Cell Segmentation Model Benchmarking Results

Generated on: 2025-08-14 22:06:44

Executive Summary

Performance Summary (Threshold = 0.5)

Model	Mean F1	Median F1	Std F1	Mean Dice	Mean Precision	Me Rec
UNET	0.3341	0.3211	0.2376	0.6836	0.3242	0.38
NNUNET	0.3833	0.3815	0.2228	0.6928	0.3619	0.48
SAC	0.0037	0	0.0082	0.1247	0.0067	0.01
LSTMUNET	0.2889	0.3163	0.206	0.6424	0.2584	0.45
MAUNET_RESNET50	0.5685	0.5993	0.2452	0.7193	0.5722	0.58
MAUNET_WIDE	0.5561	0.5978	0.2492	0.7303	0.5445	0.59
MAUNET_ENSEMBLE	0.6015	0.6471	0.264	0.7108	0.6654	0.56

Best Performing Model: MAUNET_ENSEMBLE (Mean F1: 0.6015)

Performance Visualizations

Overall Performance Comparison

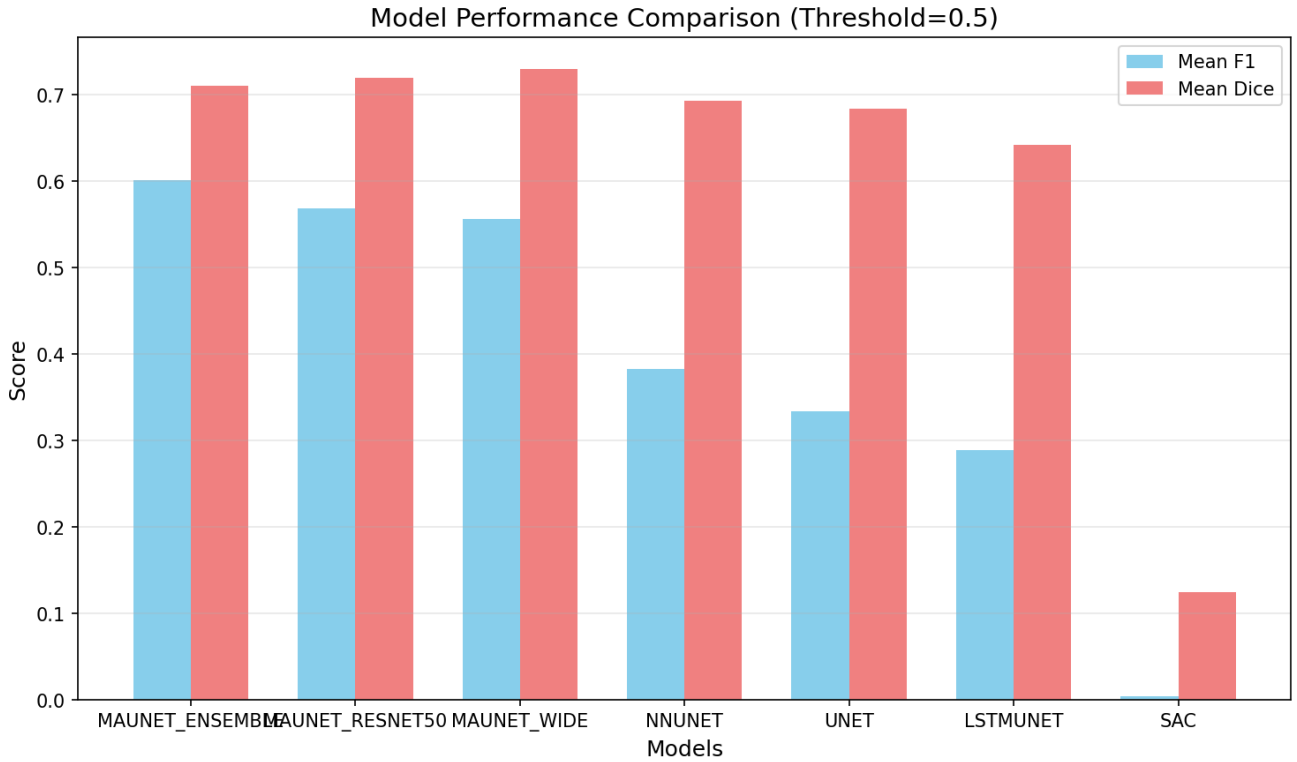


Figure 1: Comprehensive performance comparison across all metrics (F1, Dice, Precision, Recall)

