Evaluating UNET and Mask R-CNN models for their generalizability with multimodal image datasets and detecting nuclei in tumours

Course: MBP1413H
Authors:

Ajay Singh[1,2]
ajay.singh@mail.utoronto.ca

Yakup Kohen[1,3]
yakup.kohen@mail.utoronto.ca

1) Department of Medical Biophysics, University of Toronto
Toronto, ON M5G 1L7. Canada

2) Biological Sciences Platform, Sunnybrook Research Institute
Toronto, ON M4N 3M5. Canada

3) Princess Margaret Cancer Research Tower, University Health Network
Toronto, ON M5G 0A3. Canada

Abstract

Microscopy and image acquisition of a cell's nuclei are powerful tools for advancing our understanding of cellular behaviors, serving as a cornerstone in cell biology and biomedical research. The advent of automated microscopy across various image modalities has led to an increased abundance of nuclei image data. Consequently, the need for automated nuclei segmentation has become essential for subsequent analyses. This study addresses the critical challenge of segmenting nuclei by leveraging deep learning models. A focus on enhancing the efficiency and accuracy of segmentation was performed through the fitting and optimization of UNET and Mask R-CNN models. Through the applications of these models on a multimodal image dataset containing nuclei images acquired via fluorescence microscopy, as well as darkfield and brightfield microscopy of Hematoxilin and Eosin (H&E) staining, model performance metrics were compared across hyperparameter-tuned model architectures. Our testing results initially indicated that with a small dataset, UNET had a tendency to overfit. After tuning the learning parameters, it performed better learning sessions but lacked stability. To obtain a more stable segmentation, we then utilized Mask R-CNN, which showed comparable validation performance but more stable learning. Here, we demonstrate each model pipeline could be viable for segmentation usage in a diverse set of images, but both models require further training, tuning, and data for a more stable learning and generalizability. Overall, our findings hold the promise of streamlining the initial, yet crucial step in cell image analysis at a multimodal level.

GitHub Code Repository: https://github.com/ajaysingh096/2024_MBP1413_Ajay_Yakup.git

Introduction

With advancements in methodologies used to understand human diseases and their effective treatments, modern medical and cell-biological research continues to utilize the age-old technique of examining cells under a microscope to uncover answers. Particular methodologies of microscopy imaging, such as fluorescence microscopy and brightfield imaging of H&E staining upon cells have provided a targetted way at identifying subcellular structures with great detection sensitivity and specificity[1,2]. Certain structures highlighted by such imaging modalaties, such as the nucleus, can be a critical indicator of cell functionality and even pathogenicity[3]. Detailed insights into nuclear morphology provide invaluable information for the diagnosis and treatment of diseases[3].

Further advancements in the automation of microscopy has been transformative, allowing for the rapid acquisition of cellular images across fluorescence and brightfield imaging, leading to large datasets abundant with images. Although a combination of multimodal images can tackle features and aspects of the cell at multiple angles to complement eachother, these datasets are only as valuable as the insights they yield from their subsequent analyses performed upon them. Thus, the need for precise and automated object detection and segmentation of nuclei is of high value for accurate and meaningful analyses. Effective segmentation serves as a precursor to numerous downstream analyses, ranging from quantitative cell counting to phenotypic profiling and morphological classification[4]. It is also inherently vital in the study of cellular responses to pharmacological treatments or genetic alterations[5,6].

The segmentation of nuclei automatically, however, faces challenges. As manual segmentation can be time-consuming, classical methods of segmentation may not compensate for

this, as they may yield innacuracies especially when handling complex datasets containing variation in imaging conditions[7]. Images can exhibit wide variability in contrast, resolution, and staining quality, making the task of nuclei segmentation exceptionally complex. As such, an algorithm capable of generalizing across such variability without loss of accuracy and efficiency is needed.

In response to these challenges, we have comprehensively evaluated and compared the application of two deep learning architectures upon a mixed dataset comprised of fluorescence and brightfield H&E microscopy images of cells from the "*2018 Data Science Bowl*"[8]. These models are UNET and Mask R-CNN, which offer a renowned foundation for the applicaiton of segmentation that we have tailored specifically towards analyzing nuclei from the mixed dataset. In particular, The UNET model used in this study is called "Unet" and is implemented into and derived from the usage of *monai.networks.nets.unet*[9]. The Mask R-CNN model used is derived from *Detectron2* from *Facebook AI Research*[10]. UNET is an architecture designed specifically for biomedical image segmentation, consists of a contracting path to capture context and a symmetric expanding path that enables precise localization[15]. While Mask R-CNN, an extension of Faster R-CNN, is augmented with an additional mask predicting branch that adds to it's segmentation performance by identifying a region of interest, which is useful for images with multiple objects[11,12]. Both UNET and Mask R-CNN has been used successfully in nuclei detection but heterogenous microscopy data remains a challenge for both architectures[16,17].

