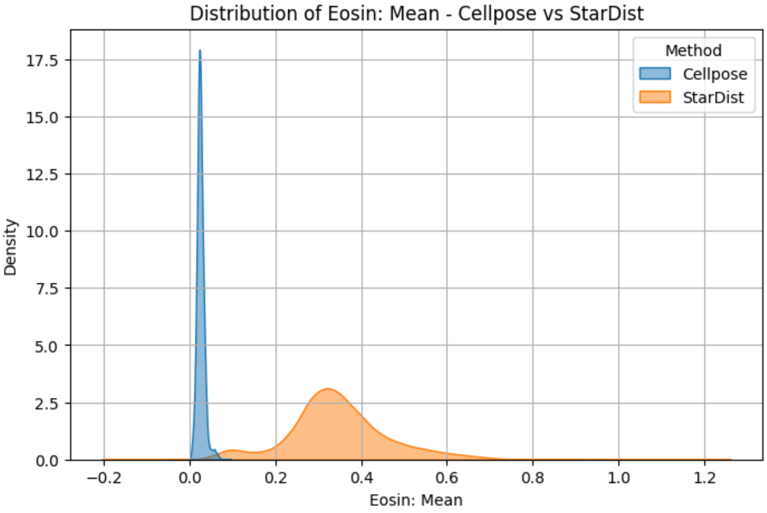


Figure 17 Eosin Mean

The distribution from StarDist was broad and shifted to higher values, suggesting it detects a wider range of cytoplasmic staining.

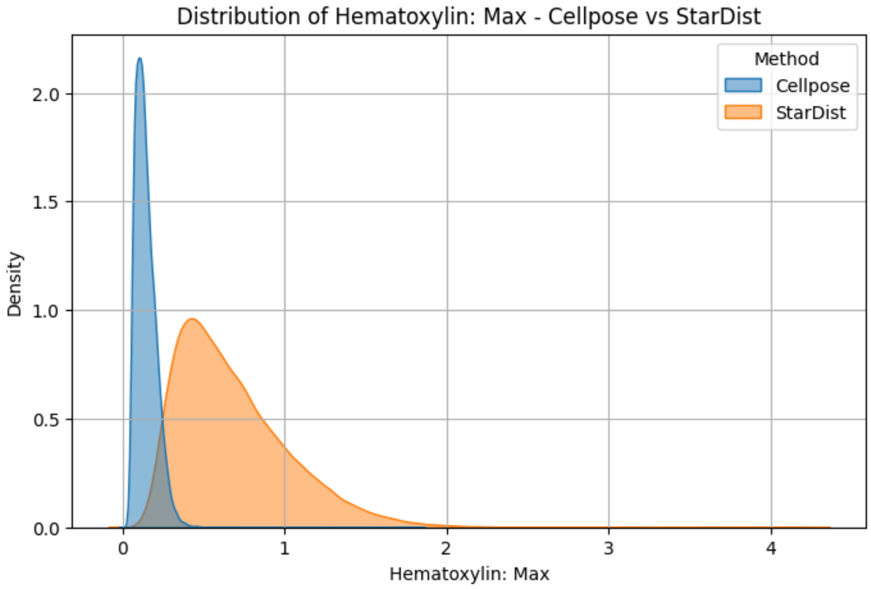


Figure 18 Haematoxylin Max

StarDist had a long-tailed distribution, while Cellpose showed a tight peak, again reflecting broader dynamic range in staining intensities.

Interpretation:

These KDE plots visually confirmed the statistical findings. The models not only differed in mean values but also in distribution shape, variance, and range of measurements. Therefore, the low p-values are valid and reflect meaningful model differences, not just artifacts of large sample size.

8.2 Cohen's d

While p-values indicate whether the difference between Cellpose and StarDist outputs is statistically significant, Cohen's d helps quantify *how large or meaningful* that difference is [6]. It is a measure of effect size, showing the standardized difference between two distributions in terms of standard deviations.

This is especially useful in our case, where the sample size is large and p-values are almost all significant. Cohen's d allows us to focus on practical significance, i.e., which features are truly different in terms of segmentation outcome.

Interpretation Guide:

- ✓ d < 0.2: Negligible difference
- ✓ d ≈ 0.5: Moderate difference
- ✓ d ≥ 0.8: Large difference

Effect size thresholds were interpreted based on conventional standards as described by Cohen and summarized in [6].