F1 Score Across Different IoU Thresholds

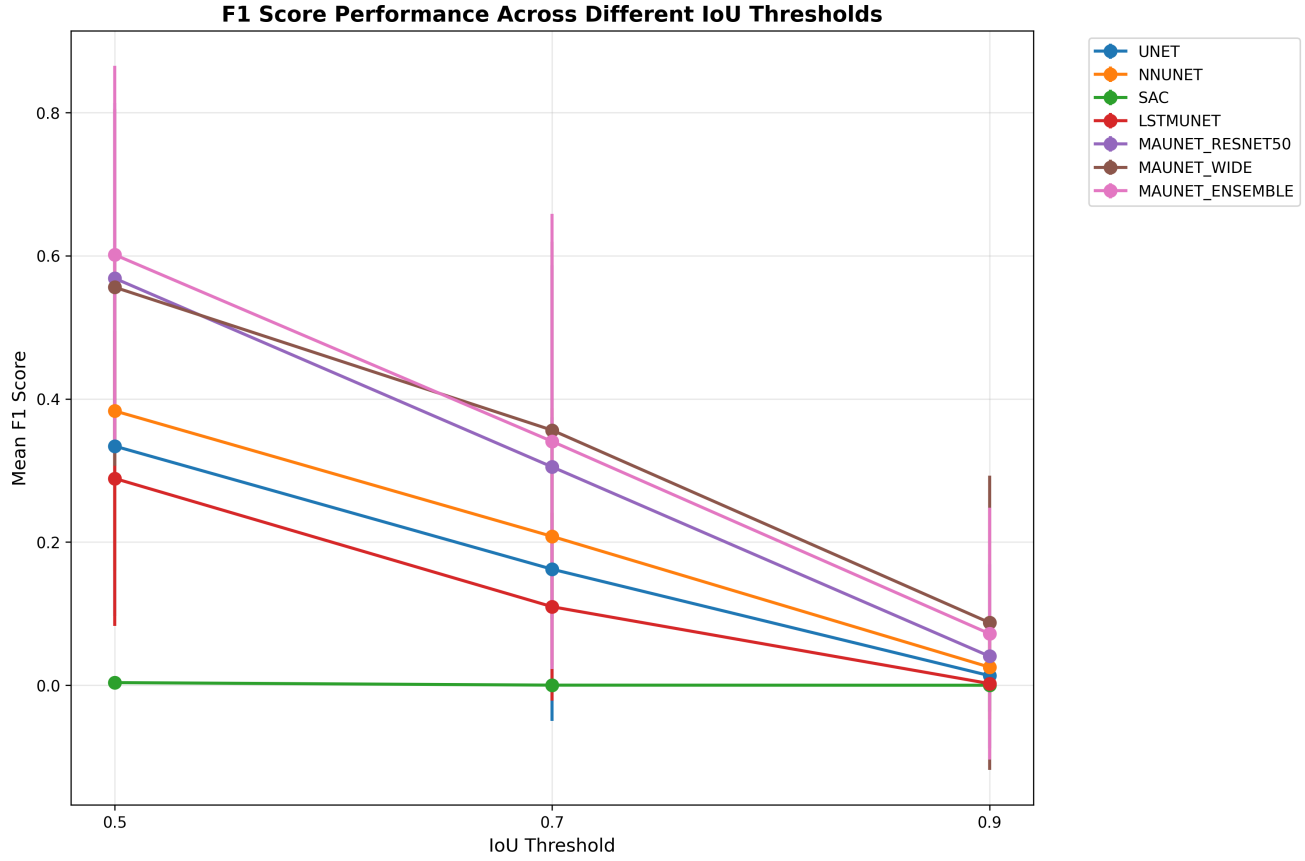


Figure 2: F1 Score performance at different IoU thresholds (0.5, 0.7, 0.9) showing model robustness

