CS230

Evaluation of U-Net vs. U-Net + VGG-16 Architecture Performance in Extracting Masks and Musculoskeletal Metrics from Ultrasound B-Mode Image Data

Shreya Narayan
Department of Electrical Engineering
Stanford University
shreyavn@stanford.edu

Abstract

Because extensive training is required to be able to read ultrasound images, automated segmentation of musculoskeletal ultrasound data as well as automated extraction of musculoskeletal metrics such as muscle thickness is a valuable tool. Specifically, this would enable patient diagnosis to be expedited and more accurate, as well as allow medical imaging to be more accessible to patients in resource limited areas. Ultrasound image data is often segmented using a U-Net architecture. However, U-Nets have limitations in modeling long term dependencies because convolutions are computed locally [2, 11]. This paper provides a U-Net + VGG-16 architecture that combines the benefits of a VGG-16 model with the advantages of a traditional U-Net Architecture. The U-Net + VGG-16 architecture provides improved results in mask accuracy created from a B-Mode image and provides improved results in extracting musculoskeletal metrics from the mask created by the algorithm.

1 Introduction

Ultrasound imaging has many advantages over other medical imaging modalities: it is relatively inexpensive, does not cause radiation exposure to the patient, can be used on patients with metallic implants and pacemakers, and can be made portable. In contrast to other imaging modalities, ultrasound also provides information about dynamic processes in the body, since imaging can be done over time [1]. Ultrasound is used commonly for musculoskeletal imaging because it is a very good method for visualizing soft tissue [1].

However, one disadvantage of ultrasound imaging is that reading the B-mode image requires expertise and is mostly done manually. This requirement for manual interpretation creates a bottleneck in the pipeline for diagnosis and treatment of medical conditions. Automated segmentation can help solve this problem and increase the efficiency of the diagnosis pipeline. By increasing this efficiency, ultrasound can in future be used in remote settings or without the need of a bedside sonographer.

In this paper, the input to the neural network is a B-Mode ultrasound image. I then use a U-Net + VGG-16 architecture to output a mask of the image, with ones in pixels where an aponeurosis or fascicle is present and zeros elsewhere. An additional algorithm takes as input the mask created by this model and outputs the fascicle[1] length, aponeurosis[2] thickness, and pennation angle[3] of muscle fibers in the image.

2 Related Work

Deep learning methods for automated segmentation of ultrasound images have included various neural network architectures, including convolutional neural networks, U-Nets, and residual neural networks (for dynamic ultrasound data) [6]. Convolutional neural networks are effective and can be made lightweight as in [12], but require large datasets, which are hard to find for ultrasound data. Recurrent neural nets or sequential prediction networks are effective when dynamic video data must be analyzed and prior images must inform the current image [13]. However, these methods are not as helpful for single image data. The U-Net architecture has been used most often with medical image data because of its ability to capture more contextual information, fast training speed, and the smaller data set size that can be used [9]. However, the U-Net architecture shows some limitations in clearly modeling dependencies because it computes convolutions locally. This makes it difficult for U-Nets to model long-range dependency. For example, in [2], the U-Net architecture used was able to generate somewhat accurate analyses of certain musculoskeletal diagnostic metrics using the extracted mask from the B-Mode data, but it relied on averaging the predictions for locations of muscle fibers to make a metric prediction. The underlying mask

[1] A fascicle is a group of muscle fibers that are bundled as a unit within a whole muscle [5]
[2] An aponeurosis is a thin sheath of connective tissue that helps connect your muscles to your bones [3]
[3] Pennation angle refers to the angle between the longitudinal axis of the entire muscle and its fibers [4].

produced was in fact often inconsistent between successive frames in a video, which should not be the case, given that successive frames are almost identical. This led to inaccuracies with some metric predictions [2]. Instead, effective U-Nets are the product of combining the U-Net architecture with another neural network, as was done cleverly with transformers in [11].

This paper aims to see if the U-Net architecture can be improved by combining it with a VGG-16 architecture. In particular, the U-Net architecture proposed by Cronin et. al [2] is evaluated against a proposed U-Net + VGG-16 architecture. The performance of the neural network is measured by the network's ability to generate accurate masks of aponeuroses and fascicles from B-Mode ultrasound data. Then, the model result is used by an algorithm that estimates muscle thickness, pennation angle, and fascicle length of muscle tissue. The estimates for these metrics with both methods are compared to evaluate the performance of each model.