We assess the built-in capabilities of these models through supervised learning, where the models are trained on labels provided in the the training dataset to learn the mapping between input data and output labels and enhance the predictive accuracy for unseen data. By advancing our ability to segment nuclei with high precision and reliability, we provide a robust framework

that can be adapted to a variety of imaging conditions which will then facilitate advancements in both biological understanding and medical discovery.

Results

Heterogeneous microscopy dataset containing labeling errors

The dataset used for training the models were obtained from *2018 Data Science Bowl* on Kaggle, containing diverse images of cellular nuclei using multiple microscopy techniques such as bright field, dark field, and fluorescence microscopy, along with H&E stained images. Incidentally, numerous competitors on Kaggle have reported error in labeling of the images[13,14]. In fact, we have discovered such an error via segmentation (Figure 2B). While we were not able to assess the impact of the erroneous labels, we suspect it might have effected the training process, as depicted in Figure 2, where nuclei have been labelled beyond normal shape. Nevertheless, the diverse nature of the data allowed the models to demonstrate their generalizability.

UNET model struggled with overfitting

In a normal training model, Dice loss; the difference in the intersection over union overlap between two sets, tends decrease when progressing through epochs. In contrast, the mean Dice coefficient; the similarity between two sets, should increase, indicating an increased similarity between prediction and label with learning. However, we trained the UNET model with approximately 450 images and had signs of overfitting, including an innitial increase in mean Dice coefficient followed by a stable decrease in Dice loss[17] (Fig 1).

UNET models generally struggled with background heterogeneity

Various UNET models with certain degree of overfitness were evaluated. One model with Epoch 2 (highest dice coefficient) with approximately 0.35 mean Dice coefficient was then used to visually inspect the model success in diverse images, including a black background microscopy image and a white background microscopy image (Fig2). While certain training models with better Dice coefficient (0.94) showed poor white background nuclei segmentation (Fig.2C), this model still gave an acceptable segmentation output visually (Fig. 2B). Although, some error can be seen. It is notable that the ground truth label and input image do not match perfectly, with large masses of nuclei not being able to be visually seen on the image but present on the label for (Fig 2B). Interestingly, some of the visually undeteclable nuclei on the original were also somewhat segmented on output. Nevertheless, such an example highlights the error with some of the labels in the data, which might have contributed to poor results.

UNET models performed well on a separate test dataset

While evaluating the final UNET model with novel parameters (see methods section), Dice loss showed a downward trend, with high Dice coefficient in early epochs. (Fig.3). While the dice coefficient fluctuated, possibly indicating that the model is not stable, visual inspection showed relatively successful segmentation on variety of cases using the model based on Epoch 2 (Figure 4). In fact, when tested with the locked dataset of 67 images, which have not been shown to the model during training or validation and that was randomly selected and separated from the Kaggle dataset, we obtained a Dice coefficient of 0.71, indicating that the model performed well on unseen data. (Fig 4A). Furthermore, the model could segment both bright background images

such as H&E-stained and dark background images with high dynamic range of intensities relatively successful compared to previous training models. (Fig. 4B and C). However, considering the instability of the learning and tendency to overfit the model to data with UNET, we also wanted to investigate segmentations with another network architecture, Mask R-CNN.

Mask R-CNN model lead to stable evaluation scores

The same amount of images that were trained upon for UNET was also used for Mask R-CNN, with that number being 450 images which was constituted from the same 70-30 training and testing split fraction. Dice scores were recorded at every 2000 iterations, which showed a consistent mean Dice coefficient of around 0.8 after the first checkpoint (Fig 5A). As well, a consistent decrease in average Dice loss over iterations was seen (Fig 5B). These results indicate that the Mask R-CNN model is learning stably and correctly with improvements over iterations and without overfitting.

Visual comparisons between trained UNET and MASK R-CNN models

When visually comparing segmentation outputs from the trained UNET and Mask R-CNN models for brightfield and fluorescence images, there is an indication that both of the models perform well in certain aspects, but also contain major pitfalls. Overall, the UNET model performed very well at object detection upon both brightfield and fluorescence images, highlighting all potential nuclei per image, but fails to create specific discernable segmentations of each nucleus in some instances (Fig 6A, B, D, and E). The performance of the Mask R-CNN model resulted in more discernible and segmentation of each nuclei throughout the dataset, although it's sensitivity with overall detection of nuclei had lacked, as many nuclei were

particularly missed in brightfield images (Fig 6A, C, D, and F). Confidence scores for the binary classification of nuclei were also implemented and reported for the Mask R-CNN model's performance based on bounding box regression, which showed a higher confidence score per nucleus in fluorescence images, versus a lower score for brightfield images (Fig 6C and F). This points towards a lack of generalizability of the model over multimodal images.