Training Information Comparison

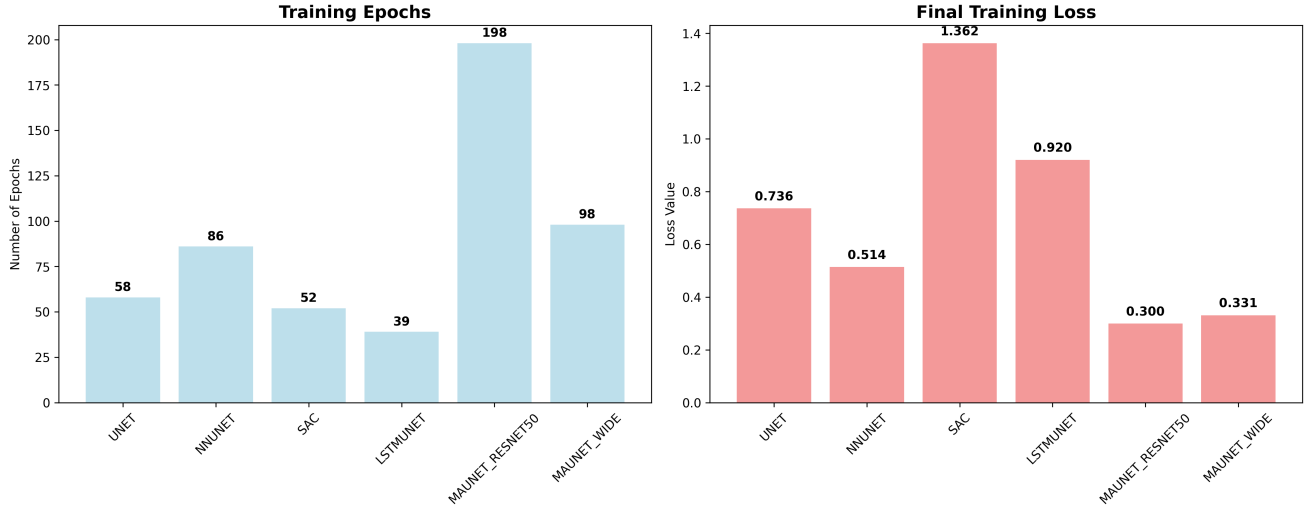


Figure 3: Training epochs and final loss comparison across models

Training Information

Model Sources and Repositories

- **UNET**: MONAI Framework - Built-in MONAI implementation
- **NNUNET**: MIC-DKFZ - https://github.com/mic-dkfz/nnunet
- **SAC**: 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
- **MAUNET_WIDE**: NeurIPS 2022 Challenge - https://github.com/Woof6/neurips22-cellseg_saltfish
- **MAUNET_ENSEMBLE**: NeurIPS 2022 Challenge (Ensemble) - https://github.com/Woof6/neurips22-cellseg_saltfish

Model Architectures and Training Parameters

Model	Architecture	Source	Repository
UNET	U-Net with ResNet blocks	MONAI Framework	Built-in MONAI implementation
NNUNET	nnU-Net (No New U-Net)	MIC-DKFZ	https://github.com/mic-dkfz/nn
SAC	Segment Anything + Custom Head	Authors via Email	Code provided by authors via e
LSTMUNET	U-Net with LSTM layers	GitLab - shaked0	https://gitlab.com/shaked0/lstm
MAUNET	MAU-Net with ResNet50 backbone	NeurIPS 2022 Challenge	https://github.com/Woof6/neuri cellseg_saltfish
MAUNET_WIDE	MAU-Net with Wide-ResNet50 backbone	NeurIPS 2022 Challenge	https://github.com/Woof6/neuri cellseg_saltfish
MAUNET_ENSEMBLE	MAU-Net Ensemble (ResNet50 + Wide-ResNet50)	NeurIPS 2022 Challenge (Ensemble)	https://github.com/Woof6/neuri cellseg_saltfish

Training Summary

- **Total Models Trained**: 6
- **Most Epochs**: 198 (MAUNET)
- **Best Training Validation Dice**: 0.6744 (NNUNET)
- **Optimizer**: AdamW (all models)

- **Learning Rate**: 6e-4 (all models)