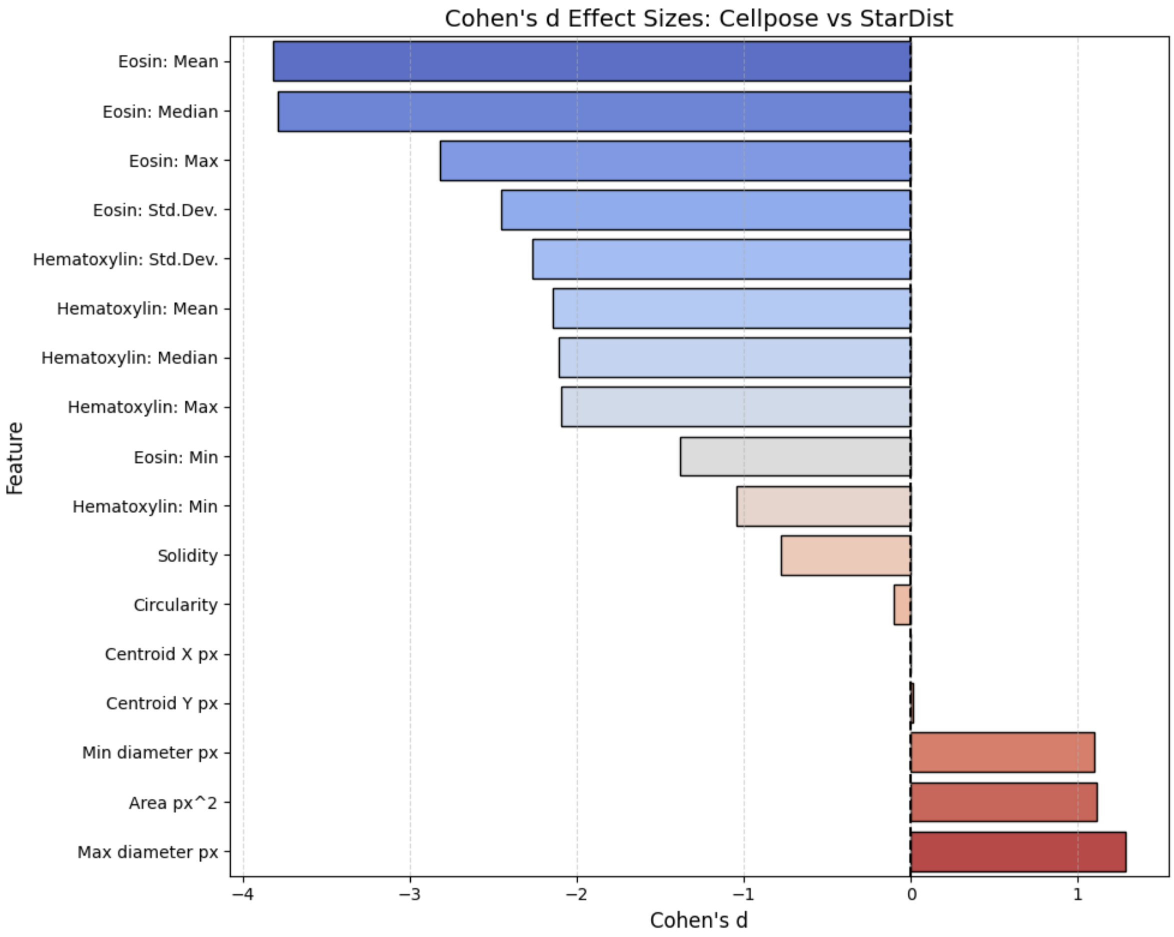


Figure 19 Cohen's d Results Summary

Key insights:

- ✓ The strongest differences were seen in Eosin-based staining metrics (e.g., Eosin: Mean, Eosin: Median) with Cohen's d values as low as -3.82, indicating much higher Eosin values detected by StarDist.
- ✓ Shape and size features like Area px^2 and Max diameter px also showed large differences (d > 1.0), favouring Cellpose.
- ✓ Centroid coordinates and Circularity had negligible effect sizes, suggesting these features are relatively consistent between models.

8.3 Total Nuclei Retained

In addition to comparing individual feature distributions, we also assessed the total number of nuclei retained after segmentation and filtering. This metric helps determine the effective coverage of each model , that is, how many nuclei were confidently identified and included in the analysis.

Why this matters?

