**Title:** Automatic Detection of Healthy and Necrotic Optic Nerve Axons using Deep Learning

**Authors**: Shelby Graham\*1, Kyle Freeman\*1, Madhusudhanan Balasubramanian2, Sophie Pilkinton1, Monica Jablonski1

\* Equal contribution

Note: Initial versions of the deep learning models in this manuscript were previously presented at the *Association for Research in Vision and Ophthalmology* (ARVO) annual meeting in May 2023

**Affiliations**

1Department of Ophthalmology, University of Tennessee Health Science Center

2Department of Electrical and Computer Engineering, The University of Memphis

**Grant Support**:

**Financial Disclosure**:

Madhusudhanan Balasubramanian: None

**Correspondence**:

Monica Jablonski, Ph.D.

Department of Ophthalmology

University of Tennessee Health Science Center

Memphis, TN

**Notes:**

***Journal Impact Factors***

* TVST (IF = 2.6, 2024)
* Scientific Reports (IF = 3.8, 2024)
* Computers in Biology and Medicine (IF = 7.0, 2024)
* Artificial Intelligence in Medicine (IF = 6.1, 2024)

## ***Scientific reports:***

## Digital image integrity and standards for journal:

High-resolution images are not required at initial submission. When a paper is accepted, the publishing team will request high-resolution files suitable for publication. All digitized images submitted with the final revision of the manuscript should be 300 DPI if possible.

For guidance, Nature's standard figure sizes are **89 mm wide (single column) and 183 mm wide (double column)**. The full depth of a Nature page is 247 mm. Figures can also be a column-and-a-half where necessary (120–136 mm).

<https://www.nature.com/nature/for-authors/final-submission>

<https://www.nature.com/srep/author-instructions/submission-guidelines#general-figure>

\*Data availability

***Financial Support:***

This work was supported by a the National Eye Institute EY001200 Grant and Challenge Grant from Research To Prevent Blindness

***Competing Interests:***

The authors declare no competing interests.

***Acknowledgements:***

***Author contributions:***

***Data availability:***

**ABSTRACT**

Retinal ganglion cell (RGCs) necrosis and loss are hallmarks of degenerative retinal diseases such as glaucoma. Characteristics of axonal distribution in the optic nerve (ON) such as axon count, morphology and the state of necrosis of each axon are useful for staging the disease and its progression. In this study, we present a deep learning approach for ON instance detection capable of counting, segmenting, and labeling axons as healthy or necrotic using two successful deep learning architectures for *instance segmentation*. ON images from two different study sites were used for training and evaluation. From University of Tennessee Health Science Center (UT dataset), ONs harvested from several generations of outbred heterogenous stock NIH rats were prepared, embedded, and stained with p-phenylenediamine. From Johns Hopkins University (JHU dataset), <include a brief description> In an initial set of 13 confocal images from 1 rat (UT dataset), all ON axons were manually annotated as healthy or necrotic by 3 trained graders with any differences adjudicated by an expert grader. A region based convolutional neural network (Mask R-CNN) architecture was initialized with pretrained ImageNet weights and was fine-tuned using annotated ON images (phase 1 training using 9 training and 2 validation images) for detecting and grading ON axons. Using this model, a separate larger set of 108 images from both study sites were automatically annotated using the trained model from phase 1. These annotations were manually reviewed and corrected by 3 trained reviewers. Mask R-CNN and a second architecture, CenterMask, based on single-shot object detection and attention-based mask generation mechanism were trained using a part of the *UT dataset.* The performance of both deep learning architectures for axon instance segmentation were assessed using average precision and recall measures. In both architectures, model parameters were optimized and fine-tuned before final testing. On both the UT and JHU test datasets, the Mask R-CNN models provided higher average precision and recall values than CenterMask. Average precision and recall of the Mask R-CNN architecture for annotating the independent JHU dataset were 52.58% and 59.58% respectively; and for the CenterMask architecture, they were 32.92% and 45.74% respectively. These deep learning architectures show promise for high throughput localization and grading of ON axons in various experimental glaucoma studies and assessing ON health with candidate drugs.

**Introduction**

Retinal ganglion cell (RGC) degeneration is a hallmark feature of several different neurodegenerative diseases of the eye. Besides aging, most of these diseases exhibit a significant amount of axonal damage and necrosis causing progressive visual loss leading to eventual blindness. Glaucoma is an optic neuropathy and is estimated to affect 111.8 million people worldwide by the year 20401. RGCs make up the innermost layer of the retina and their axons coalesce to form the optic nerve. These axons carry the visual information to the visual cortex of the brain for image formation and vision2. Previously, methods for somal damage characterization and labeling has been studied and several different molecular labeling methods have been developed3. Axonal dysfunction often can precede somal degeneration, thus it is necessary to examine both. RCG axons are smaller and more difficult to label. Understanding the disease process and research of therapeutic options requires visualization, quantification, and characterization of RGC axons and somas. Historically, the study of axonal damage has been done through manual counting and labeling of the harvested, magnified retinal cross section images. The process of manual counting is labor and time intensive and often results in great variability between expert reviewers. Automating the process of detection and grading RGC axons through the use of machine learning (ML) can mitigate these challenges. Indeed, ML in the biomedical field have already demonstrated capabilities such as automated cancerous tumor detection from mammographic images4 or detection of colon cancer from histological slides5.

A picture containing text

Description automatically generated A close-up of a grey surface

AI-generated content may be incorrect. A close-up of a microscope

AI-generated content may be incorrect. (a) (b) (c) (d)

**Figure** **1**. Deep learning-based instance segmentation of confocal optic nerve images for automated detection of axonal geometries and health. **a**) Schematic organization of axons within nerve fascicle in the optic nerve; **b**) tilted confocal images of a whole rat optic nerve; **c**) cropped optic nerve image for axon detection and classification; **d**) automated detection of optic nerve axonal geometry and health using a deep learning instance segmentation model (Mask RCNN).

Previous ML models such as AxonMaster and AxonJ have demonstrated the viability of developing a ML algorithm to recognize RGC axons in microscopy images. AxonMaster proved that an algorithm that provides axon counts in optic nerves of non-human primates ranging in damage from normal to end-stage axon loss can be generated. AxonMaster, however, was not designed for counting smaller axons, such as in mice, and its lack of segmentation limited analysis based on variability in axon area6. AxonJ is an effective tool for efficiently counting and segmenting axons in section images of healthy to moderately damaged mouse optic nerves compared to human observers. However, its performance is decreased when given samples of severely damaged optic nerves7. AxonMaster and AxonJ are also both sensitive to changes in quality of section slides and staining intensity8.

Later projects seeking to create a more robust automated axon counting system used a deep learning approach to design an algorithm less sensitive to variability in axon health, stain intensity, and image quality. A deep neural network (DNN) is a form of ML that uses a neuron as its basic computational unit. AxoNet used a deep learning approach successfully and outperformed both AxonMaster and AxonJ in a head-to-head comparison of axon counts in rat optic nerves, but this system does not provide segmentation of axons8. Another tool using deep learning, AxonDeep, was developed that was designed with both segmentation and counting capabilities. AxonDeep is validated for images of optic nerves with normal to moderate degrees of damage, but the system does not differentiate between live or necrotic axons9. Most recently AxoNet 2.0 was developed that is capable of segmenting "normal-appearing" RGC axons in both healthy and damaged ONs10.

In computer vision, *instance segmentation* algorithms belong to a class of object detection techniques used for detecting multiple instances of an object category or type present in the image11. In this work, we investigated two instance segmentation architectures namely Mask R-CNN12 and CenterMask13 to detect all optic nerve axons, to identify the geometrical boundary of each of the axons (a mask for each axon) and to classify each of the axons as either healthy or necrotic. In Mask R-CNN, an initial set of object proposals representing candidate locations of objects in each of the categories are generated using predefined geometries known as *anchors*. Using these object proposals, a final set of objects in each of the object categories and their geometric boundaries are estimated (as in a two-stage object detector). In contrast, CenterMask utilizes a single-stage object detector (FCOS architecture) for detecting objects and for estimating their geometric boundaries. We carefully identified the model parameters and trained these deep learning architectures to detect all instances of optic nerve axons in confocal images and identify their health status. We trained these networks in two-phases and evaluated their performance using two optic nerve axon confocal imaging datasets from two study centers.

**METHODS**

**Subjects and Data Collection**

Optic nerve axonal imaging data used in this study come from two independent sources namely from the University of Tennessee Health Science Center (UTHSC) and from Johns Hopkins University (JHU). A description of the study and the imaging protocol of these two data sources are as follows.