3 Dataset and Features

The data set (obtained from [2]) includes 840 anonymized images: 570 images for the aponeurosis model and 310 images for muscle fascicle model. The anonymized images are taken from single image and frames of video data obtained from various muscles including: the medial and lateral gastrocnemius (calf), vastus lateralis (quadricep/thigh), and tibialis anterior (front of calf). Four different ultrasound probes were used to collect the data; this is important for adding variety to the training data since the different settings of the probes affects how the resulting B Mode image is displayed. Data was collected from three different populations: athletes, elderly individuals, and young/healthy individuals. The images were taken when the subjects were doing different movements and contraction types [2].

The data set is augmented by using elastic deformations on the training images. This is helpful because ultrasound images of the same muscle are often different based on the patient's position and the orientation of the probe. Thus elastic deformations make the training set more robust because it simulates these variations as well as muscle contraction [2].

For each of the aponeurosis and fascicle images, researchers from [2] manually identified all instances of aponeuroses and fascicles to create a binary mask (white pixels belonging to the aponeurosis/fascicle are white, elsewhere as black). These binary masks were used as ground truth labels for the data that could then be used when training the neural networks [2].

The data was split 90/10 for training and validation respectively. This decision of 90/10 was made because of the smaller size of the datasets available. This split reflects a limitation of the dataset, which is that there is a lack of data available that is both of the desired region and has a paired mask available for training.

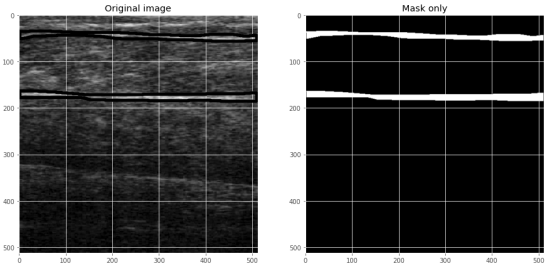


Figure 1: *Sample Image of Ultrasound B Mode image and ground truth mask of aponeurosis*

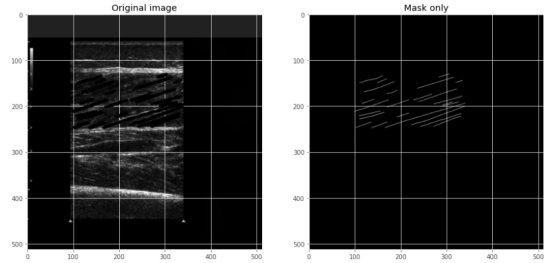


Figure 2: *Sample image of Ultrasound B Mode Image and ground truth mask of fascicle*

4 Methods

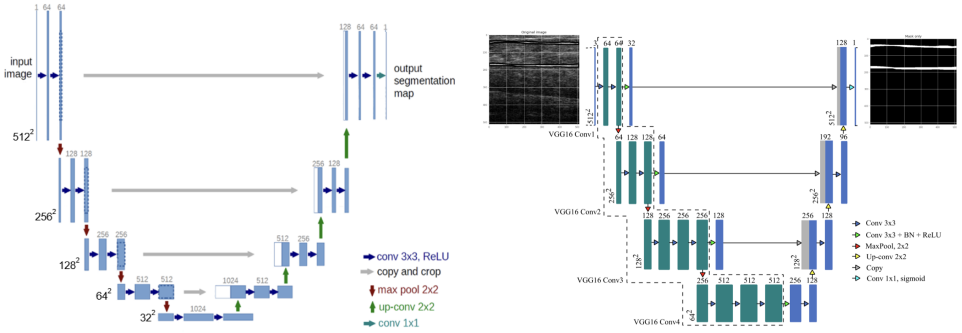
Baseline Model: U-Net Architecture

A baseline model was created based on work by [2]. This model is a U-Net architecture and is a supervised learning approach (see Figure 3a). The contracting path consists of the repeated application of two 3 x 3 unpadded convolutions followed by a ReLU and 2x2 max pooling operation with stride 2 for downsampling and a double in the number of feature channels with each downsampling step. Dropout is applied at each step. In the expansive path, the feature map is upsampled and we apply a 2x2 up-convolution (which halves the number of feature channels), concatenation with the corresponding feature map in the contracting path, a dropout layer, and two 3x3 convolutions, each followed by a ReLU activation. Finally, a 1x1 convolution maps each feature vector to one of two classes. The output is a mask with pixel wise binary labels that represent whether an aponeurosis or fascicle is present [2].