Evaluation of trained UNET and Mask R-CNN models from the training dataset

A comparison between the best and worst evaluation scores determined from performance on the training dataset can be seen from Table 1. Notably, at their best iteration/epoch checkpoints, the Mask R-CNN model had a higher Dice score compared to the UNET model, with their values being 0.811 and 0.798 respectively. Although the the higher Dice score provided by Mask R-CNN is at the highest iteration checkpoint, it is important to note that Dice score values were near-consistent throughout training. Both architectures had similar lowest Dice loss value. (Table 1).

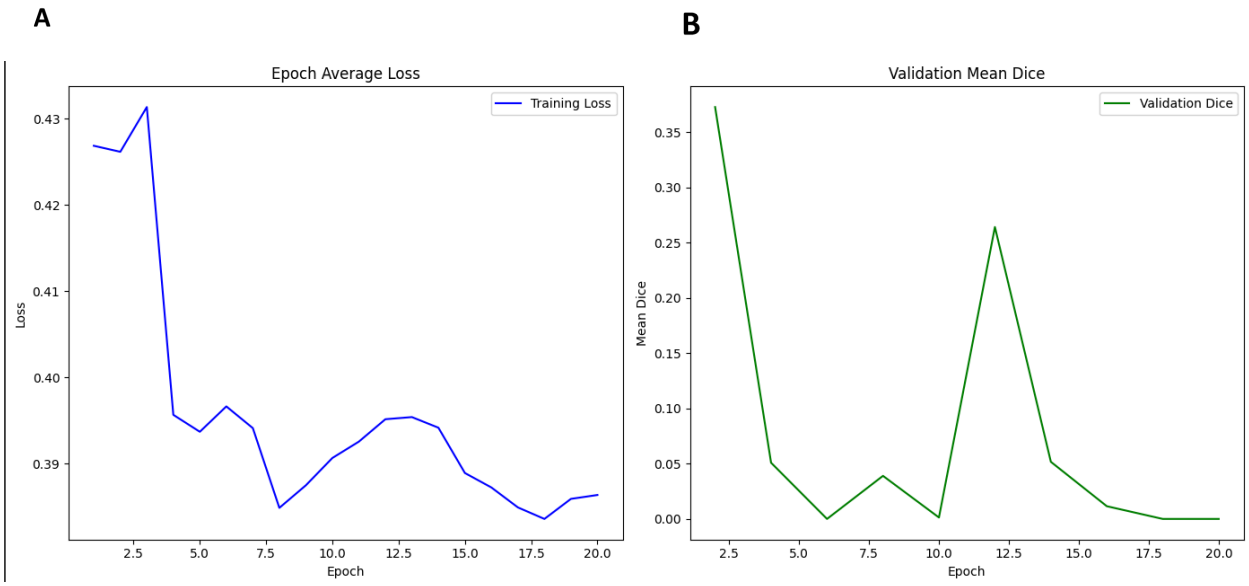

Figure 1: Signs of overfitting during UNET training. **a)** Average Dice loss per epoch during training. **b)** Mean Dice coefficient per 2 epochs on validation.