UTHSC Study:

This study used the 73rd through 80th generation from NIH heterogeneous stock outbred rats (HS rats) which were obtained from the NMcwi:HS colony (RGD Cat# RGD\_13673907, RRID:RGD\_RGD\_13673907). These rats came from eight inbred strains (ACI/N, BN/SsN, BUF/N, F344/N, M520/N, MR/N, WKY/N and WN/N) and have been maintained as an outbred population since 1984 [(Hansen and Spuhler, 1984)](https://paperpile.com/c/7j2o3F/pyo1). Breeders were given free access to the Teklad 5010 diet (Envigo, Madison, WI). Starting in 2014 through 2017, 16 batches with 50 HS rats per batch (25 males and 25 females) of 3-6 weeks of age were brought to the University of Tennessee Health Science Center (UTHSC) and used for population expansion. This designed rat population demonstrates variation in IOP and all the variant phenotypes are common in this population which allows for better study of a wide range of variants within a smaller sample size. All proper protocols and procedures pertaining to the rat population were reviewed and approved by the Animal Care and Use Board of UTHSC and were in accordance with the guidelines for laboratory animal experiments (Institute of Laboratory Animal Resources, Public Health Service Policy on Humane Care and Use of Laboratory Animals) as well as the Association of Research in Vision and Ophthalmology (ARVO) Statement for the Use of Animals in Ophthalmic and Visions Research.

Sections of optic nerves were collected from sacrificed rats and placed into a Smith-Rudt fixative for 24 hours. Nerves were then treated with 1% osmium tetroxide before being dehydrated and embedded into Epon 812. Optic nerves were sliced into one micron thick sections and stained with 1% p-phenylenediamine (PPD). The samples were then mounted under a cover glass and imaged with a Zeiss 710 Laser Scanning Microscope using ZEN Black software. A tilting function was utilized under the bounding grid method with a 10% overlap to image each nerve entirely. Each image was obtained using a 63X oil objective (1.4 NA) with 1.2 to 1.4X zoom (accounting for nerve diameter), a pixel dwell of 1.3 ms/pixel, and a 568 nm laser line (power setting 2.0 and gain 50) along with the T-PMT function with gain between 200 and 400 due to variation in tissue staining. Once the completed images were tiled, the full nerve image as well as individual tile images were exported as TIFF files for manual annotation by trained experts and for use in training the DNN.

JHU Study:

<Describe 1) the subjects used in this study, their age, etc. 2) optic nerve sectioning and staining procedures, 3) confocal specifications and 4) imaging procedures>

**Axon Morphology and Health Status**

Axonal morphology as observed from ON sections

Figure 1a shows the schematic arrangement of axons within the ON. Axons arising from RGCs coalesce and form the ON. Sections of the ON cut perpendicular to the direction of axons highlight the organization of nerves (Figure 1a, 1b). The outermost layer of the ON is called the *epineurium*. Nerve fascicles within the epineurium are surrounded by *perineurium*. Individual axons are encased in the *endoneurium*. An axon, as observed with confocal microscopy, has an identifiable membrane surrounding the axoplasm. Because the membranes of individual axons are difficult to distinguish in our confocal images, we define the geometrical boundary of an axon to be along the midpoint of the membranes of adjacent axons.

Morphology and appearance of ON axons during Necrosis:

Healthy axons have a continuous, uniform membrane surrounding a bright axoplasm with no focal points of discoloration. Necrotic axons exhibit very thick, and disproportionately dark membranes with possible abrupt changes in membrane thickness and with possible areas of central necrosis. As axons undergo necrosis, their membrane irregularly thickens and their axoplasm develops darks areas of shading or spotting. With progressive membrane breakdown over time, necrotic axonal membranes do not take up the staining pigment well during histology. We refer to the result appearance of necrotic axons as *ghosts*. Eventually, the membrane breaks down fully causing leakage of membrane fragments and axoplasm components. We refer to the appearance of this phenomenon in the confocal images as a *river of garbage*. In the later stages of necrosis, these necrotic axonal contents are remodeled into gliotic scars over time. These axonal subcategories of ghosts, rivers of garbage, and gliotic scars are ignored in this current analysis as they do not fall into either the healthy or necrotic labelling category. Figure 2 show example appearance of necrotic and healthy axons as observed with our confocal microscope in the UTHSC study.

|  |  |  |
| --- | --- | --- |
| A blurry image of a white square  AI-generated content may be incorrect. A blurry image of a black and white object  AI-generated content may be incorrect. A blurry image of a person's face  AI-generated content may be incorrect. |  | A blurry image of a person's face  AI-generated content may be incorrect. A blurry picture of a black and white picture  AI-generated content may be incorrect. |
| a) Necrotic |  | b) Healthy |

**Figure 2**. Appearance of healthy and necrotic axons as observed by a confocal microscope.

Manual Annotation of ON Axons:

For manual annotation of ON axons, annotators in this study were trained by a single expert. During training, images labeled by the annotators were reviewed, discussed and corrected weekly by the expert annotator. Discussion included incorrect geometrical annotation, missed annotations, false annotations and miscategorized axonal health. After several weeks of training, the expert annotator certified our annotators to be qualified for independent execution of manual annotation, review, and correction of annotation and classification of axons in confocal images. Figure 1d shows an example annotation of axons in the confocal image of a rat optic nerve. Geometrical boundaries of axons are marked as polygons with the color of the polygon representing their healthy status (green: healthy, red: necrotic).

**Data Preparation**

All confocal images from the UTHSC study are of dimension 1024 x 1024 pixels. These images were divided into 16 non-overlapping tiles of dimension 256 x 256 pixels and were used as individual images for training, validation and testing of the deep learning models. Images from the JHU study are of dimension 610 x 611 pixels (similar appearance and geometry of ON axons, but with a smaller field of view). These were divided into a total of four non-overlapping image tiles two of which are of dimension 305 x 306 pixels and the other two are of dimension 305 x 305 pixels.

Following the manual annotation procedure described earlier, a smaller set of 13 images from the UT dataset were annotated manually by drawing the geometrical boundaries and identifying the health status of each axon as either healthy or necrotic by the trained experts. To speed up the annotation of all remaining 108 images in the UT and JHU datasets, we trained a Mask R-CNN model MR-1 using 9 images and validated using 2 images that were manually annotated (model details provided below). Training parameters of the initial model MR-1 are presented in Table 2. We subsequently used the trained MR-1 model to automatically annotate and classify axons in all remaining images in both datasets. These model-generated annotations and classification results were manually reviewed and, when necessary, corrected and adjudicated by the trained experts. The reviewed and corrected annotations were used as ground truth description of axonal geometry and health status for training the instance segmentation models and for evaluating their performance.

For developing the final instance segmentation models, we used 30 images for training (UT training dataset), 14 images for validation (UT validation dataset) and 13 images for testing (UT testing dataset) from the UT dataset. In addition, a total of 51 images from the JHU dataset were used for conducting an independent evaluation of the final trained models (JHU testing dataset).

**Deep Learning Models for Axon Detection and Classification**

In computer vision, the task of locating all instances of optic nerve axons from confocal images, detecting their geometrical boundaries and classifying them based on their health status belongs to the class of *instance segmentation* problem. In this work, we utilize two successful deep learning architectures namely Mask Region-based CNN (Mask R-CNN) 14 and CenterMask 15 for automated detection and classification of optic nerve axons.

***1. Mask R-CNN Model***:

Mask R-CNN utilizes the same framework as the Faster R-CNN16 method along with a mask-generation module as shown in Figure 3. The deep network architecture is comprised of a backbone network and a *feature pyramid network* (FPN)17 for generating multiscale image features, a *region proposal network* (RPN) to identify promising locations and object geometries known as *regions of interest* (RoI) and fully connected layers for final detection of objects, and their classes or categories.

While the multiscale features from the backbone are limited to low- and mid-level image features such as edges and image texture, the FPN enhances these image features to gain semantic information such as object-level features. This is achieved through information flow from the top pyramidal levels with low resolution feature maps but with higher field of view to lower pyramidal levels with high resolution feature maps with lower field of view and across spatial locations within each pyramidal level. *A priori* object geometries of various sizes and aspect ratios known as *anchors* and the multiscale semantic-aware image features within these anchors are used to train a fully convolutional region proposal network (RPN) to detect RoIs in the image containing objects of interest. Multiscale features within each ROI are used to detect a box bounding each of the objects of interest as well as its category using trainable fully connected layers. In Mask R-CNN, an additional fully convolutional network (FCN)18 uses the same multiscale features within each ROI to identify the geometrical object boundaries (masks).

A black background with a black square

AI-generated content may be incorrect.