To train this model, images are first resized to 512 x 512 pixels. A larger initial size helps with training in ultrasound images because the spatial resolution of ultrasound images is already lower than other modalities. A 90/10 training/validation split was used with 25 epochs, early stopping, a batch size of 2, Adam optimizer, and a binary cross-entropy loss function [2]. Learning rate was reduced until a minimum of 0.00001 with factor 0.1 when performance plateaued for 10 epochs. Binary cross entropy loss (formula in appendix) is used because pixels are classified as either 1 or 0 based on whether or not they are part of the aponeurosis or fascicle.

In post processing, we extract muscle thickness, fascicle length, and pennation angle from the image. Aponeuroses below a threshold of 0.5 are removed. Remaining aponeuroses are extrapolated laterally to find intersection with muscle fascicles. The fascicle model identifies fascicle fragments and the extrapolation helps to determine a full structure of the fascicle. By determining the intersection between fascicles and aponeuroses, we can determine the fascicle length. Pennation angle is then computed between each fascicle and the local slope of the lower aponeurosis (defined in the study as the 50 pixel region starting from the point of intersection). Muscle thickness is determined as the shortest distance between the surface level and deep aponeuroses [2].

To determine how the model performs, we test the model on a validation set that has ground truth masks. The accuracy of the mask is evaluated by computing intersection over union of the generated mask with the ground truth mask [2].



a) U-Net architecture b) U-Net +VGG-16 architecture [8]

Figure 3: *Side by side U Net and VGG-16 architectures*

<u>Key Insight and New Architecture: U-Net + VGG-16 Implementation</u>

The model proposed in this paper is a U-Net + VGG-16 architecture with batch normalization. The VGG-16 architecture contains layers pre-trained on ImageNet. Weights from the ImageNet model are not used in this architecture. Since VGG-16 has already learned features, the performance of the network should improve [7].

$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{\text{Intersection}}{\text{Detected box}}$$

In this implementation, the fully connected layers of the VGG-16 architecture are removed because semantic segmentation on the entire image is desired [10]. The architecture contains: a pre-trained VGG-16 encoder from which the desired feature maps are extracted, and a decoder (Figure 4) that contains a 2x2 transpose convolution layer, a skip connection from the

pretrained VGG-16 architecture, and a convolution block. The convolution block contains two 3x3 convolution layers, a batch normalization layer, and a ReLU activation function. The bridge between the contracting and expansive paths is a layer from the pre-trained VGG-16 network rather than a convolution block in the traditional U-Net architecture [7]. A 2x2 max pooling layer is applied in the contracting path. As was used in the U-Net, binary cross entropy loss was used.

5 Results

Hyperparameter Tuning

Batch normalization was applied in the proposed U-Net + VGG-16 architecture in order to prevent overfitting. Learning rates of 0.01, 0.001, 0.0001, 0.00001, as well as reducing learning rate upon plateau were tested to refine the model and reduce bias while preventing overfitting. Testing showed that learning rate and batch normalization were both sensitive parameters for the mode; a static learning rate of 0.0001 was deemed best to avoid overfitting while maximizing performance of the model. A batch size of 2 was chosen to balance training speed and model generalizability. Intersection Over Union (IOU, Jaccard Index) was the primary metric used to evaluate performance of the model. The best performing model and associated metrics are shown below (plots and histograms in Appendix).

Model	Training Loss	Validation Loss	Training IOU	Validation IOU
Traditional U-Net	.0253	0.0375	0.9906	0.9905
U-Net + VGG-16, LR = 0.0001; aponeurosis model	.0222	0.0276	0.9914	0.9917
U-Net + VGG-16, LR = 0.0001; fascicle model	.0058	0.0131	0.9969	0.9961

Metric Extraction:

Finally, we observe the model's performance over two successive frames in a video. We feed the model into an algorithm that then extracts fascicle length, muscle thickness, and pennation angle for a given image. We analyze consistency in mask and metric prediction between successive frames. Enlarged images are shown in the appendix.