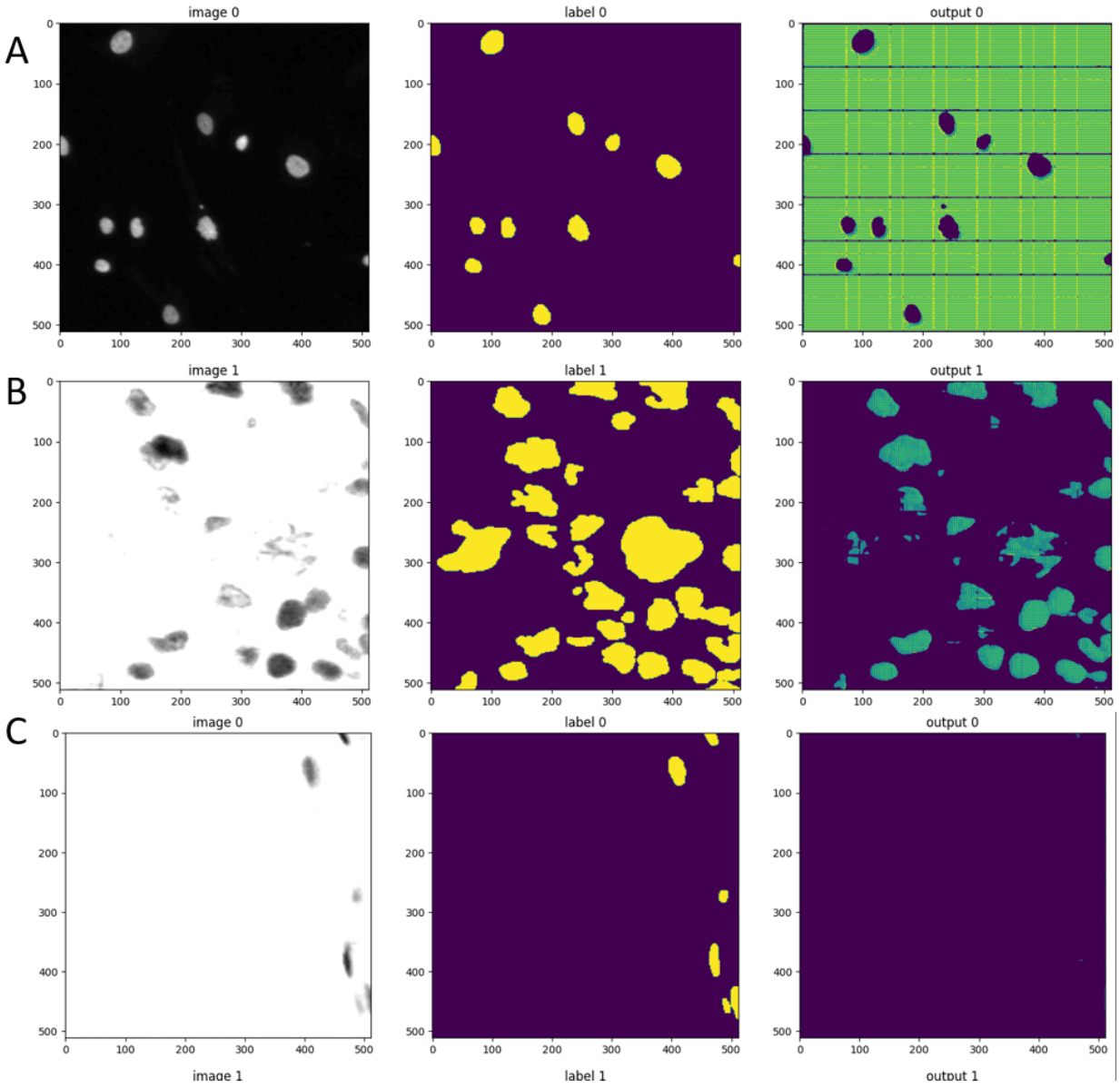

Figure 2: Example of poor learning from UNET model with heterogenous backgrounds on validation set. a) Segmentation of a black background microscopy image. **b)** Segmentation of a white background microscopy image

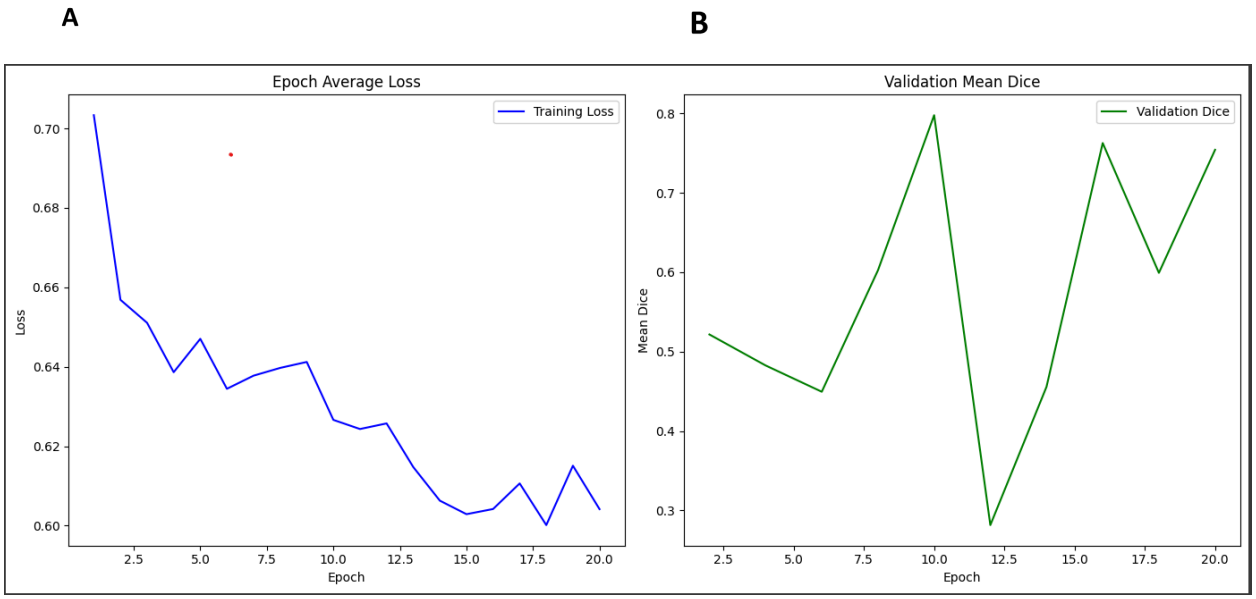

Figure 3: Final UNET model with a Dice coefficient of >0.8. a) Average Dice loss per epoch during training. **b)** Mean Dice coefficient per 2 epochs on validation.