**Figure 3**. Schematic architecture of the Mask RCNN model for object instance detection

Mask R-CNN Training:

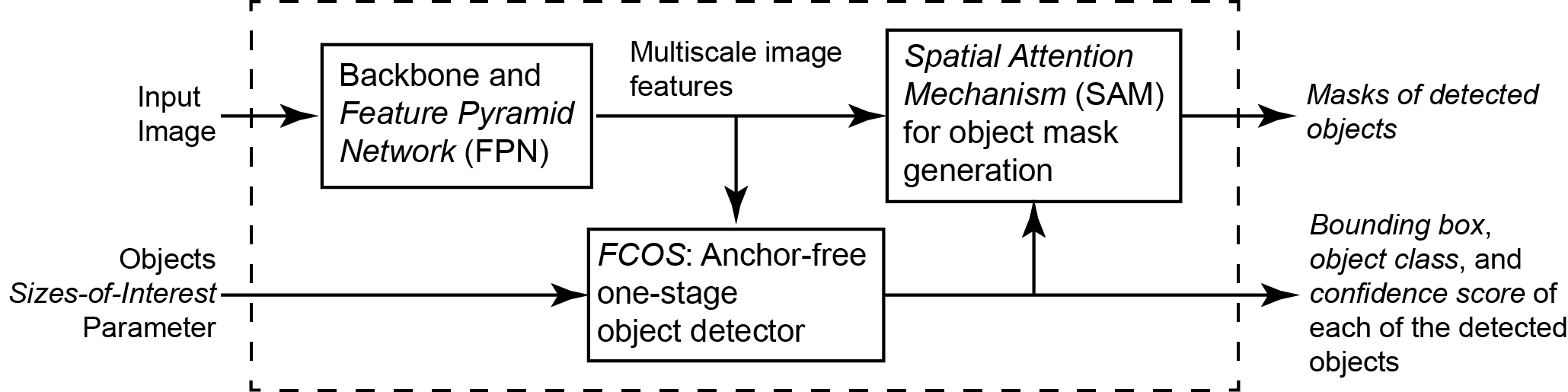
Because the data size is relatively smaller, we utilized a transfer learning technique to initialize the Mask R-CNN models with weights from larger models pre-trained for object and classification (COCO or ImageNet) and fine-tuned them on our UT training and UT validation datasets. This allowed us to retain the rich feature extraction (edges, textures, derivatives, etc.) and feature processing pathways from these generic image segmentation models as well as facilitate faster convergence of the model for axon detection and classification. A selected list of important model parameters and training parameters for the Mask R-CNN model are shown in Table 2.

***2. CenterMask2 Model***:

Figure 4 shows the schematic architecture of the CenterMask model13. CenterMask is broadly comprised of a *variety-of-view* backbone (VoVNet2)19 for extracting various image features, a *feature pyramid network* (FPN) for generating semantic-aware multiscale features, a *fully convolution one-shot object detector* (FCOS)20 for detecting objects, their locations and classifications, and a *spatial attention mechanism* (SAM) for generating object boundaries of objects detected by FCOS. The original VoVNet model addressed unnecessary and detrimental aggregation of redundant and correlated features in DenseNet21 by introducing a *one-shot aggregation* (OSA) approach wherein features from all layers within a block are aggregated only once in the output layer. CenterMask improved over the original VoVNet with VoVNet2 by incorporating a residual learning approach as in ResNet22,23 thereby addressing the degradation problem of saturated and declining accuracy with increasing network depth as well as with a *channel attention mechanism24* for optimal selection of features.

When compared to Mask R-CNN, CenterMask uses FCOS as the object detector to identify object instances (location, classification score and a centerness score) without requiring a region proposal generator (i.e. RPN) as a precursor. Further, the object detector is anchor-free and therefore does not, in principle, require parameter tuning to generate optimal anchor boxes for generating object proposals. Instead, every location in each feature scale is checked as a possible center location for each of the ground truth objects in the training dataset using a measure of intersection over union (IoU). These object coordinates (left *l*, top *t*, right *r* and bottom *b* eccentricities from the object center o) defined using bounding boxes are used as *regions of interests* (RoI) for generating the object masks. During training, candidate center locations within each ground truth object boundaries are assigned to utilize feature maps at a specific scale based on the object eccentricity (maximum of *l*, *t*, *r*, *b*).

As RoIs used as input for mask generation correspond to different scales of the feature maps, a modified RoI Align procedure is used as in Mask R-CNN to extract scale-independent fixed scaled features for mask generation. Locations within these feature maps are emphasized using the *spatial attention mechanism*. Finally, objects masks are generated using a 1 x 1 convolutional layer using these RoIs with spatial emphasis.



**Figure 4**. Schematic representation of the Centermask network for object instance detection

CenterMask training details:

CenterMask architecture was initialized with either VoVNet2-39, or ResNet-50 backbone based pretrained weights. A constant training rate of 0.01 and a constant momentum of 0.7 were used for training the CenterMask models. *L*2 norm based gradient clipping (of 10) was used to regularize the weight update process. During training all input images were rescaled with an equal scaling factor on both directions with the smaller side scaled to either 768 or 1024. During testing, the minimum side was scaled to 1024 pixels and the other side scaled proportionately.

**Performance Measures**

The standard precision measure of a method is defined as the proportion of object instances detected by the method that are true positives among all objects detected and flagged by a method as positive detections. Similarly, the recall measure is defined as the proportion of true positives detected by the method among all the true positives in the dataset. These performance measures are affected by parameter thresholds that define positive detections. For *instance-segmentation* models, thresholds on *intersection over union* (IoU) measures between the detected object geometries and true object geometries as well as a measure of confidence on the detected object (confidence score) are used to identify positive object detection. Therefore, to avoid the influence of these thresholds on how precision and recall measures are reported by various methods, average precision (AP) and average recall (AR) measures are reported for object instance segmentation models. A brief description of various AP and AR performance measures of the Mask R-CNN and the CenterMask methods reported in this study are as follows11.

Average Precision (AP50)

For an IoU threshold of 0.5, a precision-recall curve is constructed by varying the confidence threshold. The average precision AP50 measure is defined as the area under the precision-recall curve (AUC-PR). We are reporting the AP50 measure for each of the axon categories (AP50-Healthy and AP50-Necrotic) as well as the overall measure irrespective of the axon category (AP50). These measures report the model performance using a fixed IoU threshold of 50% between the geometry of each of the detected axons with its corresponding true geometry.

Mean AP (AP50To95)

To characterize the model performance under various IoU thresholds, geometric mean of AP at IoU thresholds ranging from 0.5 through 0.95 in steps of 0.05 is reported as AP50To95. The mean average precision measure is reported for each category (AP50To95-Healthy and -Necrotic) and for the overall set of axons detected by the method.

Recall at fixed IoU (R50)

Recall measures are reported at IoU=0.5 for all axonal categories (R50) and for individual categories (R50-Healthy and -Necrotic).

Average Recall (AR50To95)

Model recall values at various IoU thresholds were summarized as geometric means of recall measures at IoU thresholds from 0.5 through 0.95 in steps of 0.05 (AR50To95). The mean recall measure is reported for each category (AR50To95-Healthy and -Necrotic) and for all the categories combined (AR50To95).

To understand the performance of the model as a function of axonal area, we also investigated these AP and AR measures on all axons stratified into groups of small (0 to 1,304 sq pixels), medium (1,304 to 2,256 sq pixels), and large (2,256 to 1.1e6 sq pixels) axons.

**RESULTS**

All ON confocal images were annotated manually using the *COCO Annotator* software25. The Mask R-CNN architecture implemented using TensorFlow26 and the CenterMask architecture implemented using Detectron2 (ver. 0.6, an open-source object detection framework from Facebook) in PyTorch27 were used in this study. All models were trained and evaluated on a server with Nvidia RTX 8000 GPUs.

Data Characteristics:

The overall distribution of axonal count, their *eccentricity* and their area in each of the data groups are presented in Table 1. Approximately 65% of axons in the UT dataset and 70% axons in the JHU dataset are necrotic. The measure of axon *eccentricity* (in pixels) was defined as the farthest distance from axonal center to either the left, top, right or bottom boundary of the axon. The median eccentricity of necrotic axons was 32 pixels for the UT datasets and 30 pixels for the JHU datasets. For healthy axons, the median eccentricity was 20 pixels for the UT datasets and 24 pixels for the JHU datasets. Similarly, the axonal area was larger for necrotic axons than healthy axons in all datasets.

On an equal tertile division of axons by their areas as smaller, medium and larger area axons, the distribution of both their eccentricities and areas in each tertile group were similar in all study datasets as shown in Table 3. As shown in Table 4, the majority (47.30% to 54.58%) of necrotic axons had a larger area of at least 2,246 sq. pixels and a smaller proportion (13.33% to 14.31%) of necrotic axons had a smaller area of at most 1,304 sq. pixels. Among healthy axons, the majority (61.73% to 72.78%) had a smaller area of at most 1,304 sq. pixels and a very small portion of them (0.91% to 1.53%) had a larger area of at least 2,256 sq. pixels.

With an unequal tertile division of eccentricities skewed to the right as smaller (0 to 23 pixels), medium (23 to 28 pixels), and larger eccentricities (28 to 176 pixels), the distribution of eccentricity and area measures were skewed toward the larger eccentricity bins. Majority of the axons had larger eccentricities and larger areas (45.04% to 54.35%) as shown in Table 5. In the UT datasets, most necrotic axons (74.65% to 76.03%) had larger eccentricities, and most healthy axons (58.31% to 60.40%) had smaller eccentricities. In the JHU dataset, most necrotic axons (61.88%) had larger eccentricities and most healthy axons had either smaller eccentricities (45.07%) or medium eccentricities (48.80%). As shown in Table 6, this unequal tertile division of axonal eccentricities allowed us to achieve a larger separation between the geometrical characteristics of necrotic and healthy axons.