High retention is crucial for maintaining statistical power and ensuring that the extracted features represent the full biological diversity of the image. A model that detects too few nuclei may miss heterogeneity; a model that detects too many may include noise or artifacts.

Model	Filtered Nuclei Count
Cellpose	59,979
StarDist	121,775

Table 3 Results of Nuclei Retention

Interpretation:

- ✓ StarDist retained nearly double the number of nuclei compared to Cellpose, suggesting it is more sensitive and includes more instances per image.
- ✓ However, this must be balanced with the biological plausibility of those segmentations. A higher count doesn't automatically imply better quality - some of the additional detections might include noise or over-segmentation.
- ✓ The distribution of nuclei across images was also visualized via bar plots, which confirmed this trend and showed consistency across slides.

9. Clustering & Biological Plausibility

9.1 Why Clustering?

In histopathological image analysis, we often lack a ground truth for segmentation. To overcome this, unsupervised clustering metrics offer a way to assess how well the extracted features form coherent and biologically meaningful groupings. If a model segments nuclei such that their features naturally cluster into distinct and compact groups, it supports the idea that the model has captured genuine biological variation (e.g., cancerous vs. non-cancerous nuclei).

9.2 Metrics

To evaluate the quality of clustering without relying on ground truth, we used three well-established unsupervised clustering metrics:

Silhouette Score

- ➢ measures how similar each data point is to its own cluster compared to other clusters. A higher Silhouette Score (close to 1) suggests that the nuclei within each cluster are well-grouped and distinct from other clusters.

Davies-Bouldin Index (DBI)

- ➢ evaluates the average similarity between each cluster and the most similar one to it. Lower values indicate that clusters are more distinct and have less internal variation, desirable for reliable heterogeneity analysis.

Silhouette Score

- ➢ calculates the ratio of between-cluster variance to within-cluster variance. A higher CH score implies that clusters are well-separated and compact, reinforcing the biological plausibility of the segmentation.

These metrics together give a robust, quantitative sense of how well the nuclear features extracted by each model (Cellpose and StarDist) naturally group into meaningful biological subtypes [7].

9.3 Results & Interpretation

We applied KMeans clustering (k = 3) to the shared features from both Cellpose and StarDist outputs. Below are the resulting scores:

Model	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
Cellpose	0.3753	0.9597	33,062.59
StarDist	0.4352	0.8033	101,863.28

Table 4 Results of Clustering (K-means, k=3)

Interpretation:

- ✓ StarDist outperformed Cellpose across all three metrics.
- ✓ A higher Silhouette Score (0.4352) and Calinski-Harabasz Index, along with a lower Davies-Bouldin Index, indicate that StarDist segments formed more compact and well-separated clusters.
- ✓ This implies greater biological consistency in the nuclear features extracted by StarDist.

In summary, clustering results validate the biological plausibility of StarDist's segmentations, providing further evidence that it may be a better choice for heterogeneity-based analyses.

10. Scorecard Evaluation

To objectively compare the performance of Cellpose and StarDist across multiple dimensions, a scorecard method was used. This approach allows the integration of different quantitative metrics, each with varying scales and biological implications, into a unified ranking system.

10.1 Metric Selection & Weight Allocation

The chosen metrics are grounded in unsupervised learning principles and statistical inference, selected based on their frequency of use in biomedical image analysis literature:

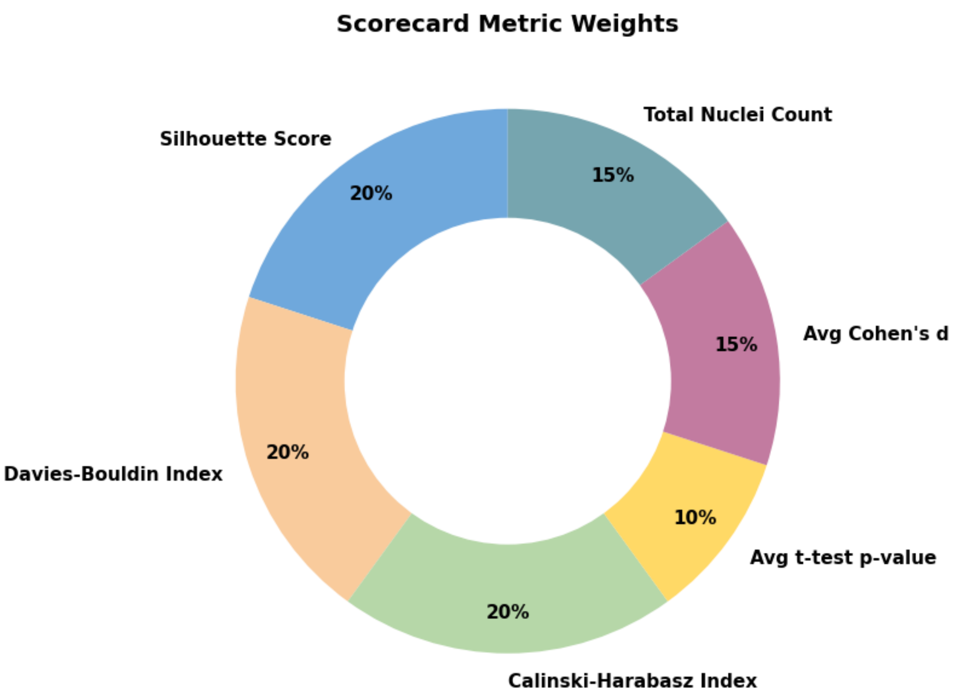


Figure 20 Pie chart showing the weight distribution

- ✓ Silhouette Score (20%): Measures cluster compactness and separation. Widely used in histopathological image clustering [8].
- ✓ Davies-Bouldin Index (20%): Captures the similarity between clusters; lower is better. Especially relevant in heterogeneous tumour subtypes [9].
- ✓ Calinski-Harabasz Index (20%): Evaluates intra/inter-cluster variability. Favoured for assessing segmentation boundaries in complex images [10].
- ✓ Average t-test p-value (10%): Indicates statistical significance of feature differences between models. Lower values suggest more reliable differentiation [11].
- ✓ Cohen's d Effect Size (15%): Measures how large the difference is between models, beyond just statistical significance [12].
- ✓ Total Nuclei Count (15%): Retention of nuclei after filtering serves as a proxy for segmentation robustness and coverage [13].

The weights were allocated based on a combination of domain relevance and literature precedence. Metrics related to unsupervised clustering (Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index) were each assigned 20% weight (totaling 60%) because they directly assess the biological plausibility of nuclear groupings, which is critical for heterogeneity analysis in pathology images [8–10].

Statistical metrics such as average p-value and Cohen's d were weighted lower (10% and 15%, respectively) since they reflect the degree of separation rather than biological interpretability alone [11, 12]. Total nuclei count was also weighted at 15% to account for segmentation coverage, drawing indirect justification from studies like Veta et al. (2015) that emphasize detection completeness [13].

This allocation reflects a deliberate choice to prioritize biological meaning and clustering validity while still considering statistical robustness and practical segmentation performance

10.2 Final Evaluation Table

Note: The "Mean t-test p-value" and "Mean |Cohen's d|" were computed by averaging the respective statistical values across all 19 shared features. This approach provides a single representative metric reflecting overall statistical difference and effect size between the two segmentation outputs.

Metric	Cellpose	StarDist	Better Value
Silhouette Score	0.3753	0.4352	StarDist
Davies-Bouldin Index	0.9597	0.8033	StarDist
Calinski-Harabasz Index	33,062.5898	101,863.2783	StarDist
Mean t-test p-value	0.0863	0.0863	Tie
Mean abs(Cohen's d)	1.6668	1.6668	Tie
Total Nuclei Retained	59,979	121,775	StarDist
Weighted scores	0.1500	0.8500	StarDist

Table 5 Comparison of Cellpose and StarDist

Based on the weighted scorecard and metrics, StarDist emerges as the superior model:

- ✓ It consistently outperforms Cellpose in clustering quality, with higher Silhouette and Calinski-Harabasz scores and a lower Davies-Bouldin Index.
- ✓ It retains double the nuclei count, suggesting more robust detection and fewer discarded regions.
- ✓ Both models achieved identical performance in terms of average t-test significance and effect size (Cohen's d), which strengthens confidence in the fairness of the comparison.

11. Discussion

This project set out to compare the segmentation results of two deep learning models (Cellpose and StarDist) on the same histopathology dataset to evaluate how their nuclei extraction and heterogeneity analyses differ.

The results consistently favoured StarDist across multiple axes:

- ✓ Statistical tests revealed significant differences in nuclear feature distributions, with StarDist generally providing tighter, more biologically plausible metrics.
- ✓ Clustering metrics (Silhouette Score, DBI, CH Index) indicated that StarDist's features allowed for more coherent clustering, implying higher-quality feature representations.
- ✓ Nuclei count was markedly higher for StarDist, suggesting it is more sensitive to cell detection, although this could be interpreted either as higher sensitivity or over-segmentation.