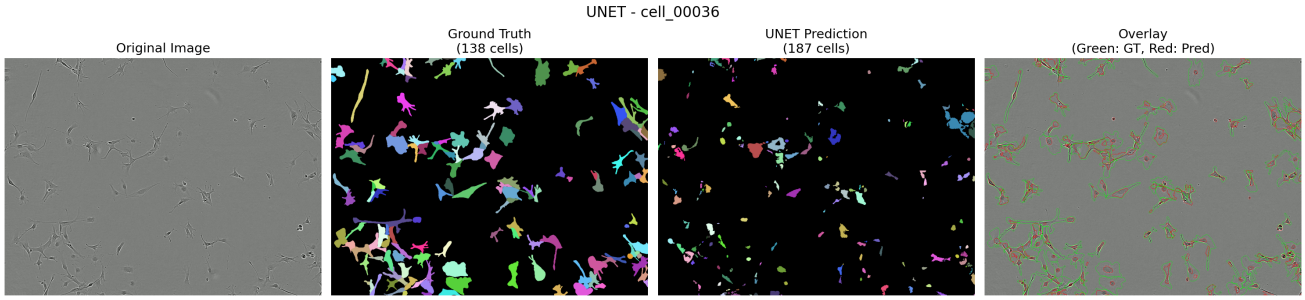
Sample Segmentation Results

Qualitative Comparison

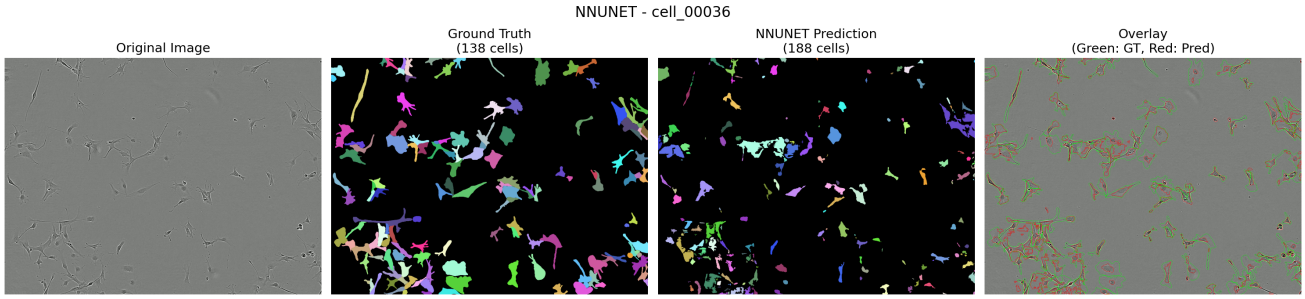
The following images show qualitative comparisons between ground truth and model predictions:

Sample 1: cell_00036

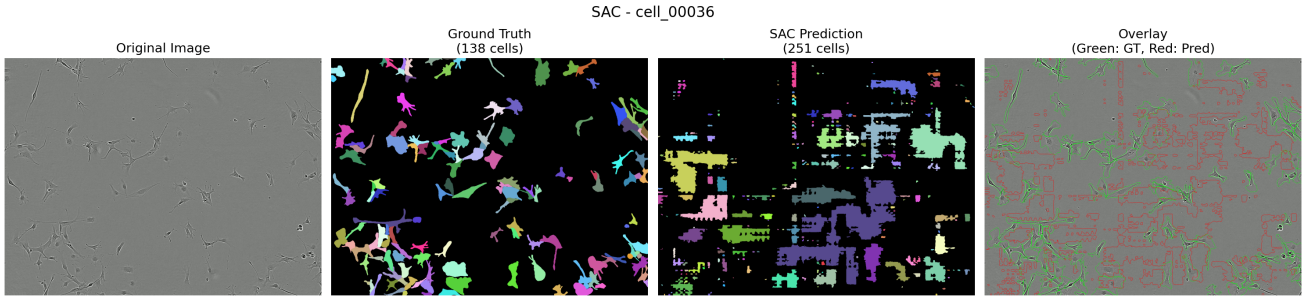
UNET:



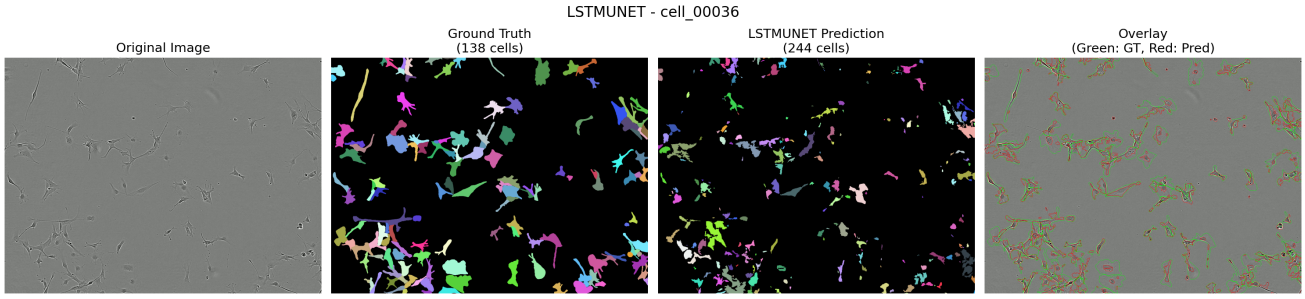
NNUNET:



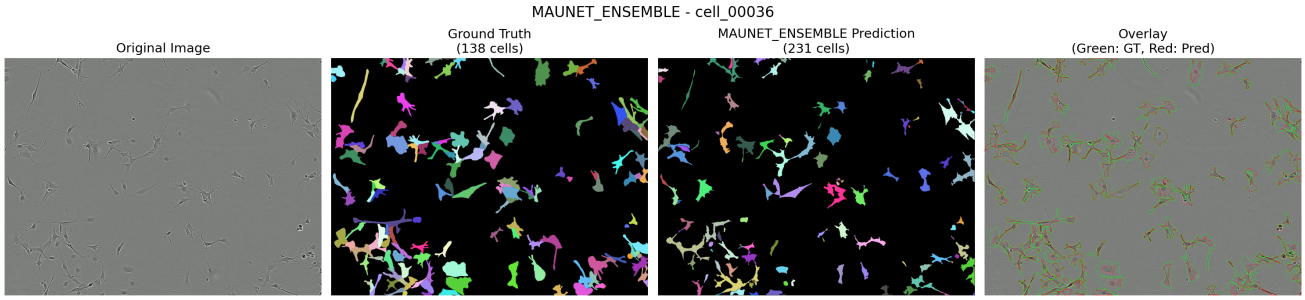
SAC:



LSTMUNET:

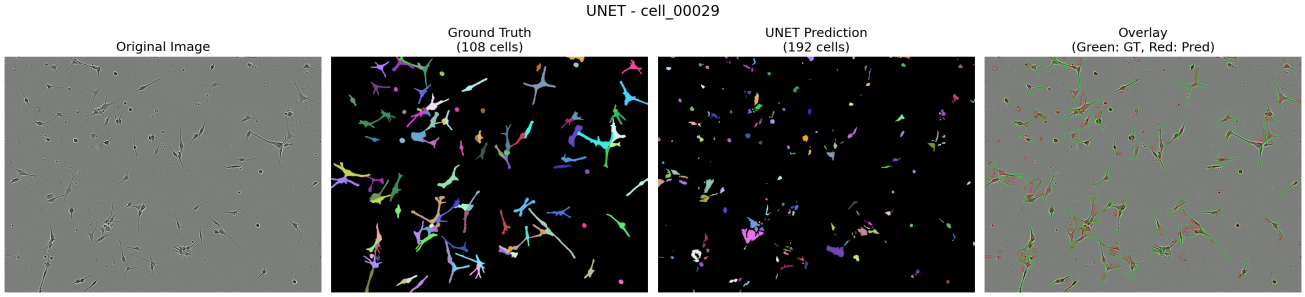


MAUNET_ENSEMBLE:

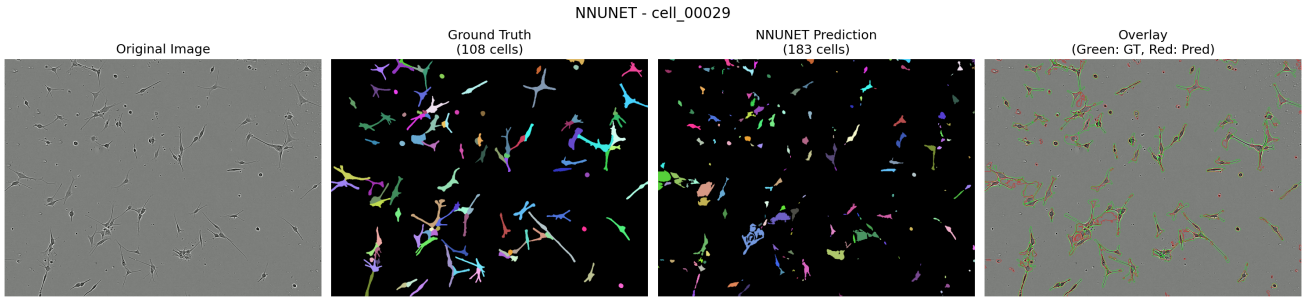


Sample 2: cell_00029

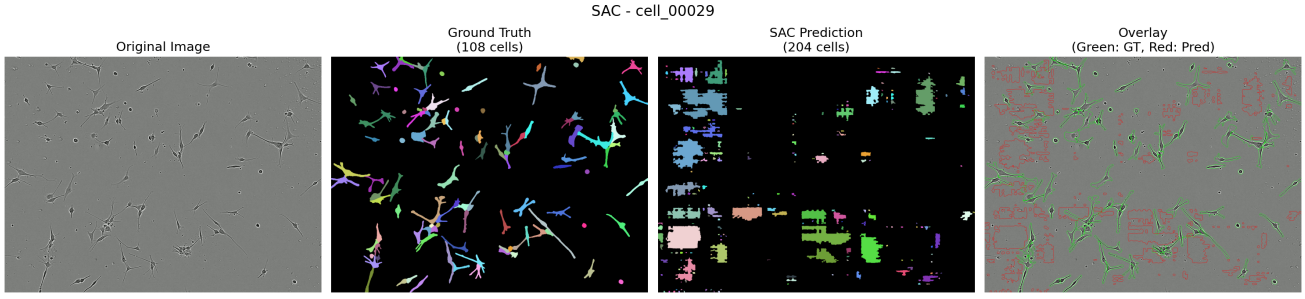
UNET:



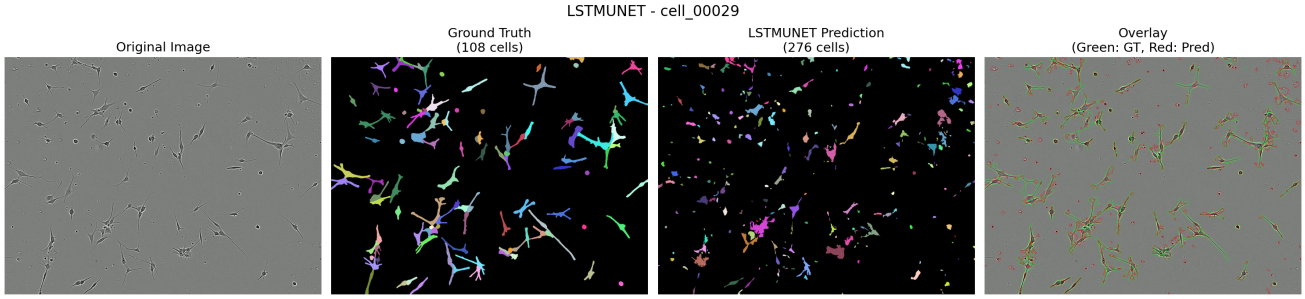
NNUNET:



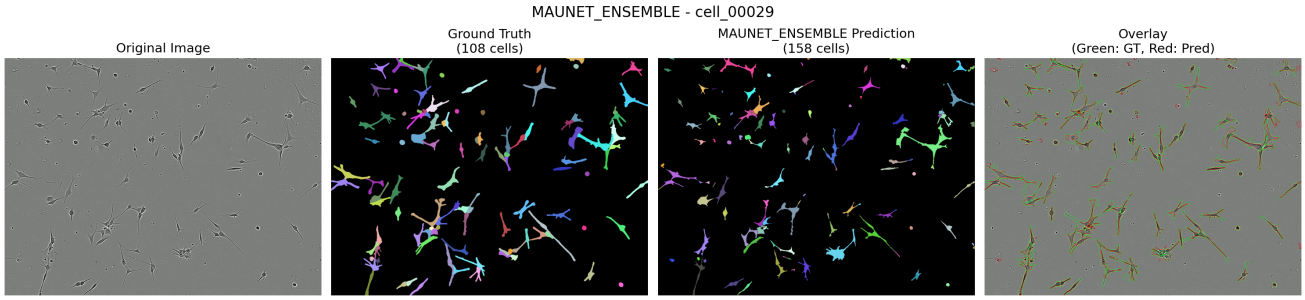
SAC:



LSTMUNET:



MAUNET_ENSEMBLE:



Detailed Results by Model

UNET

Performance Across Thresholds

Threshold	Mean F1	Mean Dice	Mean Precision	Mean Recall	Total Samples
0.5	0.3341	0.6836	0.3242	0.3854	101
0.7	0.1621	0.6836	0.1566	0.1863	101
0.9	0.0134	0.6836	0.0117	0.0168	101

Top 5 Performing Images

names	F1	dice
cell_00015_label.tiff	1	0.918
cell_00011_label.tiff	0.9655	0.8504
cell_00017_label.tiff	0.9167	0.8553
cell_00005_label.tiff	0.9	0.8749
cell_00009_label.tiff	0.8803	0.9022

Bottom 5 Performing Images

names	F1	dice
cell_00076_label.tiff	0	0.807
cell_00077_label.tiff	0	0.6852
cell_00036_label.tiff	0.0074	0.3393
cell_00072_label.tiff	0.0086	0.2054
cell_00100_label.tiff	0.0099	0.063

NNUNET

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	0.3833	0.6928	0.3619	0.4808	101
0.7	0.208	0.6928	0.1919	0.2818	101
0.9	0.0252	0.6928	0.0216	0.0423	101

Top 5 Performing Images

names	F1	dice
cell_00009_label.tiff	0.9159	0.9237
cell_00098_label.tiff	0.84	0.8228
cell_00005_label.tiff	0.8116	0.8649
cell_00094_label.tiff	0.7857	0.8157
cell_00085_label.tiff	0.7356	0.7993