Mask R-CNN Parameter Tuning and Model Performance:

A more detailed description of Mask R-CNN parameters and their choices are available elsewhere28–30. Healthy axons in both the UT and JHU datasets were in general smaller in area (approximately 1000 sq. pixels) compared to necrotic axons (approximately 2000 sq. pixels) as shown in Table 4. Therefore, through experimental combination of various anchor sizes and anchor ratios, we identified an optimal set of anchor size ([8, 16, 32, 64, 128] instead of [32, 64, 128, 256, 512]) and ratio ([0.25, 0.5, 1, 2, 4] instead of [0.5, 1, 2]) parameters to identify both healthy and necrotic axons in the RPN module. A detailed list of important model parameters of the two Mask R-CNN models namely MR-1 and MR-2 are presented in Table 2. As described earlier, model MR-1 was used for automated annotation of all study datasets which served as the ground truth of axonal geometry annotations and health status after review and correction by trained experts.

After training the Mask R-CNN architecture using the *UT Training* dataset, an optimal model MR-2 was chosen from the training iteration with the optimal precision and recall performance on the *UT Validation* dataset. For the *UT Training* dataset, the AP50to95 average precision measure was 41.67% overall, 48.48% for annotating necrotic axons and 34.86% for healthy axons. For the *UT Testing* dataset, the average precision for all axons, necrotic and healthy axons were, respectively, 44.32%, 51.78% and 36.86%. For the *JHU Testing* dataset, the average precision for all axons, necrotic and healthy axons were 52.58%, 58.83% and 46.33% respectively. A comprehensive set of all performance measures for the optimized Mask R-CNN model CM-2 is presented in Table 7. Figures 1, 2, and 3 show the MR-2 model-generated annotations and grading of representative images from the *UT Validation*, *UT Testing*, and *JHU Testing* datasets.

CenterMask Parameter Tuning and Model Performance:

Multiresolution and multiscale methods, in general, utilize higher resolution features from lower pyramidal levels to detect smaller objects and lower resolution feature maps from higher pyramidal levels to detect larger objects. In CenterMask, we used axonal eccentricity measures to assign specific scales of the feature maps trained to detect object instances with specific geometric extent or dimension. By default, CenterMask utilizes five pyramidal level features (from the FPN) as input for object instance detection namely “p3”, “p4”, “p5”, “p6” and “p7” resulting in featuring maps with a stride length of 8, 16, 32, 64 and 128 respectively in the order of decreasing feature map resolution. For an input image of size 1024 x 1024 pixels, feature maps will be of sizes 256 x 256, 128 x 128, …, 8 x 8 pixels at the respective pyramidal levels. With the default eccentricity bins (FCOS.SIZES\_OF\_INTEREST = [64, 128, 256, 512]), we observed that 1) only feature maps from levels “p3” and “p4” were used to detect majority of the objects with minimal or no contribution from the other levels; and 2) significant number of healthy and necrotic axons were assigned to utilize the same FPN level features. We determined an optimal eccentricity binning (as FCOS.SIZES\_OF\_INTEREST = [23, 28] based on the eccentricity distribution of healthy and necrotic axons (Table 6). Further, in an attempt to facilitate larger feature maps for detecting smaller axons, we limited the FPN features to levels “p2”, “p3” and “p4” only.

As in Mask R-CNN, the CenterMask models were trained using the *UT Training* dataset and validated independently using the *UT Testing* and *JHU Testing* datasets. Based on the average precision and recall measures observed in the *UT Validation* dataset, the optimal CenterMask model CM-2 was chosen. Table 8 shows the default and optimal CenterMask parameters used for training and testing the models. Similar to the Mask R-CNN model MR-2, the optimal CenterMask model CM-2 provided the highest average precision and recall measures of 32.92% and 45.74% respectively for the *JHU Testing* dataset when compared to the *UT Testing* dataset with an average precision of 28.86% and an average recall of 41.52%. The average precision and recall of detecting necrotic axons were 38.28% and 50.75%, respectively, for the *UT Testing dataset*; and 41.18% and 52.40%, respectively, for the *JHU Testing* dataset. For detecting healthy axons, the average precision and recall measures were 19.33% and 32.39%, respectively, for the *UT Testing dataset*; and 24.66% and 39.07%, respectively, for the *JHU Testing* dataset. A detailed summary of performance measures for the optimized CM-2 model is presented in Table 9. Figures 1, 2 and 3 show the annotations and grading of representative images from the *UT Validation*, *UT Testing*, and *JHU Testing* datasets generated by the optimal CenterMask model CM-2 and by the optimal Mask R-CNN model MR-2.

**DISCUSSION**

Performance analyses were conducted on the *Testing* datasets only once at the end of the study after finalizing an optimal model for both Mask R-CNN and CenterMask architectures. While the *UT Testing* dataset is part of the same UTHSC study from which the models were trained (using *UT Training*), the distributions of axonal counts, and geometrical characteristics were similar in both *UT Testing* and *JHU Testing* datasets. Though the final testing datasets came from different study sites, both deep learning instance segmentation models were able to achieve a similar detection performance in both datasets. This improves our confidence on the ability of the trained models to translate to other optic nerve scans without incurring overfitting during training. Thus, these architectures show promise for high throughput localization and grading of optic nerve axons.

The highest average precision measure (AP50to95) of annotating axons from the *JHU Testing* dataset was 52.58% for the optimal Mask R-CNN model (MR-2) and 32.92% for the optimal CenterMask model (CM-2). The highest average recall measure (AR50to95) of annotating axons from the *JHU Testing* dataset was 59.58% for the MR-2 model and 45.74% for the CM-2 model. Performance of both models were lower for annotating the *UT Testing* dataset than that of annotating the *JHU Testing* dataset. As observed in this study, the average precision measures (AP50to95) of instance segmentation models are in the range of 30% to 60% as they reflect the average detection ability of the model with a varying choices of classification parameters and their thresholds31.

In this study, an initial Mask R-CNN model (MR-1) was used to automatically generate annotations for all datasets (*UT Training, UT Validation, UT Testing,* and *JHU Testing*). These annotations were reviewed and corrected before conducting the final phase of Mask R-CNN and CenterMask model developments to identify MR-2 and CM-2 respectively. Because the same Mask R-CNN architecture was initially used to generate the annotations, it is likely that the optimal Mask R-CNN model MR-2 performed significantly better than the CenterMask model CM-2 (e.g. AP50to95 of 44.32% vs 28.86% for *UT Testing* dataset). To test this hypothesis, we initialized an independent Mask R-CNN model with pretrained ImageNet weights and trained the model using *UT Training* dataset for 300 iterations. An optimal model was selected (MR-3 from iteration 263) based on its performance for the *UT Validation* dataset. A detailed performance summary of model MR-3 is presented in Table 10. For the *UT Testing* dataset, the MR-3 model provided an average precision of 41.48% very similar to the optimal model MR-2. This confirmed that Mask R-CNN architecture based on our current fine-tuning procedures and model parameter choices provided the best overall performance when compared to the CenterMask models.

To further understand the possible sources of parameter optimization for CenterMask, we first investigated the performance using the default FPN parameters (model CM-1). Model parameters of CM-1 are presented in Table 8 and its performance measures are presented in Table 11. Specifically, the model CM-1 was trained using the default scale assignment parameter with five pyramidal levels (*k* = 3, 4, 5, 6, 7 and FPN\_LEVELS = [“p3”, “p4”, “p5”, “p6”, “p7”]) each with a feature scale ratio of 1/2k. Upon inspection, we observed that the model utilized only features from scales with *k* = 3 and 4 for detecting both necrotic and healthy axons. In contrast, with an optimal parameter choice, CM-2 was able to utilize a specific scale (higher resolution features to detect smaller healthy axons) for each axon class (i.e. only a lower proportion of both classes was assigned to any given feature scale) and utilize higher resolution features from “p2” pyramidal level (instead of “p3”). This provided a balance between the field of view of the output layers and the feature resolution to maximize the detection accuracy. Despite these advantages, the performance of the default model CM-1 was also similar (AP50to95 of CM1 = 41.43% and of CM2 = 45.74% for the *JHU Testing* dataset) requiring further investigation.

The FCOS module for object detection in CenterMask uses the measures of classification error, *centerness* error, and location or *objectness* error during training. While Mask R-CNN utilizes corner coordinates to specify boundary coordinates of rectangular regions (boxes) enclosing the detected axons, the FCOS detector within the CenterMask utilizes an object center coordinate and the corresponding eccentricities to the left, top, right and bottom boundaries. The FCOS module, therefore, learns to identify the centerness of each object in addition to correctly identifying the resulting bounding box (centerness of each detection measured using its aspect ratio). In every CenterMask model we trained, the centerness error did not improve (both the initial and final centerness errors were around 0.6). A higher loss related to the centerness detection indicated the consistent difficulty the network had in learning to correctly detect the axonal centers. It is highly likely that this loss is higher for smaller axons. Theoretically, the reference point within an object need not to be at the center of the object as the eccentricities to the object boundaries will compensate for candidate locations that are off-center. This is the most likely cause of the observed lower performance of the CenterMask models CM-1 and CM-2 in all *Testing* datasets.

