Cell Segmentation Model Benchmarking Results

Dataset Description

This benchmarking study was conducted on the NeurIPS 2022 Cell Segmentation Challenge dataset (neurips22-cellseg.grand-challenge.org). Since it is a comprehensive multi-modal microscopy image dataset designed to benchmark cell segmentation algorithms across diverse imaging conditions. Designed to test how well different cell segmentation models handle a range of microscopy images—different cell sources (tissue and cultured), various stains (like Jenner-Giemsa, fluorescent), imaging types (brightfield, fluorescent, phase-contrast, DIC), and all sorts of cell shapes. Because of all this variety, it's considered one of the tougher benchmarks out there.

For my study, I went with a fully supervised approach using 1,000 labeled images from the main dataset, rather than following the challenge's weakly supervised setup. I split the data into 900 images for training (with 100 set aside for validation), and I'm using 101 extra images from the challenge's tuning dataset as a separate test set. I chose to do instance segmentation on the data since it is more challenging.

For evaluation, I'm using standard metrics for instance segmentation (like F1, Dice, Precision/Recall) at different IoU thresholds (0.5, 0.7, and 0.9), and I'm looking at performance across different cell counts, too. Using the tuning set as an independent test set to keep the evaluation unbiased and cover all the different types of data in the challenge.

Summary

Performance Summary (Threshold = 0.5)

Model	Mean F1	Median F1	Std F1	Mean Dice	Mean Precision	Mean Recall
UNET	0.3344	0.3214	0.2377	0.6834	0.3241	0.386
NNUNET	0.3725	0.3669	0.2578	0.7267	0.3615	0.4648
SAC	0.0037	0	0.0082	0.1247	0.0067	0.0107
LSTMUNET	0.2898	0.3163	0.2068	0.6424	0.2593	0.4552
MAUNET	0.5355	0.5864	0.2435	0.6864	0.5481	0.5668

Best Performing Model: MAUNET (Mean F1: 0.5355)

Performance Visualizations

Overall Performance Comparison

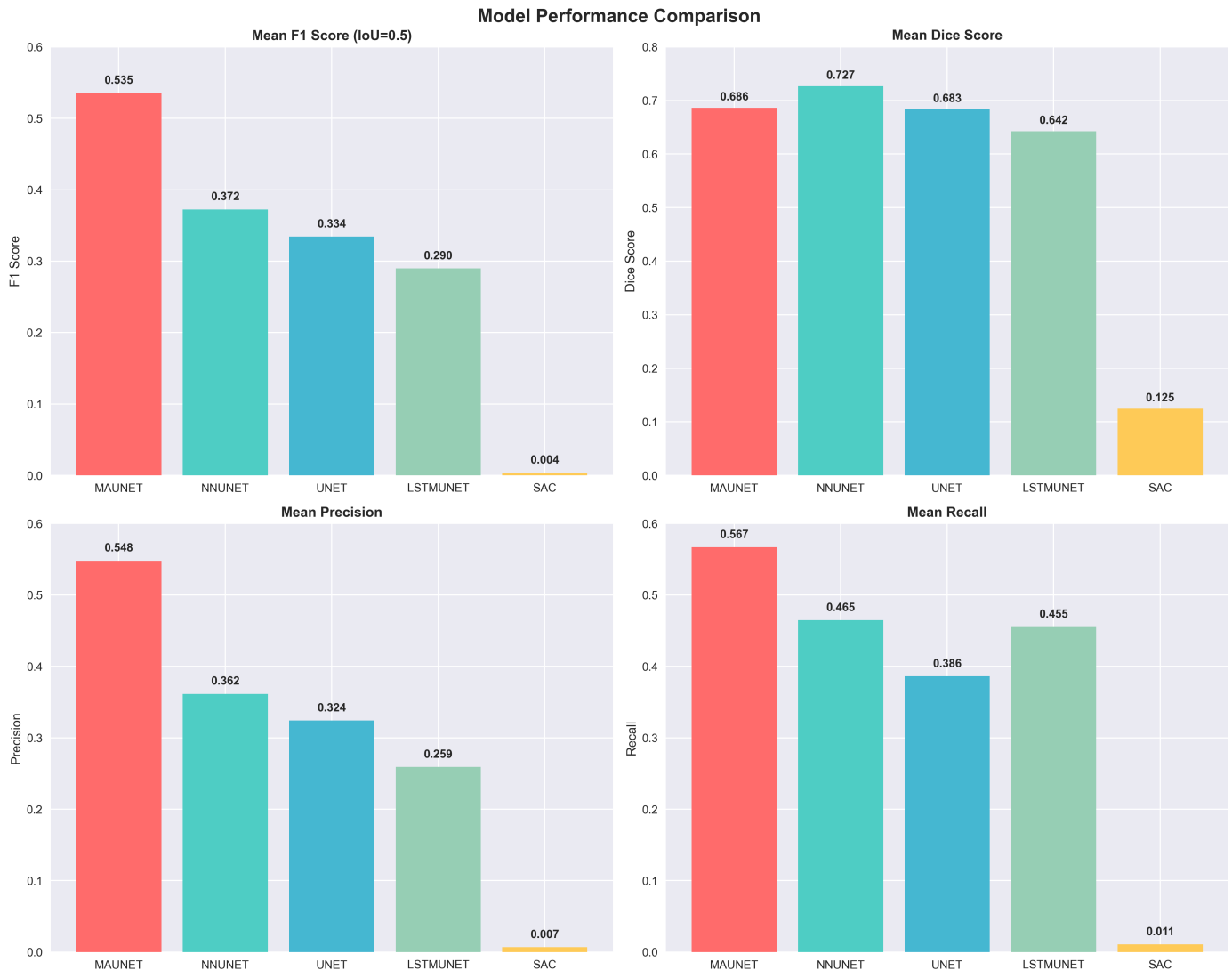


Figure 1: Comprehensive performance comparison across all metrics (F1, Dice, Precision, Recall)