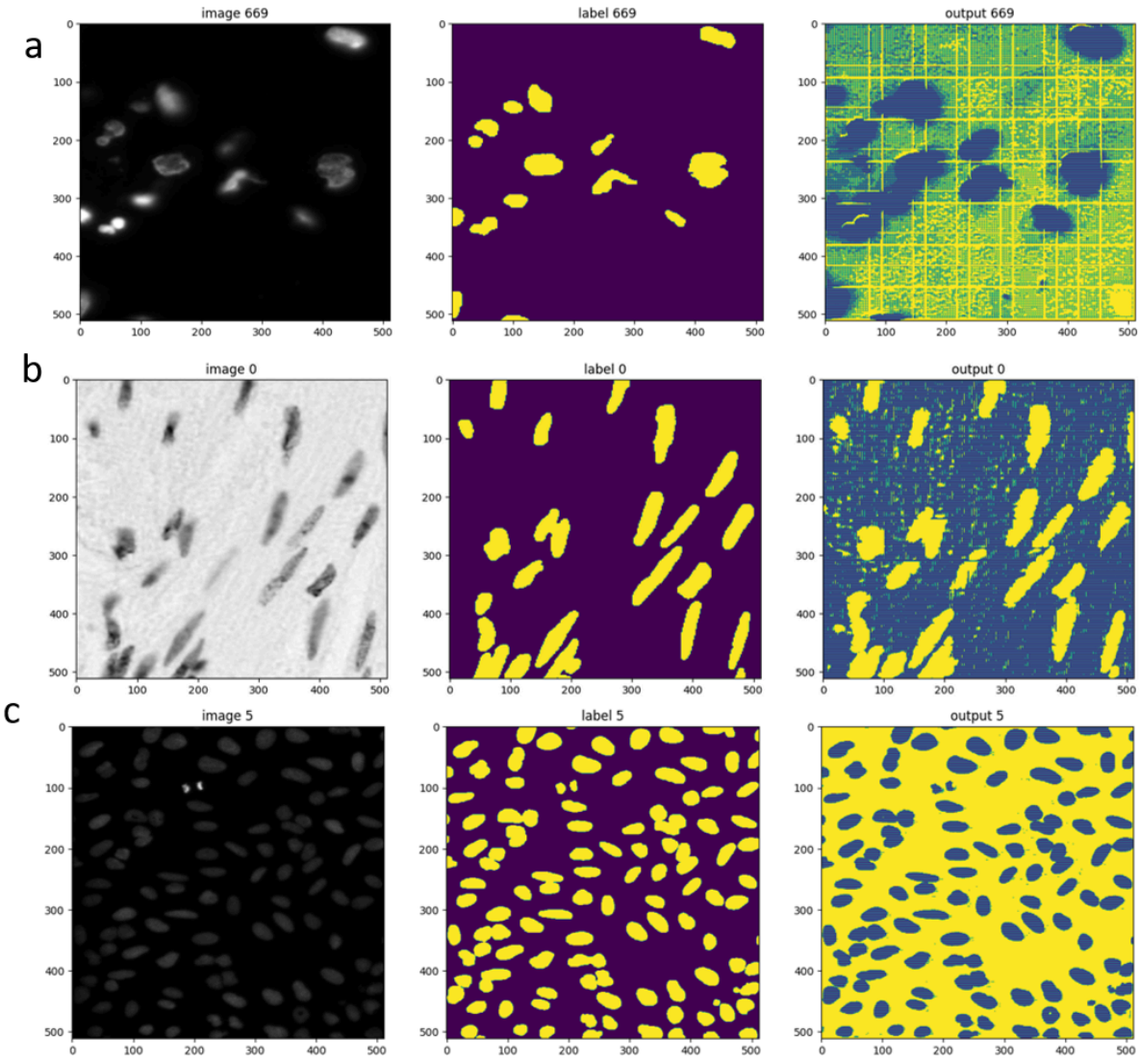

Figure 4: Final UNET model example segmentations. a) An image from the locked testing set, previously unseen to the model, Dice coefficient 0.82. **b)** An H&E staining image from the training dataset. **c)** A fluorescent microscopy image segmentation with high intensity spots from the training dataset.