For both models, we experimented with learning rate adjustment strategies such as a linear warm up period leading to a base learning rate followed by either a constant learning rate, a cyclically adjusted learning rate or damped cosine annealing. For the optimal Mask R-CNN models, both the learning rate and momentum were adjusted throughout the training period, typically in the range of 5e-5 to 1e-2 and 0.7 to 0.9 respectively, after a linear warm up period of 1000 iterations. For CenterMask models, a simpler warm up phase with a constant learning rate of 0.01 provided the best performance on the validation dataset. Though the ResNet-50 backbone provided higher performance for Mask R-CNN models, we did not observe any performance improvement in CenterMask models with either ResNet-50 or -101 backbones.

**CONCLUSION**

Labeling retinal ganglion (RGC) cell axons when studying neurodegenerative ocular pathologies is time consuming and often fraught with large inter-observer variation. While a few computational methods are available for detecting axons from ON images, no methods are available at present for automatic markup as well as grading of ON axons. In this study, we have developed optimal deep learning models for annotating and classifying ON axons based on the Mask R-CNN and CenterMask object instance segmentation architectures. We carefully optimized all the model parameters including the object proposal generation parameters in the Mask R-CNN model and the multiscale feature parameters for detecting axons with differing areas in CenterMask. These models provided higher average precision and recall values for two different study datasets and thus reflect their ability to annotate axons from the confocal scans of the ON. Though the CenterMask architecture does not require object proposals as in Mask R-CNN, it has difficulty learning to identify axon centers (more likely in smaller, healthy axons) leading to its lower performance than Mask R-CNN. Overall, the deep learning instance segmentation techniques show promise for high throughput annotation of confocal ON scans and grading of ON axons. They have likely applications in studying the efficacy of glaucoma drugs and understanding the pathophysiological mechanisms of axonal loss in glaucoma.

**SOFTWARE**

In this study, we utilized Matterport’s implementation of Mask R-CNN using TensorFlow available at <https://github.com/matterport/Mask_RCNN/> and CenterMask available from the original study using PyTorch and Detectron2 at <https://github.com/youngwanLEE/CenterMask>. Optimized model parameters, configuration files and utility programs are available at <URL>.

**DATA AVAILABILITY**

Data may be made available upon request to the corresponding author of this manuscript.

**ACKNOWLEDGEMENT**

The authors thank Andrew B. Stiemke, Ph.D. for his efforts in training the personnel for 1) reading confocal images of the optic nerve, 2) manual annotation of axonal geometries in the confocal scans, and 3) categorizing axons as healthy or necrotic; and also for adjudicating the earlier manual annotations as well as during review of initial model-generated axonal annotations and classifications.

**BIBLIOGRAPHY**

1. Allison, K., Patel, D. & Alabi, O. Epidemiology of Glaucoma: The Past, Present, and Predictions for the Future. *Cureus* **12**, e11686 (2020).

2. Fudalej, E. *et al.* Neuroprotective Factors of the Retina and Their Role in Promoting Survival of Retinal Ganglion Cells: A Review. *Ophthalmic Res* **64**, 345–355 (2021).

3. Stiemke, A. B. *et al.* Systems Genetics of Optic Nerve Axon Necrosis During Glaucoma. *Front Genet* **11**, 31 (2020).

4. Wang, Z., Yu, G., Kang, Y., Zhao, Y. & Qu, Q. Breast tumor detection in digital mammography based on extreme learning machine. *Neurocomputing* **128**, 175–184 (2014).

5. Sirinukunwattana, K. *et al.* Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images. *IEEE Trans Med Imaging* **35**, 1196–1206 (2016).

6. Reynaud, J. *et al.* Automated quantification of optic nerve axons in primate glaucomatous and normal eyes--method and comparison to semi-automated manual quantification. *Invest Ophthalmol Vis Sci* **53**, 2951–9 (2012).

7. Zarei, K. *et al.* Automated Axon Counting in Rodent Optic Nerve Sections with AxonJ. *Sci Rep* **6**, 26559 (2016).

8. Ritch, M. D. *et al.* AxoNet: A deep learning-based tool to count retinal ganglion cell axons. *Sci Rep* **10**, 8034 (2020).

9. Deng, W. *et al.* AxonDeep: Automated Optic Nerve Axon Segmentation in Mice With Deep Learning. *Transl Vis Sci Technol* **10**, 22 (2021).

10. Goyal, V. *et al.* AxoNet 2.0: A Deep Learning-Based Tool for Morphometric Analysis of Retinal Ganglion Cell Axons. *Transl Vis Sci Technol* **12**, 9–9 (2023).

11. Jung, S., Heo, H., Park, S., Jung, S. U. & Lee, K. Benchmarking Deep Learning Models for Instance Segmentation. *Applied Sciences 2022, Vol. 12, Page 8856* **12**, 8856 (2022).

12. He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask R-CNN. *IEEE Trans Pattern Anal Mach Intell* **42**, 386–397 (2017).

13. Lee, Y. & Park, J. CenterMask : Real-Time Anchor-Free Instance Segmentation.

14. He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask R-CNN. *IEEE Trans Pattern Anal Mach Intell* **42**, 386–397 (2017).

15. Lee, Y. & Park, J. CenterMask : Real-Time Anchor-Free Instance Segmentation.

16. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.

17. Lin, T.-Y. *et al.* Feature Pyramid Networks for Object Detection.

18. Long, J., Shelhamer, E. & Darrell, T. Fully Convolutional Networks for Semantic Segmentation.

19. Lee, Y. & Bae ETRI, Y. *An Energy and GPU-Computation Efficient Backbone Network for Real-Time Object Detection Joong-Won Hwang \* ETRI*.

20. Tian, Z., Shen, C., Chen, H. & He, T. FCOS: Fully Convolutional One-Stage Object Detection.

21. Iandola, F. *et al.* DenseNet: Implementing Efficient ConvNet Descriptor Pyramids. (2014).

22. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **2016-Decem**, 770–778 (2016).

23. He, K., Zhang, X., Ren, S. & Sun, J. Identify Mappings in Deep Residual Networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **9914 LNCS**, V (2016).

24. Chorowski, J. & Bahdanau, D. Attention-Based Models for Speech Recognition.

25. Brooks, J. COCO Annotator. https://github.com/jsbroks/coco-annotator/ (2019).

26. Abdulla, W. Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow. https://github.com/matterport/Mask\_RCNN (2017).

27. Lee, Y. & Park, J. CenterMask. https://github.com/youngwanLEE/CenterMask (2020).

28. Machefer, M., Lemarchand, F., Bonnefond, V., Hitchins, A. & Sidiropoulos, P. Mask R-CNN refitting strategy for plant counting and sizing in uav imagery. *Remote Sens (Basel)* **12**, 1–23 (2020).

29. Melo, F. De & Lameir, G. Diagnosis of Apple Fruit Diseases in the Wild with Mask R-CNN. (2022).

30. Manda-Handzlik, A., Fiok, K., Cieloch, A., Heropolitanska-Pliszka, E. & Demkow, U. Convolutional Neural Networks–Based Image Analysis for the Detection and Quantification of Neutrophil Extracellular Traps. *Cells 2020, Vol. 9, Page 508* **9**, 508 (2020).

31. Jung, S., Heo, H., Park, S., Jung, S. U. & Lee, K. Benchmarking Deep Learning Models for Instance Segmentation. *Applied Sciences 2022, Vol. 12, Page 8856* **12**, 8856 (2022).

**Table 1** Geometric characteristics of ON axons in the study datasets. Majority of axons are necrotic (65 to 70%) with larger eccentricities and axonal area in all datasets. UT datasets were used for model training and validation. Both UT dataset (testing) and JHU dataset (JHU testing) were used for final evaluation of the model performance. The dimension of all UT images was 256 x 256 pixels and the dimension of JHU images was either 305 x 305 or 305 x 306 pixels. For training, validation and testing, images were scaled to 1024 x 1024 pixels. All axonal characteristics depicted here are from images of size 1024 x 1024 pixels.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | # Images | Axon Distribution | | | | | | | | |
| **# Axons** | | | **Eccentricity in # pixels**  (median / range) | | | **Area in sq. pixels**  (median / range) | | |
| Overall | Necrotic | Live | Overall | Necrotic | Live | Overall | Necrotic | Live |
| **UT Training** | 30 | 4148 | 2773 / 66.85% | 1375 / 33.15% | 28 / 12 to 140 | 32 / 12 to 140 | 20 / 12 to 64 | 1,680 / 24 to 35,388 | 2,232 / 24 to 35,288 | 1,088 / 222 to 4,304 |
| **UT Validation** | 14 | 2793 | 1836 / 65.74% | 957 / 34.26% | 28 / 12 to 176 | 32 / 12 to 176 | 20 / 12 to 52 | 1,712 / 172 to 19,744 | 2,376 / 172 to 19,745 | 1,088 / 256 to 4,728 |
| **UT Testing** | 13 | 2573 | 1,684 / 65.45% | 889 / 34.55% | 28 / 12 to 284 | 32 / 12 to 108 | 20 / 12 to 284 | 1,696 / 32 to 20,856 | 2,320 / 32 to 20,856 | 1,073 / 248 to 5,453 |
| **JHU Testing** | 51 | 14,671 | 10,278 70.06% | 4,393 / 29.94% | 27 / 10 to 154 | 30 / 10 to 154 | 24 / 10 to 57 | 1,764 / 11 to 25,001 | 2,187 / 11 to 25,001 | 1,201 / 11 to 3,292 |