F1 Score Across Different IoU Thresholds

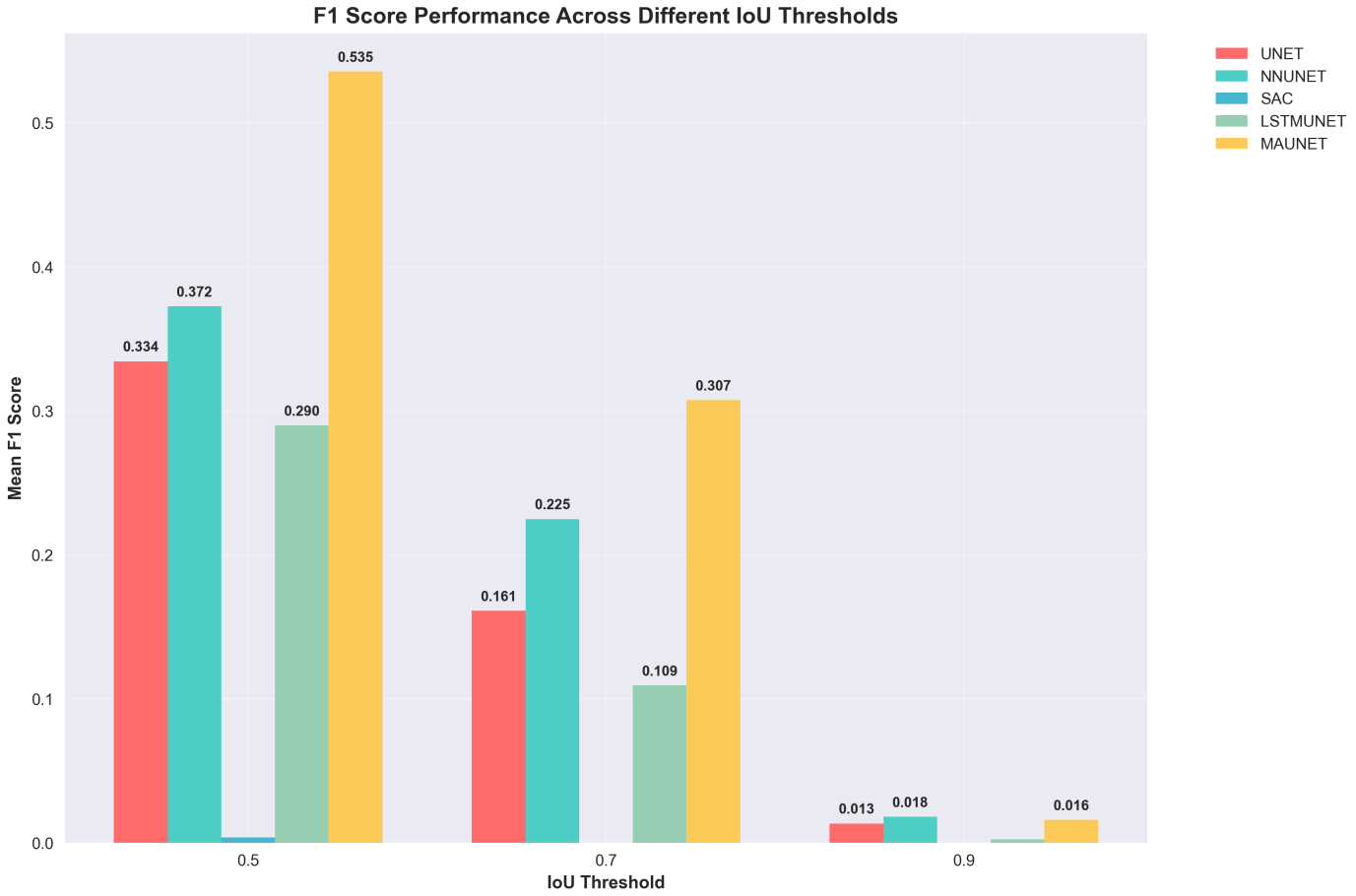


Figure 2: F1 Score performance at different IoU thresholds (0.5, 0.7, 0.9) showing model robustness