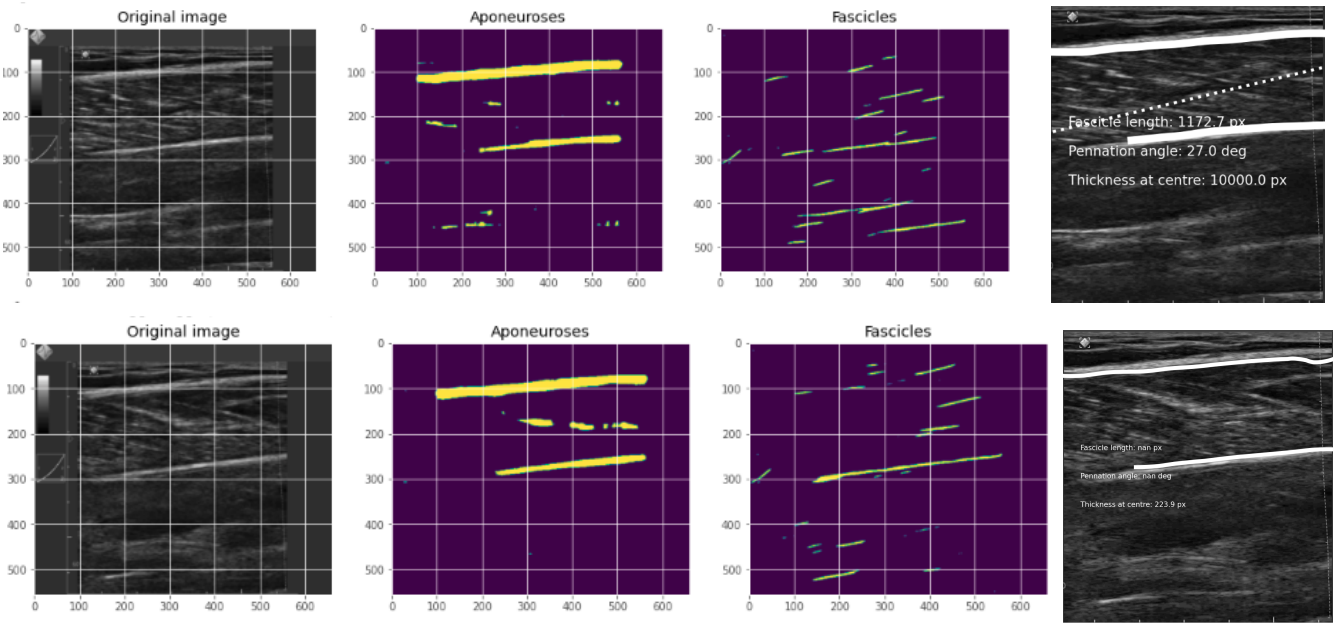


Figure 4 (above): *U-Net model from [2]. Top: Image 349. Thickness is incorrect. Bottom: Image 350. Differing number of fascicles detected. No measurements extractable for fascicles for image 350.*

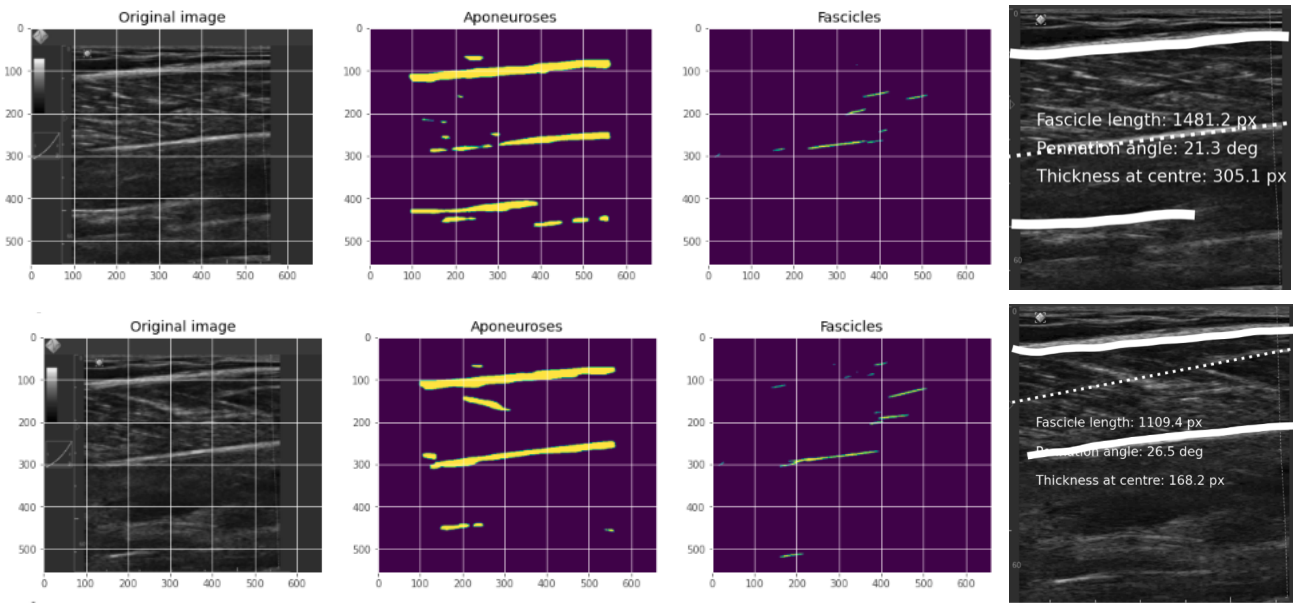


Figure 5: *Developed U-Net + VGG-16 model. Top: Image 349. Bottom: Image 350. Masks and measurements are more similar between the two successive frames and fascicle measurements are extractable from image 350. Similar fascicles are detected in image 350 as in 349.*