**Table 2** Mask R-CNN model parameters optimized for optic nerve axon detection based on the distribution of axonal geometrical characteristics. Model MR-1 was trained with default RPN anchor parameters using a smaller set of training data (9 training images and 2 validation images). In model MR-2, the RPN anchor parameters were optimized based on axonal geometric characteristics. The model was initially trained using a limited training dataset and finetuned using a larger set of training data.

|  |  |  |
| --- | --- | --- |
| Model Parameters | Model MR-1 | Model MR-2 |
| **Network Initialization** | COCO Weights | ImageNet weights |
| **BACKBONE** | ResNet50 | ResNet50 |
| **BACKBONE\_STRIDES** | [4,8,16,32,64] | [4,8,16,32,64] |
| **IMAGE\_SHAPE** | [1024,1024,3] | [1024,1024,3] |
| **IMAGE\_MIN/MAX\_DIM** | 800/1024 | 800/1024 |
| **MAX\_GT\_INSTANCES** | 400 | 400 |
| **RPN\_ANCHOR\_RATIOS** | [0.5, 1, 2] | **[0.25, 0.5, 1, 2, 4]** |
| **RPN\_ANCHOR\_SCALES** | (32, 64, 128, 256, 512) | **(8, 16, 32, 64, 128)** |
| **RPN\_NMS\_THRESHOLD** | 0.7 | 0.9 |
| **RPN\_TRAIN\_ANCHORS\_PER\_IMAGE** | 400 | 400 |
| **PRE\_NMS\_LIMIT** | 6000 | 6000 |
| **POST\_NMS\_ROIS\_INFERENCE / TRAINING** | 1800 / 8000 | 1800 / 8000 |
| **ROI\_POSITIVE\_RATIO** | 0.33 | 0.33 |
| **LEARNING STAGES** | [Heads only, 4+ layers, all layers] | **[Heads only, 5+ layers, 4+ layers, 3+ layers, all layers]** |
| **NUM\_GPUS** | 1 | 1 |
| **BATCH\_SIZE** | 1 | 3 |
| **Number of Epochs** | [20, 20, 5] | [40, 15, 214, 210, 820] |
| **STEPS\_PER\_EPOCH** | 75 | 3 |
| **LEARNING\_RATE** | [1e-3, 1e-3, 5e-5] | **[1e-3,1e-3, 1e-2, 1e-2, 1e-2]** |
| **LEARNING MOMENTUM** | [0.9] | [0.7, 0.7, 0.9, 0.7, 0.9] |
| **WEIGHT\_DECAY** | 1e-4 | 1e-4 |
| **GRADIENT\_CLIP\_NORM** | 5.0 | **10.0** |
| **TRAIN\_ROIS\_PER\_IMAGE** | 400 | 400 |
| **DETECTION\_NMS\_THRESHOLD** | 0.3 | 0.3 |
| **DETECTION\_MIN\_CONFIDENCE** | 0.7 | **0.9** |
| **DETECTION\_MAX\_INSTANCES** | 400 | 400 |

**Table 3** Axonal attribute (eccentricity and area) distribution in the training, validation and testing datasets as a function of axon area distribution in the training dataset categorized into tertiles namely smaller area (0 to 1,304 sq pixels), medium area (1,304 to 2256 sq pixels) and larger area (2,256 to 1.1e6 sq pixels).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset | **Eccentricity:** Median / range in # pixels  (proportion in %) | | | **Area:** Median / range in sq pixels  (proportion in %) | | |
| Smaller Area | Medium Area | Larger Area | Smaller Area | Medium Area | Larger Area |
| **UT Training** | 20 / 12 to 60  (32.88%) | 28 / 20 to 52  (33.63%) | 40 / 28 to 140 (33.49%) | 960 / 24 to 1,303  (32.88%) | 1,675 / 1,304 to 2,252 (33.63%) | 3,288 / 2,256 to 35,288 (33.49%) |
| **UT Validation** | 20 / 12 to 76 (35.20%) | 28 / 20 to 44 (28.64%) | 40 / 28 to 176 (36.16%) | 952 / 172 to 1,296 (35.20%) | 1,703 / 1,304 to 2,249 (28.64%) | 3,367 / 2,256 to 19,745 (36.16%) |
| **UT Testing** | 20 / 12 to 284 (34.51%) | 28 / 20 to 48 (31.33%) | 40 / 28 to 108 (34.16%) | 984 / 32 to 1,301 (34.51%) | 1,704 / 1,304 to 2,248 (31.33%) | 3,472 / 2,256 to 20,856 (34.16%) |
| **JHU Testing** | 20 / 10 to 54 (27.82%) | 27 / 20 to 54 (38.77%) | 37 / 27 to 154 (33.41%) | 1,026 / 11 to 1,303 (27.82%) | 1,702 / 1,305 to 2,255 (38.77%) | 3,145 / 2,260 to 25,001 (33.41%) |

**Table 4** Axonal area distribution by axonal health status (necrotic or healthy) in the training, validation and testing datasets as a function of axon area distribution in the training dataset categorized into equal tertiles namely smaller area (0 to 1,304 sq pixels), medium area (1,304 to 2256 sq pixels) and larger area (2,256 to 1.1e6 sq pixels). In general, most necrotic axons (>45%) are of larger area and most healthy axons (> 60%) are of smaller area. In JHU testing dataset, a slightly higher proportion in the medium area for both necrotic and healthy axons i.e. higher overlap between the number of necrotic and healthy axons of medium area.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset | **Necrotic:** Median / range in sq pixels  (proportion in %) | | | **Healthy:** Median / range in sq pixels  (proportion in %) | | |
| Smaller Area | Medium Area | Larger Area | Smaller Area | Medium Area | Larger Area |
| **UT Training** | 1,008 / 24 to 1,303 (14.17%) | 1,769 / 1,304 to 2,252 (36.49%) | 3,296 / 2,256 to 35,288 (49.33%) | 944 / 222 to 1,299 (70.62%) | 1,480 / 1,304 to 2,208 (27.85%) | 2,664 / 2,256 to 4,304 (1.53%) |
| **UT Validation** | 995 / 172 to 1,296 (16.39%) | 1,792 / 1,304 to 2,249 (29.03%) | 3,380 / 2,256 to 19,745 (54.58%) | 930 / 256 to 1,296 (71.26%) | 1,528 / 1,304 to 2,216 (27.90%) | 2,440 / 2,280 to 4,728 (0.84%) |
| **UT Testing** | 1,040 / 32 to 1,301 (14.31%) | 1,792 / 1,304 to 2,248 (34.26%) | 3,496 / 2,256 to 20,856 (51.43%) | 968 / 248 to 1,298 (72.78%) | 1,520 / 1,304 to 2,200 (25.76%) | 2,376 / 2,272 to 5,453 (1.46%) |
| **JHU Testing** | 1,037 / 11 to 1,303 (13.33%) | 1,787 / 1,305 to 2,255 (39.38%) | 3,156 / 2,260 to 25,001 (47.30%) | 1,026 / 11 to 1,302 (61.73%) | 1,522 / 1,306 to 2,254 (37.35%) | 2,440 / 2,260 to 3,292 (0.91%) |

**Table 5** Axonal attribute (eccentricity and area) distribution in the training, validation and testing datasets as a function of axon eccentricity distribution in the training dataset categorized into tertiles namely smaller (0 to 23 pixels), medium area (23 to 28 pixels) and larger area (28 to 176 pixels). Bins are skewed by choice to optimal allocation of pyramid levels from the feature pyramid network (FPN) to the FCOS module for specializing in detecting necrotic (generally larger) and healthy axons.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset | Eccentricity(Ecc.)**:** Median / range in # pixels  (proportion in %) | | | Area**:** Median / range in sq pixels  (proportion in %) | | |
| Smaller Ecc. | Medium Ecc. | Larger Ecc. | Smaller Ecc. | Medium Ecc. | Larger Ecc. |
| **UT Training** | 20 / 12 to 20 (25.70%) | 24 / 24 to 24 (20.40%) | 32 / 28 to 140 (53.91%) | 912 / 222 to 1,424 (25.70%) | 1,392 / 536 to 2,184 (20.40%) | 2,562 / 24 to 35,288 (53.91%) |
| **UT Validation** | 20 / 12 to 20 (26.82%) | 24 / 24 to 24 (18.83%) | 36 / 28 to 176 (54.35%) | 875 / 171 to 1,400 (26.82%) | 1,360 / 552 to 1,992 (18.83%) | 2,712 / 358 to 19,745 (54.35%) |
| **UT Testing** | 20 / 12 to 20 (27.21%) | 24 / 24 to 24 (19.94%) | 36 / 28 to 284 (52.86%) | 928 / 248 to 1,400 (27.21%) | 1,392 / 659 to 2,104 (19.94%) | 2,649 / 32 to 20,856 (52.86%) |
| **JHU Testing** | 20 / 10 to 20 (18.99%) | 24 / 24 to 27 (35.97%) | 34 / 30 to 154 (45.04%) | 955 / 11 to 1,460 (18.99%) | 1,511 / 23 to 2,502 (35.97%) | 2,750 / 11 to 25,001 (45.04%) |