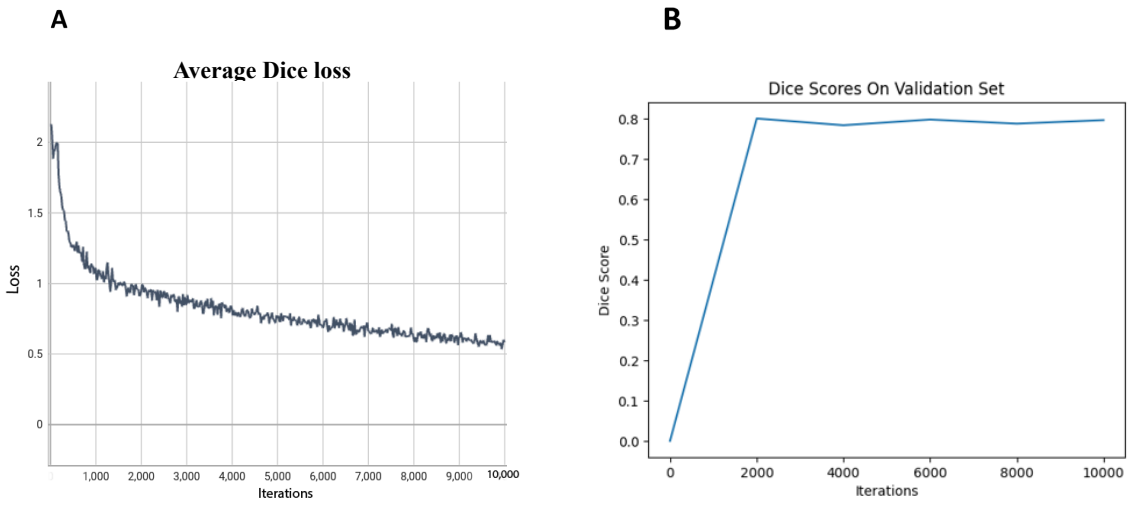

Figure 5: Performance metrics for Mask R-CNN. a) Average Dice loss per iteration during training. **b)** Mean Dice coefficient per 2000 interactions on validation.

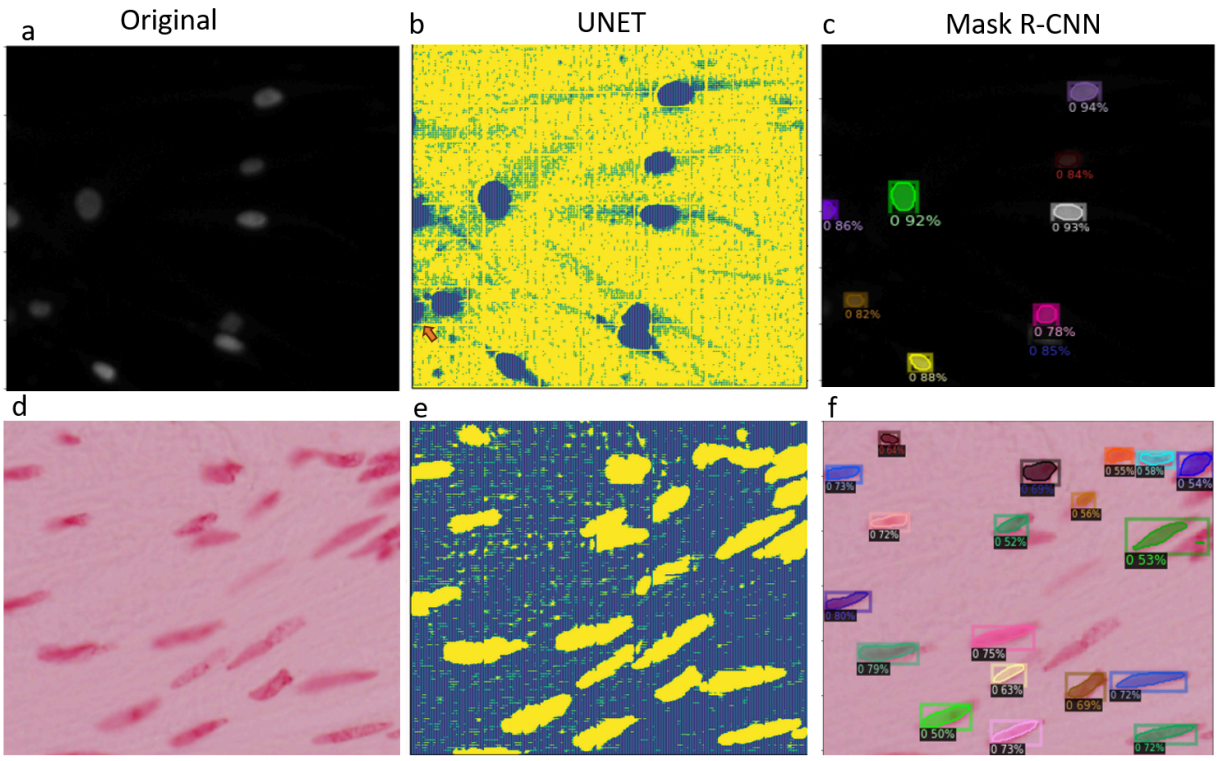

Figure 6: Comparison of image segmentations between UNET and Mask R-CNN models. **a)** and **d)** original input images. **b)** and **e)** UNET-generated segmentation. Red arrow on b indicates a non-existing nuclei detected by UNET. **c)** and **f)** Mask R-CNN-generated segmentations.