Bottom 5 Performing Images

names	F1	dice
cell_00076_label.tiff	0	0.7035
cell_00100_label.tiff	0.0027	0.0217
cell_00077_label.tiff	0.0032	0.5186
cell_00041_label.tiff	0.0133	0.0861
cell_00078_label.tiff	0.0203	0.6948

SAC

Performance Across Thresholds

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

names	F1	dice
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	0.2889	0.6424	0.2584	0.4549	101
0.7	0.1097	0.6424	0.0924	0.2105	101
0.9	0.0023	0.6424	0.0016	0.0266	101

Top 5 Performing Images

names	F1	dice
cell_00009_label.tiff	0.8426	0.8672
cell_00085_label.tiff	0.66	0.8377
cell_00098_label.tiff	0.6364	0.8117
cell_00086_label.tiff	0.6316	0.8307
cell_00096_label.tiff	0.6296	0.78

Bottom 5 Performing Images

names	F1	dice
cell_00076_label.tiff	0	0.3965
cell_00077_label.tiff	0	0.7563
cell_00078_label.tiff	0	0.6792
cell_00070_label.tiff	0.0049	0.0597

names	F1	dice
cell_00074_label.tiff	0.0058	0.5468

MAUNET_RESNET50

Performance Across Thresholds

Threshold	Mean F1	Mean Dice	Mean Precision	Mean Recall	Total Samples
0.5	0.5685	0.7193	0.5722	0.5803	101
0.7	0.305	0.7193	0.3001	0.3188	101
0.9	0.0405	0.7193	0.0381	0.0443	101

Top 5 Performing Images

names	F1	dice
cell_00005_label.tiff	1	0.9237
cell_00009_label.tiff	0.9762	0.9364
cell_00015_label.tiff	0.9697	0.9088
cell_00004_label.tiff	0.9677	0.9397
cell_00011_label.tiff	0.9655	0.8807

Bottom 5 Performing Images

names	F1	dice
cell_00074_label.tiff	0.0063	0.0268
cell_00078_label.tiff	0.0161	0.5046
cell_00077_label.tiff	0.0241	0.6233
cell_00076_label.tiff	0.0357	0.7116
cell_00100_label.tiff	0.092	0.4295

MAUNET_WIDE

Performance Across Thresholds

Threshold	Mean F1	Mean Dice	Mean Precision	Mean Recall	Total Samples
0.5	0.5561	0.7303	0.5445	0.5985	101
0.7	0.3563	0.7303	0.3408	0.3945	101
0.9	0.0875	0.7303	0.0799	0.1014	101

Top 5 Performing Images

names	F1	dice
cell_00005_label.tiff	1	0.955
cell_00011_label.tiff	1	0.912
cell_00015_label.tiff	1	0.9462
cell_00009_label.tiff	0.9635	0.9423
cell_00004_label.tiff	0.9375	0.9465

Bottom 5 Performing Images

names	F1	dice
cell_00078_label.tiff	0.0034	0.4541
cell_00077_label.tiff	0.0264	0.5097
cell_00099_label.tiff	0.0287	0.3119
cell_00074_label.tiff	0.0334	0.097
cell_00100_label.tiff	0.0362	0.2966

MAUNET_ENSEMBLE

Performance Across Thresholds

Threshold	Mean F1	Mean Dice	Mean Precision	Mean Recall	Total Samples
0.5	0.6015	0.7108	0.6654	0.5638	101
0.7	0.3407	0.7108	0.3743	0.3206	101
0.9	0.072	0.7108	0.0779	0.0688	101

Top 5 Performing Images

names	F1	dice
cell_00004_label.tiff	1	0.9514
cell_00011_label.tiff	1	0.8979
cell_00015_label.tiff	1	0.9279
cell_00042_label.tiff	1	0.9004
cell_00070_label.tiff	1	0.9474

Bottom 5 Performing Images

names	F1	dice
cell_00074_label.tiff	0	0

names	F1	dice
cell_00078_label.tiff	0.0036	0.3769
cell_00100_label.tiff	0.0081	0.0341
cell_00077_label.tiff	0.0201	0.4121
cell_00076_label.tiff	0.0351	0.7876

Dataset Analysis

Performance vs. Ground Truth Cell Count

UNET - Performance by Cell Count

GT Count Range	mean	count
1-5	nan	0
6-10	0.418667	3
11-20	0.353818	17
21-50	0.452732	22
51-100	0.2258	11
100+	0.26601	29

NNUNET - Performance by Cell Count

GT Count Range	mean	count
1-5	nan	0
6-10	0.399233	3
11-20	0.377288	17
21-50	0.469936	22
51-100	0.371818	11
100+	0.30761	29