**Table 6** Axonal eccentricity distribution by axonal health status (necrotic or healthy) in the training, validation and testing datasets as a function of axon eccentricity distribution in the training dataset categorized into unequal / skewed tertiles namely smaller (0 to 23 pixels), medium (23 to 28 pixels) and larger eccentricities (28 to 176 pixels). Bins are skewed by choice to achieve optimal allocation of pyramid level features from the feature pyramid network (FPN) to the FCOS module within the CenterMask architecture for specializing in detecting necrotic (generally larger) and healthy axons. In the UT datasets (training, validation and testing), majority of the necrotic axons (>45%) have larger eccentricities and majority of the healthy axons (> 58%) have smaller eccentricities. In JHU testing dataset, significantly higher proportion of both necrotic and healthy axons have medium eccentricities i.e. higher overlap between the number of necrotic and healthy axons with a medium level of eccentricity.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset | **Necrotic:** Eccentricity (Ecc.) - Median / range in # pixels (proportion in %) | | | **Healthy:** Eccentricity (Ecc.) - Median / range in # pixels (proportion in %) | | |
| Smaller Ecc. | Medium Ecc. | Larger Ecc. | Smaller Ecc. | Medium Ecc. | Larger Ecc. |
| ***UT Training*** | 20 / 12 to 20 (9.45%) | 24 / 24 to 24 (15.90%) | 32 / 28 to 140 (74.65%) | 20 / 12 to 20 (58.47%) | 24 / 24 to 24 (29.45%) | 28 / 28 to 64 (12.07%) |
| ***UT Validation*** | 20 / 12 to 20 (10.40%) | 24 / 24 to 24 (13.56%) | 36 / 28 to 176 (76.03%) | 20 / 12 to 20 (58.31%) | 24 / 24 to 24 (28.94%) | 28 / 28 to 52 (12.75%) |
| ***UT Testing*** | 20 / 12 to 20 (9.68%) | 24 / 24 to 20 (15.08%) | 36 / 28 to 108 (75.24%) | 20 / 12 to 20 (60.40%) | 24 / 24 to 24 (29.13%) | 28 / 28 to 284 (10.46%) |
| ***JHU Testing*** | 20 / 10 to 20 (7.84%) | 27 / 24 to 27 (30.48%) | 37 / 30 to 154 (61.68%) | 20 / 10 to 20 (45.07%) | 24 / 24 to 27 (48.80%) | 30 / 30 to 57 (6.12%) |

**Table 7** Performance of the Mask R-CNN model MR-2 on UT testing and JHU testing datasets. AP50: area under the precision-recall curve at IOU=0.5. AP50to95: Mean of AP values for IoU=0.5, 0.55, …, 0.95. R50: Recall value at IoU = 0.5. AR50to95: Average of recall values at IoU=0.5, 0.55, …, 0.95. AP and AR performance of CenterMask2 on UT Testing dataset is similar as those of the UT Validation dataset. Performance on the external evaluation dataset JHU Testing is much higher than that of the observed performance on UT Validation and UT Testing datasets.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | Metric | Overall | Small Area | Medium Area | Large Area | Necrotic | Healthy |
| UT Validation | **AP50** | 85.27% | 69.87% | 88.83% | 71.26% | 82.29% | 87.59% |
| **AP50to95** | 41.67% | 26.40% | 43.94% | 41.49% | 48.48% | 34.86% |
| **R50** | 87.43% | 72.73% | 91.20% | 90.54% | 84.37% | 90.49% |
| **AR50to95** | 48.27% | 31.15% | 52.91% | 59.76% | 53.21% | 43.32% |
| UT Testing | **AP50** | 88.97% | 81.92% | 90.86% | 62.04% | 87.40% | 90.55% |
| **AP50to95** | 44.32% | 32.70% | 44.87% | 37.20% | 51.78% | 36.86% |
| **R50** | 92.18% | 83.32% | 95.21% | 87.66% | 89.43% | 94.94% |
| **AR50to95** | 51.17% | **37.28%** | **54.83%** | 60.78% | 56.54% | 45.80% |
| JHU Testing | **AP50** | 85.67% | 70.38% | 83.22% | 77.99% | 87.25% | 84.09% |
| **AP50to95** | **52.58%** | **32.79%** | **47.61%** | **49.44%** | **58.83%** | **46.33%** |
| **R50** | 88.90% | 72.61% | 84.63% | 89.69% | 88.94% | 88.87% |
| **AR50to95** | **59.58%** | 36.80% | 53.26% | **64.23%** | **63.56%** | **55.59%** |

**Table 8** In Centermask models, FCOS parameters related to the pyramidal multiscale features were optimized based on the distribution of axonal geometries in the UT training dataset. In model CM-1, default FCOS parameters (mainly SIZES\_OF\_INTEREST) were used and in model CM-2, these parameters were optimized based on axonal geometries.

|  |  |  |
| --- | --- | --- |
| Model Parameters | Model CM-1 | Model CM-2 |
| **VOVNET.CONV\_BODY** | V-39-eSE | V-39-eSE |
| **VOVNET.OUT\_FEATURES** | [“stage3”, “stage4”, “stage5”] | **[“stage2”, “stage3”, “stage4”]** |
| **FPN.IN\_FEATURES** | [“stage3”, “stage4”, “stage5”] | **[“stage2”, “stage3”, “stage4”]** |
| **PROPOSAL GENERATOR** | “FCOS” | “FCOS” |
| **FCOS.IN\_FEATURES** | [“p3”, “p4”, “p5”, “p6”, “p7] | **[“p2”, “p3”, “p4”]** |
| **FCOS.FPN\_STRIDES** | [8, 16, 32, 64, 128] | **[4, 8, 16]** |
| **FCOS.SIZES\_OF\_INTEREST** | [32, 256, 512, 2048] | **[23, 28]** |
| **FCOS.TOP\_LEVELS** | 2 | **0** |
| **FCOS.PRE\_NMS\_TOPK\_TRAIN** | 1200 | 1200 |
| **FCOS.POST\_NMS\_TOP\_TRAIN** | 600 | 600 |
| **FCOS.PRE\_NMS\_TOPK\_TEST** | 1200 | 1200 |
| **FCOS.POST\_NMS\_TOPK\_TEST** | 600 | 600 |
| **FCOS.INFERENCE\_TH\_TRAIN** | 0.3 | 0.3 |
| **FCOS.INFERENCE\_TH\_TEST** | 0.7 | 0.7 |
| **FCOS.NMS\_TH** | 0.3 | 0.3 |
| **FCOS.NUM\_CLASSES** | 2 | 2 |
| **FCOS.PRIOR\_PROB** | 0.005 | 0.005 |
| **PIXEL\_MEAN** | 150.9695 | 150.9695 |
| **PIXEL\_STD** | 65.8090 | 65.8090 |
| **ROI\_HEADS.NUM\_CLASSES** | 2 | 2 |
| **ROI\_HEADS.IN\_FEATURES** | [“p3”, “p4”, “p5”] | **[“p2”, “p3”, “p4”]** |
| **ROI\_HEADS.BATCH\_SIZE\_PER\_IMAGE** | 600 | 600 |
| **ROI\_HEADS.POSITIVE\_FRACTION** | 0.5 | 0.5 |
| **SOLVER.IMS\_PER\_BATCH** | 3 | 3 |
| **SOLVER.BASE\_LR** | 0.01 | 0.01 |
| **SOLVER.GAMMA** | 0.1 | 0.1 |
| **SOLVER.MOMENTUM** | 0.7 | 0.7 |
| **SOLVER.CLIP\_GRADIENTS.CLIP\_TYPE** | Norm | Norm |
| **SOLVER.CLIP\_GRADIENTS.NORM\_TYPE** | 2.0 | 2.0 |
| **SOLVER.CLIP\_GRADIENTS.CLIP\_VALUE** | 10.00 | 10.0 |
| **INPUT.MIN\_SIZE\_TRAIN** | (768, 1024) | (768, 1024) |
| **INPUT.MIN\_SIZE\_TEST** | 1024 | 1024 |
| **TEST.DETECTIONS\_PER\_IMAGE** | 600 | 600 |