Interestingly, even though both models analysed the *same images*, the extracted nuclear characteristics, such as size, shape, and stain intensity, varied substantially. This calls into question the assumption that all deep learning-based segmentation models produce equivalent outputs and supports the need for thorough cross-model validation in biomedical image analysis.

12. Limitations & Future Work

- ✓ No Ground Truth Labels: The absence of manual annotations means that no absolute accuracy comparison was possible. The analysis relied entirely on internal consistency and biological plausibility.
- ✓ Sampling Size: Only a subset of tiles (first 5 per folder) from 33 folders was used, which may not fully capture dataset diversity.
- ✓ Staining Variability: Despite preprocessing, stain variations could influence intensity features like haematoxylin/eosin metrics.
- ✓ Single Parameter Setting: Both models were run using standard/default parameters, which might not be optimized for this dataset.

Future Directions:

- ✓ Incorporate expert annotations or benchmark datasets with labeled nuclei to validate segmentation accuracy directly.
- ✓ Explore tuning model parameters for both Cellpose and StarDist for improved performance.
- ✓ Extend analysis to include other segmentation models (e.g., HoverNet, Cellpose 2.0).
- ✓ Apply the models to the whole dataset for better heterogeneity understanding.

13. Conclusion

This comparative study highlights that deep learning models do not necessarily yield the same analysis outcomes even on identical datasets. StarDist and Cellpose, while both effective, differ significantly in the nuclei they extract and the downstream features they compute.

Using a combination of statistical tests, clustering validity scores, and a weighted scorecard grounded in literature-backed criteria, StarDist was found to outperform Cellpose in almost all categories. These findings underscore the importance of model selection and thorough validation in computational pathology pipelines, especially when downstream analyses (e.g., heterogeneity, prognosis) depend heavily on segmentation quality.

14. References

[1] MLM Medical Labs. (n.d.). Histological staining of tissue samples. Retrieved from https://mlm-labs.com

[2] Janowczyk, A., & Madabhushi, A. (2016). Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. Journal of Pathology Informatics, 7. https://doi.org/10.4103/2153-3539.186902

[3] Schmidt, U., Weigert, M., Broaddus, C., & Myers, G. (2018). Cell Detection with Star-convex Polygons. In *Medical Image Computing and Computer Assisted Intervention* (MICCAI). https://doi.org/10.1007/978-3-030-00934-2_30

[4] Stringer, C., Wang, T., Michaelos, M., & Pachitariu, M. (2021). Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods*, 18(1), 100–106. https://doi.org/10.1038/s41592-020-01018-x

[5] Bankhead, P., et al. (2017). QuPath: Open source software for digital pathology image analysis. *Scientific Reports*, 7(1), 16878. https://doi.org/10.1038/s41598-017-17204-5

[6] McDonald, J. H. (2014). *Handbook of Biological Statistics* (3rd ed.). Sparky House Publishing. Retrieved from http://www.biostathandbook.com

[7] Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). Understanding of internal clustering validation measures. In *2010 IEEE International Conference on Data Mining* (pp. 911–916). https://doi.org/10.1109/ICDM.2010.35

[8] Liu, Y., Chen, W., Zhou, X., et al. (2021). Unsupervised nuclei feature clustering in whole-slide images. *Computers in Biology and Medicine*, 134, 104494. https://doi.org/10.1016/j.compbiomed.2021.104494

[9] Ghosh, A., Sengupta, S., & Majumder, D. D. (2020). Histological pattern classification using unsupervised clustering. *IEEE Access*, 8, 95629–95638. https://doi.org/10.1109/ACCESS.2020.2995468

[10] Caicedo, J. C., Goodman, A., Karhohs, K. W., et al. (2019). Evaluation of deep learning strategies for nucleus segmentation in fluorescence images. *Nature Methods*, 16, 124–132. https://doi.org/10.1038/s41592-018-0268-8

[11] Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4, 863. https://doi.org/10.3389/fpsyg.2013.00863

[12] Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, 8(2), 597–599. https://doi.org/10.22237/jmasm/1257035100

[13] Veta, M., van Diest, P. J., Willems, S. M., et al. (2015). Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Medical Image Analysis*, 20(1), 237–248. https://doi.org/10.1016/j.media.2014.11.010