SAC - Performance by Cell Count

GT Count Range	mean	count
1-5	nan	0
6-10	0	3
11-20	0.00850588	17
21-50	0.00404545	22

GT Count Range	mean	count
51-100	0	11
100+	0.00294138	29

LSTMUNET - Performance by Cell Count

GT Count Range	mean	count
1-5	nan	0
6-10	0.144467	3
11-20	0.182706	17
21-50	0.350159	22
51-100	0.286691	11
100+	0.289928	29

MAUNET_RESNET50 - Performance by Cell Count

GT Count Range	mean	count
1-5	nan	0
6-10	0.408133	3
11-20	0.679341	17
21-50	0.675041	22
51-100	0.444318	11
100+	0.524207	29

MAUNET_WIDE - Performance by Cell Count

GT Count Range	mean	count
1-5	nan	0
6-10	0.385	3
11-20	0.624594	17
21-50	0.650927	22
51-100	0.461609	11
100+	0.512693	29

MAUNET_ENSEMBLE - Performance by Cell Count

GT Count Range	mean	count
1-5	nan	0

GT Count Range	mean	count
6-10	0.666667	3
11-20	0.750794	17
21-50	0.737732	22
51-100	0.471982	11
100+	0.544072	29

Model Comparison

Metrics Comparison (Threshold = 0.5)

Metric	UNET	NNUNET	SAC	LSTMUNET	MAUNET_RESNET50
F1 Score	0.3341	0.3833	0.0037	0.2889	0.5685
Dice Score	0.6836	0.6928	0.1247	0.6424	0.7193
Precision	0.3242	0.3619	0.0067	0.2584	0.5722
Recall	0.3854	0.4808	0.0107	0.4549	0.5803

Statistical Significance Analysis

F1 Score Descriptive Statistics

Model	Mean ± Std	Median	95% CI	Range
UNET	0.3341 ± 0.2364	0.3211	[0.0080, 0.9083]	[0.0000, 1.0000]
NNUNET	0.3833 ± 0.2217	0.3815	[0.0083, 0.7986]	[0.0000, 0.9159]
SAC	0.0037 ± 0.0081	0.0000	[0.0000, 0.0289]	[0.0000, 0.0440]
LSTMUNET	0.2889 ± 0.2050	0.3163	[0.0024, 0.6340]	[0.0000, 0.8426]
MAUNET_RESNET50	0.5685 ± 0.2440	0.5993	[0.0299, 0.9687]	[0.0063, 1.0000]
MAUNET_WIDE	0.5561 ± 0.2480	0.5978	[0.0311, 0.9818]	[0.0034, 1.0000]
MAUNET_ENSEMBLE	0.6015 ± 0.2626	0.6471	[0.0141, 1.0000]	[0.0000, 1.0000]

F1 Score ANOVA Results

- **F-statistic**: 92.3365
- **p-value**: 0.000000
- **Significant**: Yes

F1 Score Kruskal-Wallis Results

- **H-statistic**: 346.0892
- **p-value**: 0.000000
- **Significant**: Yes

Key Pairwise Comparisons (F1 Score)

Comparison	T-test p-value	Mann-Whitney p-value	Cohen's d	Effect Size
NNUNET VS MAUNET ENSEMBLE	0.000000	0.000000	-0.8934	Large
UNET VS MAUNET ENSEMBLE	0.000000	0.000000	-1.0647	Large

Dice Score Descriptive Statistics

Model	Mean ± Std	Median	95% CI
UNET	0.6836 ± 0.1902	0.7344	[0.1772, 0.9275]
NNUNET	0.6928 ± 0.1616	0.7364	[0.2346, 0.8939]
SAC	0.1247 ± 0.1258	0.0723	[0.0000, 0.3986]
LSTMUNET	0.6424 ± 0.1695	0.6792	[0.0794, 0.8576]
MAUNET_RESNET50	0.7193 ± 0.1395	0.7399	[0.4458, 0.9321]
MAUNET_WIDE	0.7303 ± 0.1553	0.7585	[0.3043, 0.9443]
MAUNET_ENSEMBLE	0.7108 ± 0.1878	0.7543	[0.2096, 0.9423]

Recommendations

Based on the benchmarking results:

1. **MAUNET_ENSEMBLE** shows the best overall performance with highest mean F1 score
2. Consider using threshold = 0.5 for optimal balance between precision and recall
3. Models perform better on images with moderate cell counts (10-50 cells)
4. Further training or fine-tuning may improve performance on densely populated images