**Table 9** Performance of the CenterMask model CM-2 using optimized model parameters. AP50: area under the precision-recall curve at IOU=0.5. AP50to95: Mean of AP values for IoU=0.5, 0.55, …, 0.95. R50: Recall value at IoU = 0.5. AR50to95: Average of recall values at IoU=0.5, 0.55, …, 0.95. AP and AR performance on the *UT Testing* dataset is similar as those of the *UT Validation* dataset. Performance on the external evaluation dataset *JHU Testing* is much higher than that of the observed performance on UT Validation and UT Testing datasets.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | Metric | Overall | Small Area | Medium Area | Large Area | Necrotic | Healthy |
| UT Validation | **AP50** | 55.93% | 41.84% | 42.58% | 41.36% | 65.39% | 46.47% |
| **AP50to95** | 30.94% | 19.92% | 23.97% | 25.91% | 38.81% | 23.06% |
| **R50** | 69.73% | 53.18% | 65.06% | 49.86% | 76.14% | 63.32% |
| **AR50to95** | 43.08% | 28.59% | 40.88% | 35.16% | 50.34% | 35.82% |
| UT Testing | **AP50** | 56.45% | 46.02% | 44.31% | 40.97% | 67.90% | 45.01% |
| **AP50to95** | 28.86% | 20.27% | 22.29% | 24.70% | 38.28% | 19.44% |
| **R50** | 71.30% | 58.94% | 67.28% | 48.38% | 80.29% | 62.32% |
| **AR50to95** | 41.52% | 30.37% | 38.75% | 31.85% | 50.75% | 32.39% |
| JHU Testing | **AP50** | 59.41% | 39.17% | 52.40% | 45.39% | 68.80% | 50.02% |
| **AP50to95** | 32.92% | 18.10% | 26.81% | 27.27% | 41.18% | 24.66% |
| **R50** | 73.91% | 46.77% | 67.47% | 67.66% | 79.70% | 68.13% |
| **AR50to95** | 45.74% | 23.45% | 39.39% | 43.95% | 52.40% | 39.07% |

**Table 10** An independent Mask R-CNN model (MR-3) initialized with pretrained ImageNet weights and trained using *UT Training* dataset. AP50: area under the precision-recall curve at IOU=0.5. AP50to95: Mean of AP values for IoU=0.5, 0.55, …, 0.95. R50: Recall value at IoU = 0.5. AR50to95: Average of recall values at IoU=0.5, 0.55, …, 0.95. Performance of the independent Mask R-CNN model MR-3 was lower than that of the optimized Mask R-CNN model MR-2, but significantly higher than that of the optimized CenterMask model CM-2.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | Metric | Overall | Small Area | Medium Area | Large Area | Necrotic | Healthy |
| UT Validation | **AP50** | 65.83% | 50.88% | 55.76% | 47.52% | 75.06% | 56.60% |
| **AP50to95** | 39.89% | 27.11% | 34.63% | 32.44% | 49.56% | 30.23% |
| **R50** | 72.45% | 54.51% | 66.54% | 57.31% | 78.65% | 66.25% |
| **AR50to95** | 47.79% | 31.17% | 45.54% | 42.69% | 55.17% | 40.41% |
| UT Testing | **AP50** | 67.87% | 56.06% | 55.98% | 45.08% | 77.51% | 58.23% |
| **AP50to95** | 41.48% | 30.87% | 34.95% | 30.96% | 51.74% | 31.22% |
| **R50** | 75.50% | 60.32% | 69.14% | 50.67% | 81.83% | 69.18% |
| **AR50to95** | 50.22% | 35.62% | 47.95% | 37.68% | 57.56% | 42.88% |
| JHU Testing | **AP50** | 68.17% | 50.71% | 61.36% | 49.84% | 77.71% | 58.62% |
| **AP50to95** | 37.75% | 22.87% | 32.18% | 30.36% | 47.85% | 27.64% |
| **R50** | 75.62% | 53.07% | 65.86% | 68.15% | 80.72% | 70.52% |
| **AR50to95** | 46.45% | 24.97% | 37.83% | 45.29% | 53.45% | 39.45% |

**Table 11** Performance of the CenterMask model CM-1 using default model parameters. AP50: area under the precision-recall curve at IOU=0.5. AP50to95: Mean of AP values for IoU=0.5, 0.55, …, 0.95. R50: Recall value at IoU = 0.5. AR50to95: Average of recall values at IoU=0.5, 0.55, …, 0.95. AP and AR performance of the model CM-1 were similar among *UT Validation, UT Testing,* and *JHU Testing* datasets. Additionally, performance of the default model CM-1 was similar to that of the optimized model CM-2.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | Metric | Overall | Small Area | Medium Area | Large Area | Necrotic | Healthy |
| UT Validation | **AP50** | 54.76% | 37.75% | 42.24% | 42.29% | 66.08% | 43.44% |
| **AP50to95** | 31.86% | 19.35% | 24.47% | 27.08% | 40.88% | 22.84% |
| **R50** | 65.46% | 46.60% | 59.38% | 49.31% | 72.93% | 57.99% |
| **AR50to95** | 41.72% | 26.68% | 38.22% | 35.46% | 49.08% | 34.36% |
| UT Testing | **AP50** | 55.97% | 40.78% | 42.88% | 44.58% | 67.98% | 43.94% |
| **AP50to95** | 29.91% | 18.32% | 22.82% | 27.10% | 39.95% | 19.86% |
| **R50** | 68.77% | 51.10% | 63.06% | 49.41% | 76.90% | 60.63% |
| **AR50to95** | 41.08% | 27.09% | 38.37% | 32.10% | 49.49% | 32.67% |
| JHU Testing | **AP50** | 55.73% | 34.52% | 44.75% | 45.96% | 66.48% | 44.97% |
| **AP50to95** | 30.76% | 15.82% | 22.63% | 27.05% | 39.36% | 22.15% |
| **R50** | 67.48% | 40.80% | 56.91% | 63.68% | 72.36% | 62.60% |
| **AR50to95** | 41.43% | 20.59% | 32.95% | 40.72% | 46.91% | 35.95% |

|  |  |  |  |
| --- | --- | --- | --- |
|  | A close-up of a black and white image  AI-generated content may be incorrect. | A close-up of a grey surface  AI-generated content may be incorrect. |  |
|  | **UT Val 1** | **UT Val 2** |  |
| **UT Val 1** | A close-up of a colorful pattern  AI-generated content may be incorrect. | A close-up of a colorful image  AI-generated content may be incorrect. | A close-up of a colorful pattern  AI-generated content may be incorrect. |
|  | | | |
| **UT Val 2** | A close-up of a red and green cell  AI-generated content may be incorrect. | A close-up of a colorful pattern  AI-generated content may be incorrect. | A close-up of a grey surface  AI-generated content may be incorrect. |
|  | a) True axonal locations and geometries | b) Optimal Mask R-CNN model MR-2 predictions | c) Optimal CenterMask model CM-2 predictions |

**Figure 5**. Two example confocal imaging sections of rat optic nerve from the *UT Validation* dataset (UT Val 1, 2); (a) true axonal locations and geometries; (b) predictions from the optimized Mask R-CNN MR-2 model; and (c) from the optimized CenterMask CM-2 instance segmentation methods.

|  |  |  |  |
| --- | --- | --- | --- |
|  | A close-up of a rock  AI-generated content may be incorrect. | A close-up of a grey surface  AI-generated content may be incorrect. |  |
|  | **UT Test 1** | **UT Test 2** |  |
| **UT Test 1** | A close-up of a red and green pattern  AI-generated content may be incorrect. | A close-up of a rock  AI-generated content may be incorrect. | A close-up of a grey surface  AI-generated content may be incorrect. |
|  |  |  |  |
| **UT Test 2** | A close-up of a colorful pattern  AI-generated content may be incorrect. | A close-up of a grey surface  AI-generated content may be incorrect. | A close-up of a map  AI-generated content may be incorrect. |
|  | a) True axonal locations and geometries | b) Model MRCNN (416\_03) predictions | c) Model CM1 (819\_2\_init) predictions |

**Figure 6**. Two example confocal imaging sections of rat optic nerve from the *UT Testing* dataset (UT Test 1, 2); (a) true axonal locations and geometries; (b) predictions from the optimized Mask R-CNN MR-2 model; and (c) from the optimized CenterMask CM-2 instance segmentation methods.

|  |  |  |  |
| --- | --- | --- | --- |
|  | A close-up of a grey surface  AI-generated content may be incorrect. | A close-up of a black and white image  AI-generated content may be incorrect. |  |
|  | **JHU Test 1** | **JHU Test 2** |  |
| **JHU Test 1** | A close-up of a microscope  AI-generated content may be incorrect. | A close-up of a microscope  AI-generated content may be incorrect. | A close-up of a grey surface  AI-generated content may be incorrect. |
|  |  |  |  |
| **JHU Test 2** | A close-up of a red and green pattern  AI-generated content may be incorrect. | A close-up of a colorful pattern  AI-generated content may be incorrect. | A close-up of a colorful pattern  AI-generated content may be incorrect. |
|  | a) True axonal locations and geometries | b) Model MRCNN (416\_03) predictions | c) Model CM1 (819\_2\_init) predictions |

**Figure 7**. Two example confocal imaging sections of rat optic nerve from the *JHU Testing* dataset (JHU Test 1, 2); (a) true axonal locations and geometries; (b) predictions from the optimized Mask R-CNN MR-2 model; and (c) from the optimized CenterMask CM-2 instance segmentation methods.