Model	**Best Dice score**	**Lowest Dice Loss score**	**Best Dice score: Iteration/ Epoch**	**Lowest Dice loss: Iteration/ Epoch**	**Loss at best Dice score**
Mask-RCNN	0.811	0.584	10,000	2000	0.930
UNET	0.798	0.600	10	18	0.627

Table 1: Comparison of Dice losses and Dice scores of the final models

Methods

UNET model

Input images and labels sizes were 512x512x3. Dataset for UNET training was not pre-processed, except transformation with Monai transform tools; ScaleIntensityRanged[for normalizing pixel values between 0-255 to be in range of 0-1), RandRotate90d. Using RandCropByPosNegLabeld training labels and images were cropped to 4x (96 x96) images to enhance training. Validation was performed on ROI size of (96x96).

Unet model was generated using monai.networks.nets.unet. 5 layer network with 3 input channels and 3 output channels with channel sizes of (4,8,16,32). The stride size for convolution was 2 and kernel size was 3. Residual unit was set to 2. Activation function was ReLu. Loss function was set as DiceLoss from monai.lossess.DiceLoss, excluding background. Metric was DiceMetric from monai.metrics, reduction was set as mean. Learning rate was 0.01.

Model was trained on Google Collab CPU. More details and trained model can be found on github.

Mask R-CNN model

We trained a Mask-RCNN model based on the unbiasedteacher repo. Unfortunately we did not have the resources to create a semi-supervised model, which was our original plan. However, we used the detectron2 features in this repo, including dataset registration, config, logging and train-time validation to organize training and tune hyperparameters. We imported our config which contained important paths for I/O, controls training parameters, types of ROI heads (i.e. inclusion of masks). Then, we splited and registered our dataset based on our annotation json.

We used the RCNN training engine. Because we originally intended to have a semi-supervised model, the student/teacher models were still assembled into an ensemble, however only the student model was trained. For inference, we assembled the model as an ensemble despite the fact that the teacher model has not been trained. For our purposes (fully supervised), we only used the trained student model(ubteacher-rcnn). The confidence threshold was set to filter low confidence predictions. Finally, the image is converted to a torch-compatible tensor and our Mask-RCNN model undergoes a forward pass to create predictions. ROI threshold test was 0.5. Mask setting was ON and number of classes for detection was set as 1. Learning rate was 0.005, and iteration number was 10000 with returning metrics every 2000 iterations and evaluation every 1000.

Discussion

Overall, it can be visually seen that both model architectures of Unet and Mask R-CNN provide a robust performance for segmentation of the nucle. Mask R-CNN had a more stable learning performance, making it a more suitable choice for this application on this data set. It has previously been reported that both UNET and Mask R-CNN strugles with this dataset. Our observations were similar to the previous reports that Mask R-CNN performs better than UNET with this dataset[17]. When determining whether Unet or Mask R-CNN should be chosen for the purposes of segmenting nuclei from cells in a multimodal image dataset, the considerations extend beyond accuracy metrics. Factors such as computational efficiency, ease of integration, and adaptability to dataset variability must be weighed. While in this study we did not assess the computational loads of training these models, in the future, we would like to compare the computational burden of these architectures. We would also like to introduce strategic image

pre-proccessing modifications that can enhance their performance and compare model outputs and performance metrics before and after the application of such techniques. The pre-processing pipeline would encompass a range of techniques designed to homogenize and enhance image quality across image types, thereby facilitating more effective model learning as there may be less nuances for the models to learn.

The limited size and labeling errors of our dataset, comprising 1005 samples, introduces a significant challenge in terms of model training, particularly in the case of overfitting. This can lead to an unwanted development of a model that learns the training data too well, including its noise and outliers, at the expense of its ability to generalize to new data. We also do not know the composition of training and validation tests and their respective distribution of bright field and dark field images. Systematic investigation of dataset could be useful to ensure both training, validation, and testing sets have similar distribution of white-dark background images. Our study was also limited by the computational resources for the training the UNET model. Due to time and usage constrains on Google Collab, we were unable to extend the Epoch number to greater values. A longer training session might yield better results, which could be better for comparisons to Mask R-CNN. We also did not systematically assess different learning rates, loss functions, and other hyperparameters to fine tune the performance of the models. While this might be a limitation to see the true capacities of these models in this challenging task, we believe, such alterations would make it difficult to interpret the inert differences and strengths of these models with this task as there is no proper way of keeping the tuning consistent between the models. However, this is something that could be beneficial for investigation in the future, such as by using learning rate schedulers and adaptive learning rate methods[18,19].

In conclusion, our findings highlight the trade-offs inherent in model selection for tasks involving complex image segmentation. While Unet and Mask R-CNN models both show promise for their application upon the multimodal microscopy dataset, careful consideration must be given to dataset characteristics and the potential for overfitting, with a focus on including strategies to counteract these issues.