Training Information Comparison

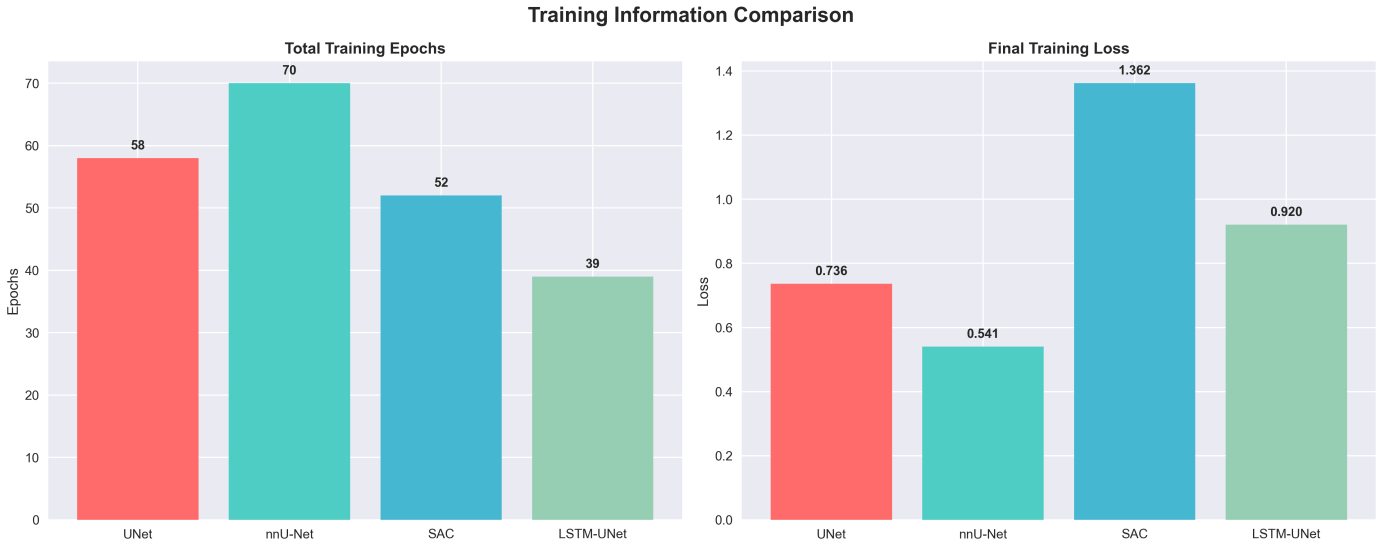


Figure 3: Training epochs and final loss comparison across models

Training Information

Model Sources and Repositories

- **UNET**: MONAI Framework - Built-in MONAI implementation
- **NNUNET**: MIC-DKFZ - https://github.com/mic-dkfz/nnunet
- **SAC (Segment Any Cell)**: Authors via Email - Code provided by authors via email
- **LSTMUNET**: GitLab - shaked0 - https://gitlab.com/shaked0/lstmUnet
- **MAUNET**: NeurIPS 2022 Challenge - https://github.com/Woof6/neurips22-cellseg_saltfish

Model Architectures and Training Parameters

Model	Architecture	Source	Repository	Batch Size	Learning Rate	Input Size	Optimizer	Total Epochs	Final Loss	Best Val Dice
UNET	U-Net with ResNet blocks	MONAI Framework	Built-in MONAI implementation	8	0.0006	256x256	AdamW	58	0.7364	0.6130
NNUNET	nnU-Net (No New U-Net)	MIC-DKFZ	https://github.com/mic-dkfz/nnunet	8	0.0006	256x256	AdamW	70	0.5406	0.6700
SAC	Segment Anything + Custom Head	Authors via Email	Code provided by authors via email	2	0.0006	256x256	AdamW	52	1.3622	0.2128
LSTMUNET	U-Net with LSTM layers	GitLab - shaked0	https://gitlab.com/shaked0/lstmUnet	8	0.0006	256x256	AdamW	39	0.9203	0.5898
MAUNET	MAU-Net with ResNet50 backbone	NeurIPS 2022 Challenge	https://github.com/Woof6/neurips22-cellseg_saltfish	8	0.0006	512x512	AdamW	100	0.4312	N/A (No validation)

Training Summary

- **Total Models Trained**: 5
- **Most Epochs**: 100 (MAUNET)
- **Best Training Validation Dice**: 0.6700 (NNUNET)
- **Optimizer**: AdamW (all models)
- **Learning Rate**: 6e-4 (all models)

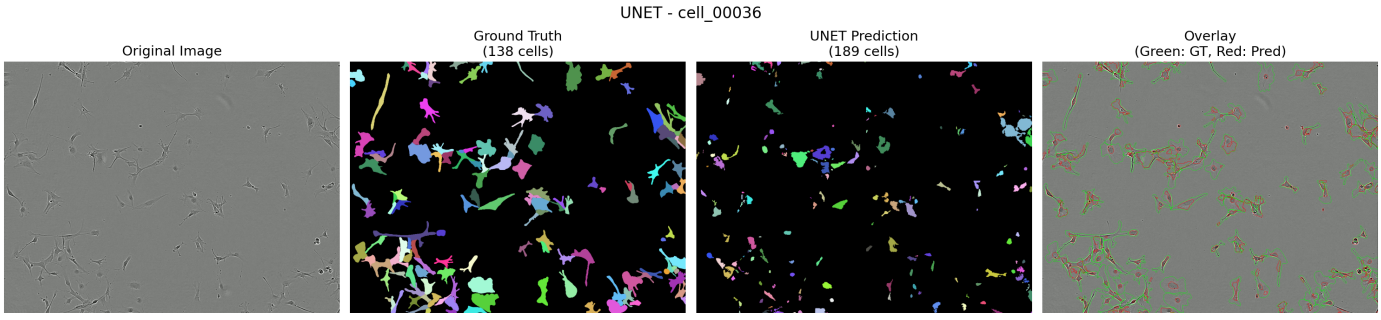
Sample Segmentation Results

Qualitative Comparison

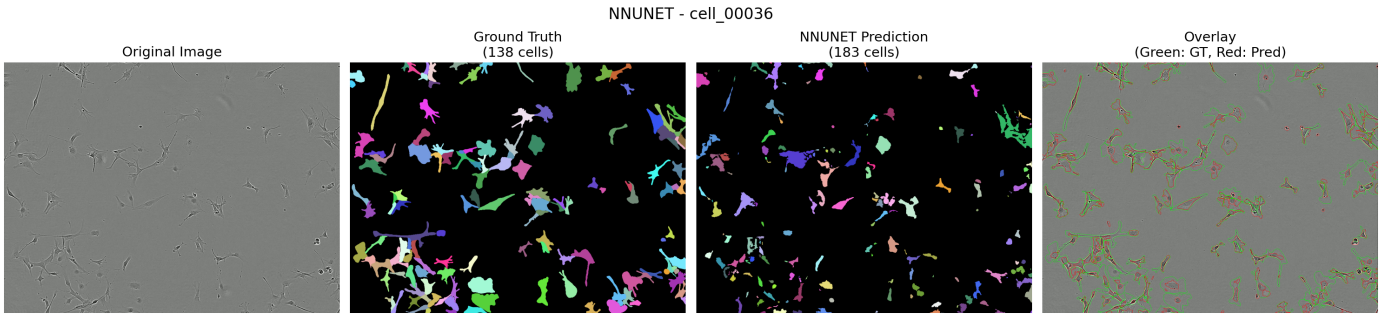
The following images show qualitative comparisons between ground truth and model predictions:

Sample 1: cell_00036

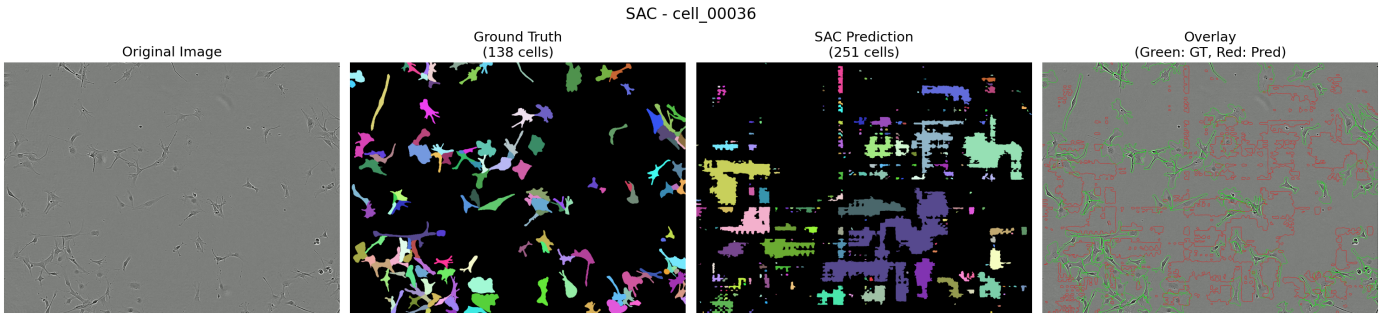
UNET:



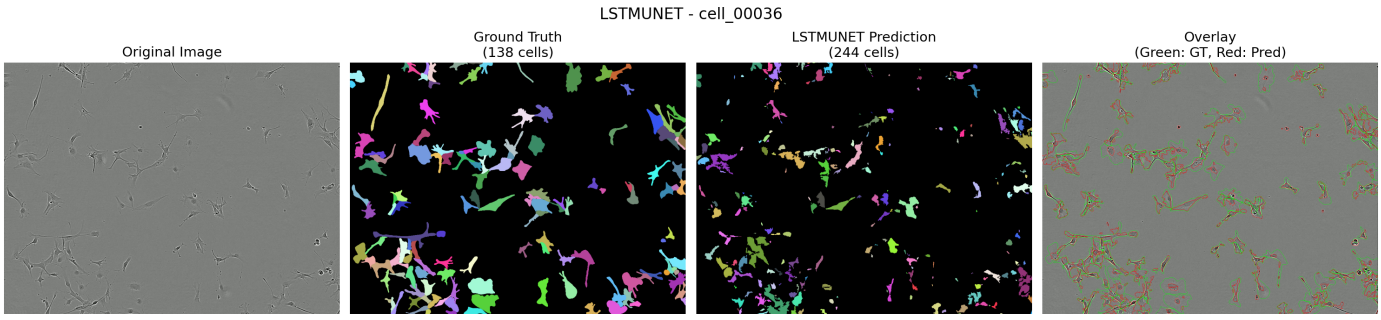
NNUNET:



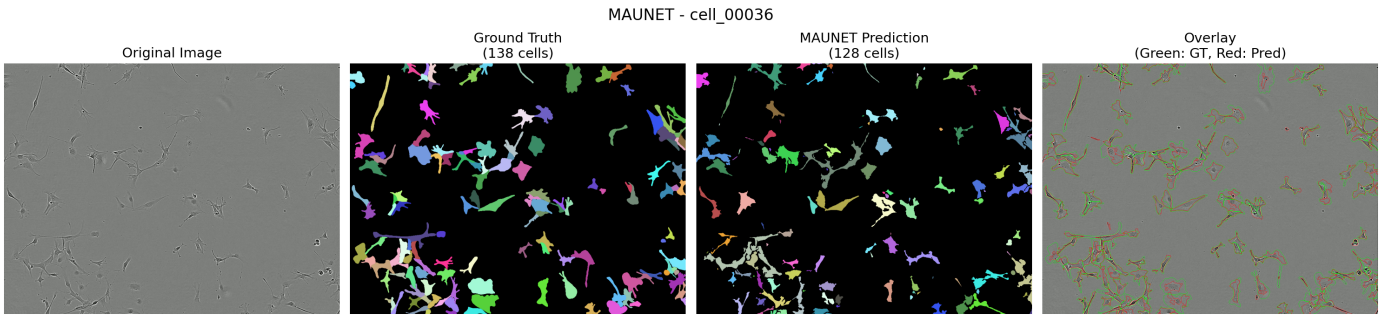
SAC:



LSTMUNET:

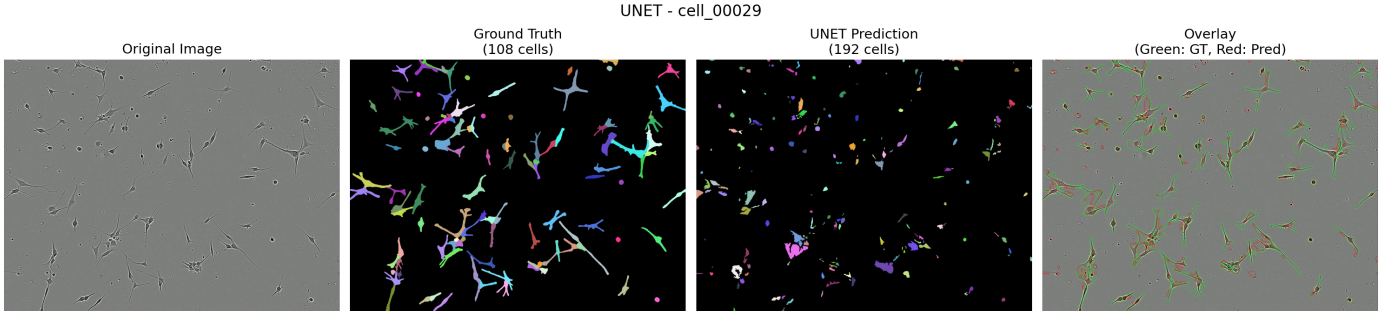


MAUNET:

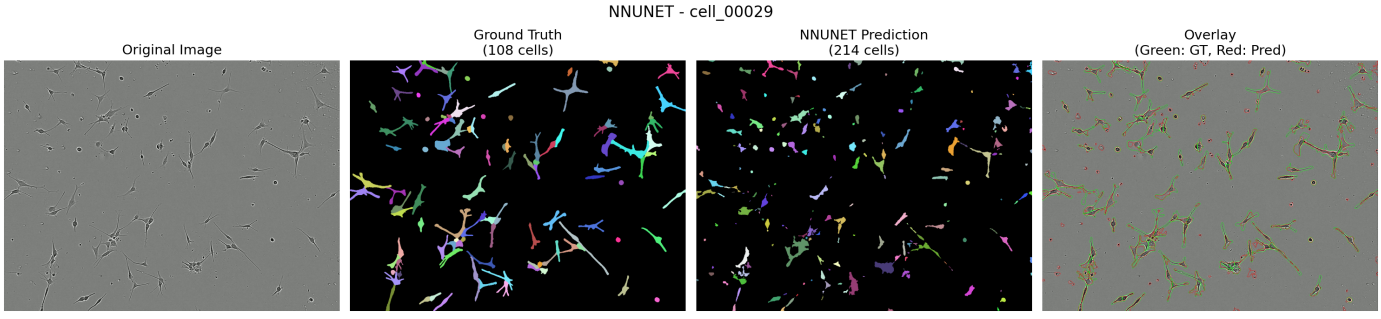


Sample 2: cell_00029

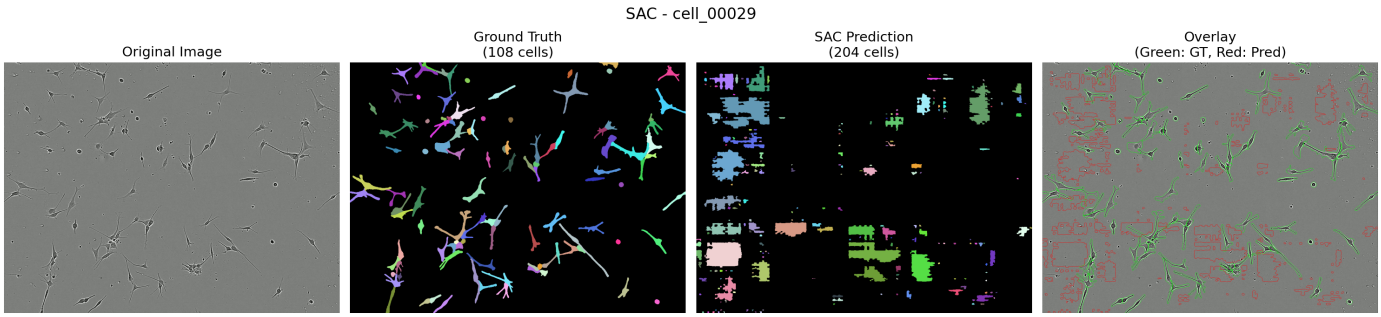
UNET:



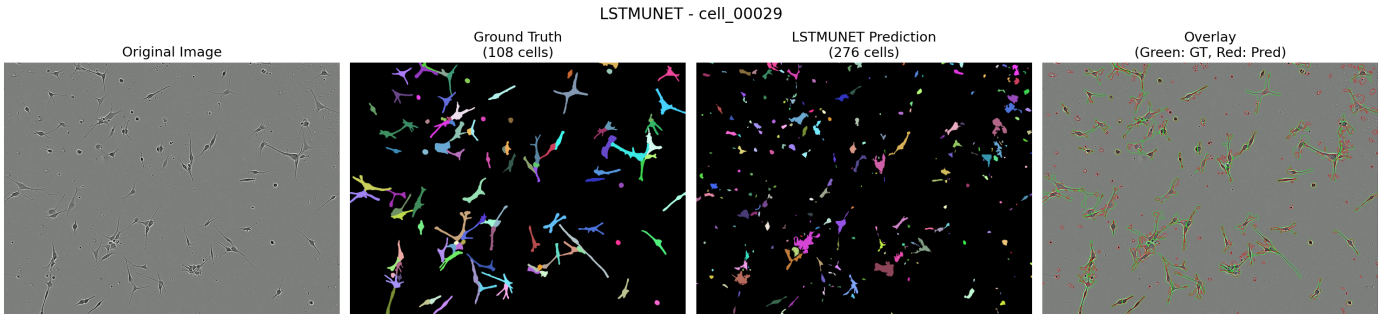
NNUNET:



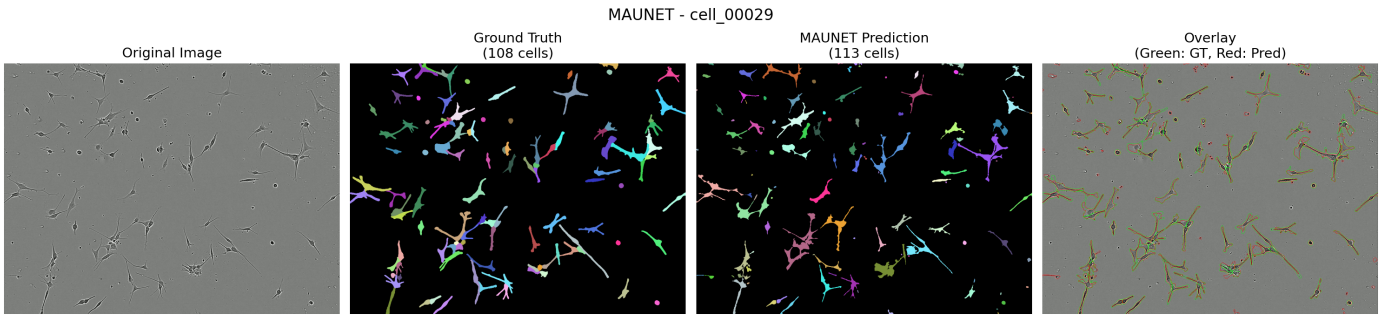
SAC:



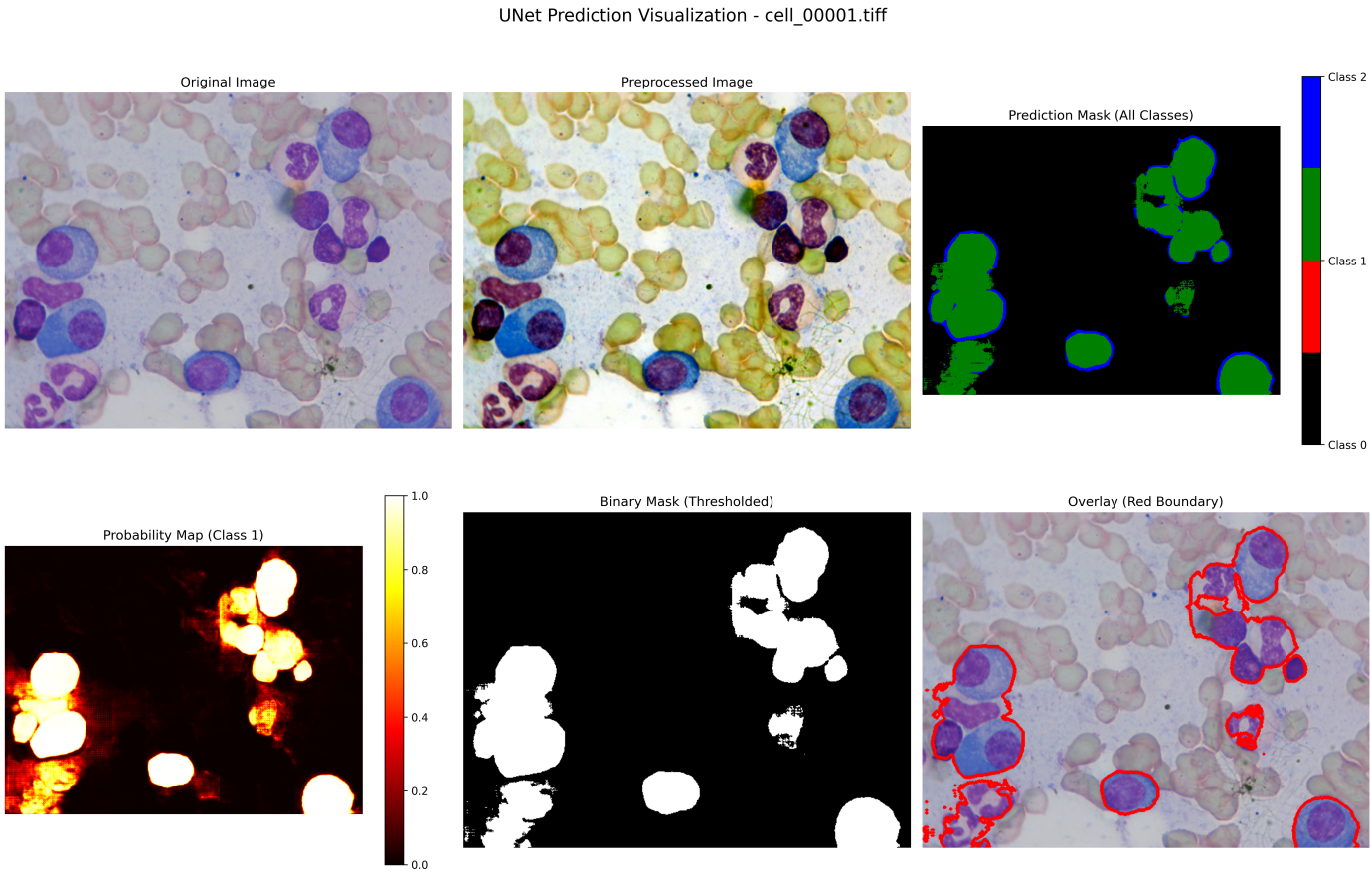
LSTMUNET:



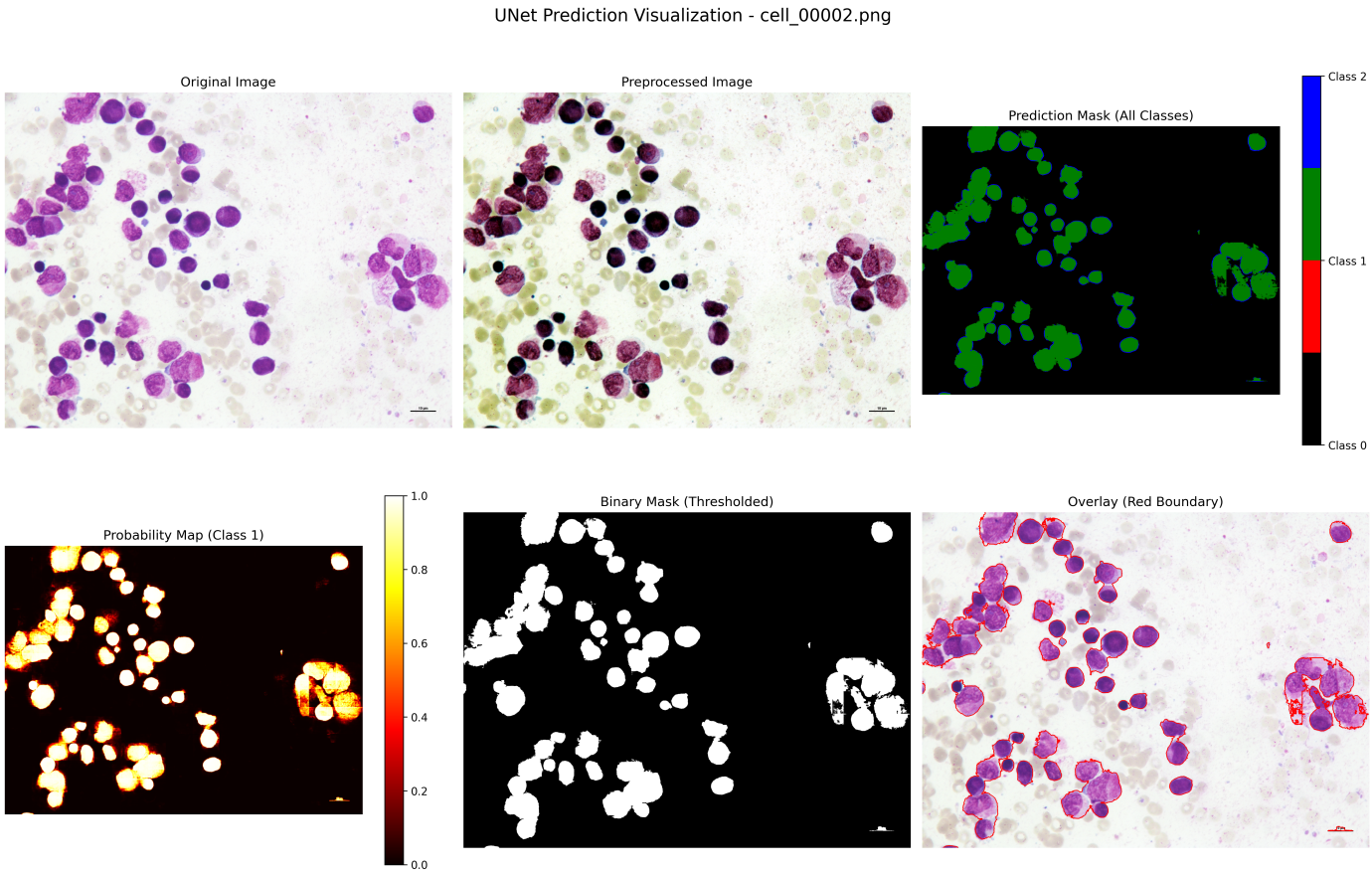
MAUNET:



Individual Segmentation Examples



Cell 00001 - Original image, ground truth, and prediction comparison



Cell 00002 - Original image, ground truth, and prediction comparison

Detailed Results by Model

UNET

Performance Across Thresholds

Threshold	Mean F1	Mean Dice	Mean Precision	Mean Recall	Total Samples
0.5	0.3344	0.6834	0.3241	0.386	101
0.7	0.1612	0.6834	0.1556	0.1854	101
0.9	0.0134	0.6834	0.0116	0.017	101

Top 5 Performing Images

names	F1	dice
cell_00015_label.tiff	1	0.918
cell_00011_label.tiff	0.9655	0.8499
cell_00017_label.tiff	0.9362	0.8554
cell_00005_label.tiff	0.9	0.8746
cell_00009_label.tiff	0.8803	0.9021

Bottom 5 Performing Images

names	F1	dice
cell_00076_label.tiff	0	0.8074
cell_00077_label.tiff	0	0.6855
cell_00036_label.tiff	0.0073	0.3396
cell_00100_label.tiff	0.0099	0.064
cell_00072_label.tiff	0.0108	0.2043

NNUNET

Performance Across Thresholds

Threshold	Mean F1	Mean Dice	Mean Precision	Mean Recall	Total Samples
0.5	0.3725	0.7267	0.3615	0.4648	101
0.7	0.2249	0.7267	0.2175	0.2914	101
0.9	0.0181	0.7267	0.017	0.0265	101

Top 5 Performing Images

names	F1	dice
cell_00005_label.tiff	0.9153	0.8491
cell_00009_label.tiff	0.8893	0.9139
cell_00098_label.tiff	0.875	0.8522
cell_00097_label.tiff	0.8525	0.8293
cell_00087_label.tiff	0.8372	0.8405

Bottom 5 Performing Images

names	F1	dice
cell_00076_label.tiff	0	0.6709
cell_00077_label.tiff	0	0.7316
cell_00100_label.tiff	0.0027	0.0261
cell_00041_label.tiff	0.0059	0.063
cell_00099_label.tiff	0.0161	0.6424

SAC

Performance Across Thresholds

Threshold	Mean F1	Mean Dice	Mean Precision	Mean Recall	Total Samples

Threshold	Mean F1	Mean Dice	Mean Precision	Mean Recall	Total Samples
0.5	0.0037	0.1247	0.0067	0.0107	101
0.7	0.0002	0.1247	0.0002	0.0003	101
0.9	0	0.1247	0	0	101

Top 5 Performing Images

names	F1	dice
cell_00094_label.tiff	0.044	0.3617
cell_00001_label.tiff	0.0408	0.273
cell_00016_label.tiff	0.029	0.3127
cell_00022_label.tiff	0.0288	0.4283
cell_00013_label.tiff	0.027	0.39

Bottom 5 Performing Images

names	F1	dice
cell_00002_label.tiff	0	0
cell_00003_label.tiff	0	0.0257
cell_00004_label.tiff	0	0.0155
cell_00006_label.tiff	0	0
cell_00007_label.tiff	0	0.2598