References

1. Achilefu S. 2010. Introduction to concepts and strategies for molecular imaging. *Chem Rev*. May 12;110(5):2575-8. doi: 10.1021/cr1001113.
2. Fischer AH, Jacobson KA, Rose J, Zeller R. 2008. Hematoxylin and eosin staining of tissue and cell sections. *CSH Protoc*. May 1;2008:pdb.prot4986. doi: 10.1101/pdb.prot4986.
3. Caicedo JC, Singh S, Carpenter AE. 2016. Applications in image-based profiling of perturbations. *Curr Opin Biotechnol.* Jun;39:134-142. doi: 10.1016/j.copbio.2016.04.003.
4. Caicedo JC, Goodman A, Karhohs KW, Cimini BA, Ackerman J, Haghighi M, Heng C, Becker T, Doan M, McQuin C, Rohban M, Singh S, Carpenter AE. 2020. Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. *Nat Methods.* Dec;16(12):1247-1253. doi: 10.1038/s41592-019-0612-7.
5. Krause J, Grabsch HI, Kloor M, Jendrusch M, Echle A, Buelow RD, Boor P, Luedde T, Brinker TJ, Trautwein C, Pearson AT, Quirke P, Jenniskens J, Offermans K, van den Brandt PA, Kather JN. 2021. Deep learning detects genetic alterations in cancer histology generated by adversarial networks. *J Pathol*. ;254(1):70-79. doi: https://doi.org/10.1002/path.5638
6. Nawabi AK, Jinfang S, Abbasi R, Iqbal MS, Heyat MBB, Akhtar F, Wu K, Twumasi BA. 2022. Segmentation of Drug-Treated Cell Image and Mitochondrial-Oxidative Stress Using Deep Convolutional Neural Network. Oxid Med Cell Longev. May 26;2022:5641727. doi: 10.1155/2022/5641727.
7. Dimopoulos S, Mayer CE, Rudolf F, Stelling J. 2014. Accurate cell segmentation in microscopy images using membrane patterns. Bioinformatics. 30(18):2644-2651. doi: https://doi.org/10.1093/bioinformatics/btu302
8. 2018 Data Science Bowl. Accessed March 31, 2024. https://kaggle.com/competitions/data-science-bowl-2018/discussion/56326
9. Kerfoot, E., Clough, J., Oksuz, I., Lee, J., King, A.P., Schnabel, J.A. 2019. Left-Ventricle Quantification Using Residual U-Net. In: Pop, M., et al. Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges. *Springer, Cham*. STACOM 2018. Lecture Notes in Computer Science(), vol 11395. doi: https://doi.org/10.1007/978-3-030-12029-0_40
10. Wu Y, Kirillov A, Lo F, Girshick R. 2019. Detectron2. Source: https://github.com/facebookresearch/detectron2
11. Iqbal, A., Sharif, M., Khan, M.A. et al. 2022. FF-UNet: a U-Shaped Deep Convolutional Neural Network for Multimodal Biomedical Image Segmentation. Cogn Comput 14, 1287–1302. https://doi.org/10.1007/s12559-022-10038-y
12. Bharati, P., Pramanik, A. (2020). Deep Learning Techniques—R-CNN to Mask R-CNN: A Survey. Computational Intelligence in Pattern Recognition. Advances in Intelligent

Systems and Computing, vol 999. Springer, Singapore. doi: https://doi.org/10.1007/978-981-13-9042-5_56

13. Li Z, Kamnitsas K, Glocker B. Analyzing Overfitting under Class Imbalance in Neural Networks for Image Segmentation. *IEEE Trans Med Imaging*. 2021;40(3):1065-1077. doi:10.1109/TMI.2020.3046692
14. Lopuhin K. lopuhin/kaggle-dsbowl-2018-dataset-fixes. Published online March 21, 2024. Accessed March 31, 2024. https://github.com/lopuhin/kaggle-dsbowl-2018-dataset-fixes
15. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, eds. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Vol 9351. Lecture Notes in Computer Science. Springer International Publishing; 2015:234-241. doi:10.1007/978-3-319-24574-4_28.
16. Wang J, Zhou J, Wang M. Pan-cancer image segmentation based on feature pyramids and Mask R-CNN framework. *Med Phys*. Published online March 4, 2024. doi:10.1002/mp.17014
17. Mela CA, Liu Y. Application of convolutional neural networks towards nuclei segmentation in localization-based super-resolution fluorescence microscopy images. *BMC Bioinformatics*. 2021;22(1):325. doi:10.1186/s12859-021-04245-x
18. Shu J, Zhu Y, Zhao Q, Meng D, Xu Z. MLR-SNet: Transferable LR Schedules for Heterogeneous Tasks. *IEEE Trans Pattern Anal Mach Intell*. 2023;45(3):3505-3521. doi:10.1109/TPAMI.2022.3184315
19. Takase T, Oyama S, Kurihara M. Effective neural network training with adaptive learning rate based on training loss. *Neural Netw*. 2018;101:68-78. doi:10.1016/j.neunet.2018.01.016