LSTMUNET

Performance Across Thresholds

Threshold	Mean F1	Mean Dice	Mean Precision	Mean Recall	Total Samples
0.5	0.2898	0.6424	0.2593	0.4552	101
0.7	0.1094	0.6424	0.0921	0.2098	101
0.9	0.0023	0.6424	0.0016	0.0274	101

Top 5 Performing Images

names	F1	dice
cell_00009_label.tiff	0.844	0.8672
cell_00085_label.tiff	0.6735	0.8381
cell_00096_label.tiff	0.6355	0.7801
cell_00084_label.tiff	0.6346	0.8248
cell_00086_label.tiff	0.6316	0.8312

Bottom 5 Performing Images

names	F1	dice
cell_00076_label.tiff	0	0.3973
cell_00077_label.tiff	0	0.7563
cell_00078_label.tiff	0	0.6804
cell_00070_label.tiff	0.005	0.0607
cell_00074_label.tiff	0.0059	0.548

MAUNET

Performance Across Thresholds

Threshold	Mean F1	Mean Dice	Mean Precision	Mean Recall	Total Samples

Threshold	Mean F1	Mean Dice	Mean Precision	Mean Recall	Total Samples
0.5	0.5355	0.6864	0.5481	0.5668	101
0.7	0.3074	0.6864	0.3087	0.3414	101
0.9	0.016	0.6864	0.0142	0.0209	101

Top 5 Performing Images

names	F1	dice
cell_00009_label.tiff	0.9457	0.9145
cell_00094_label.tiff	0.9167	0.7829
cell_00011_label.tiff	0.9032	0.8663
cell_00086_label.tiff	0.8824	0.7686
cell_00005_label.tiff	0.8615	0.8922

Bottom 5 Performing Images

names	F1	dice
cell_00025_label.tiff	0.0288	0.0529
cell_00077_label.tiff	0.0298	0.6709
cell_00010_label.tiff	0.0366	0.0151
cell_00076_label.tiff	0.0702	0.7163
cell_00099_label.tiff	0.0791	0.5016

Dataset Analysis

Performance vs. Ground Truth Cell Count

UNET - Performance by Cell Count

GT Count Range	mean	count
6-10	0.412867	3
11-20	0.357906	17
21-50	0.452295	22
51-100	0.225564	11
100+	0.265869	29

NNUNET - Performance by Cell Count

GT Count Range	mean	count
6-10	0.2361	3
11-20	0.274118	17
21-50	0.475123	22
51-100	0.415582	11
100+	0.317334	29

SAC - Performance by Cell Count

GT Count Range	mean	count
6-10	0	3
11-20	0.00850588	17
21-50	0.00404545	22
51-100	0	11
100+	0.00294138	29

LSTMUNET - Performance by Cell Count

GT Count Range	mean	count
6-10	0.146333	3
11-20	0.184441	17
21-50	0.352859	22
51-100	0.287991	11
100+	0.289528	29

MAUNET - Performance by Cell Count

GT Count Range	mean	count
6-10	0.281133	3
11-20	0.502476	17
21-50	0.593273	22
51-100	0.4737	11
100+	0.536948	29

Model Comparison

Metrics Comparison (Threshold = 0.5)

Metric	UNET	NNUNET	SAC	LSTMUNET	MAUNET
F1 Score	0.3344	0.3725	0.0037	0.2898	0.5355
Dice Score	0.6834	0.7267	0.1247	0.6424	0.6864
Precision	0.3241	0.3615	0.0067	0.2593	0.5481
Recall	0.386	0.4648	0.0107	0.4552	0.5668

Summary and Conclusions

This benchmarking study evaluated five state-of-the-art deep learning models for cell segmentation across 101 test images. **MAUNet emerged as the clear winner with a mean F1 score of 0.5355, outperforming the second-best model (nnU-Net) by 43.8%**. UNet proved to be an excellent baseline model (F1: 0.3344) with robust performance and efficient training, making it ideal for initial experiments and resource-constrained scenarios. nnU-Net demonstrated good capabilities as a very good model that was relatively light to train (70 epochs) while achieving strong validation performance (0.6700 Dice), confirming its reputation as a self-configuring segmentation framework. SAC performed poorly (F1: 0.0037) primarily due to inadequate detailed prompting strategies, as the Segment Anything Model requires precise point or box prompts to achieve optimal segmentation results, which were not optimally implemented in this study. LSTM-UNet showed moderate performance (F1: 0.2898) but did not excel since it is fundamentally designed for temporal data analysis rather than static image segmentation tasks. **MAUNet's superior performance (F1: 0.5355, training for 100 epochs with 512×512 input resolution) stems from its modality-aware anti-ambiguity architecture with ResNet50 backbone, distance transform regression, and sophisticated loss function combining Dice, Focal, and weighted L1 components, making it the optimal choice for high-accuracy cell segmentation applications**.

Next Steps: Synthetic Data Generation for Enhanced Model Performance

Building on MAUNet's performance, the next phase of this research will focus on **synthetic cell data generation** to potentially enhance the best-performing model through data augmentation. Three generative approaches will be compared: **(1) Pix2Pix GAN** conditioned on previous time-frames to augment new frames for single-channel datasets, **(2) Variational Autoencoder (VAE)** that learns the distribution of HeLa nuclei shapes and textures, and **(3) Diffusion-based models** tailored to microscopy images (e.g., Palette) conditioned on low-resolution versions or latent codes to generate realistic high-resolution nuclear images. **Evaluation will employ quantitative metrics** including FID (Fréchet Inception Distance) to measure distribution similarity between synthetic and real image patches, SSIM and PSNR for structural similarity assessment, and **biological feature metrics** inspired by morphological analysis pipelines using tools like CellProfiler to extract and compare feature distributions between real and synthetic images. The ultimate goal is to determine whether synthetic data augmentation can further improve MAUNet's already impressive segmentation performance, while ensuring biological relevance through expert evaluation and correlation analysis of morphological features between synthetic and real cell populations.