Discussion

The results show improved performance in mask extraction in the U-Net + VGG-16 models, as shown in validation loss and IOU metrics shown above. In metric extraction, the problems associated with detecting fascicles over successive frames in a video are significantly improved with the proposed U-Net + VGG-16 model. Measurements can be extracted where they could not be with the U-Net model (no NaNs are reported) and the same number of fascicles are extracted between successive frames.

6 Conclusion

Today, it is very difficult to read an ultrasound image without a healthcare professional. Enabling deep learning methods to read and produce diagnostically relevant information from ultrasound images can enable remote ultrasound imaging and ultrasound imaging in resource limited areas to be much more accessible and helpful for patient care. The use of a U-Net + VGG-16 with batch normalization and a learning rate 0.0001 for semantic segmentation benefits from pretrained layers of VGG-16 while also benefiting from the U-Net advantages of capturing contextual data, faster training speed, and ability to train from a smaller data set. The architecture is an effective automated way to obtain musculoskeletal metrics from B-Mode image data and is an improved neural network over the traditional U-Net model alone. Future work would benefit from incorporating a recurrent neural network to improve the accuracy of metrics from frame to frame in video data.

7 Contributions:

All components were completed by Shreya Narayan.

References

[1] Shah AB, Bhatnagar N. (July 2019). Ultrasound imaging in musculoskeletal injuries-What the Orthopaedic surgeon needs to know. *J Clin Orthop Trauma*. PMID: 31316235; PMCID: PMC6611988.Retrieved from https://pubmed.ncbi.nlm.nih.gov/31316235/

[2] Cronin, N., Finni, T., Seynnes, O. (September 2020) Fully automated analysis of muscle architecture from B-mode ultrasound images with deep learning. *Electrical Engineering and Systems Science → Image and Video Processing.* Retrieved from https://arxiv.org/pdf/2009.04790v1.pdf

[3] Cleveland Clinic. Aponeurosis. Last updated 2022. Retrieved from https://my.clevelandclinic.org/health/body/23407-aponeurosis

[4] Lee D, Li Z, Sohail QZ, Jackson K, Fiume E, Agur A. A three-dimensional approach to pennation angle estimation for human skeletal muscle (2015). Comput Methods Biomech Biomed Engin.;18(13):1474-84. doi: 10.1080/10255842.2014.917294. Epub 2014 May 21. PMID: 24849037. Retrieved from https://www.cs.toronto.edu/pub/reports/na/Lee.PA.2013.pdf

[5] National Cancer Institute. Structure of Skeletal Muscle. Retrieved from https://training.seer.cancer.gov/anatomy/muscular/structure.html

[6[Shereena, V. Medical Ultrasound Image Segmentation Using U-Net Architecture. (July 2022) *International Conference on Advances in Computing and Data Sciences*. Retrieved from https://link.springer.com/chapter/10.1007/978-3-031-12638-3_30

[7] Tomar, Nikhil. (December 2021). VGG16 UNET Implementation in TensorFlow. Idiot Developer. Retrieved from https://idiotdeveloper.com/vgg16-unet-implementation-in-tensorflow/

[8] Kanaeva, I. Ivanova, J. (January 2021). Road pavement crack detection using deep learning with synthetic data. *IOP Conference Series Materials Science and Engineering.* Retrieved from https://www.researchgate.net/publication/348661034_Road_pavement_crack_detection_using_deep_learning_with_synthetic_data

[9] Yin, X., Sun, L. Fu, Y., Lu, R., Zhang, Y. (April 2022). U-Net Based Medical Image Segmentation. *Journal of Healthcare Engineering.* Retrieved from https://www.hindawi.com/journals/jhe/2022/4189781/

[10] Le, James (2020). How to Do Semantic Segmentation Using Deep Learning. *Nanonets*. Retrieved from https://nanonets.com/blog/how-to-do-semantic-segmentation-using-deep-learning/

[11] Chen, J. Lu, Y., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A., Zhou, Y. (February 2021). TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. Retrieved from: https://arxiv.org/abs/2102.04306

[12] Malhotra P, Gupta S, Koundal D, Zaguia A, Enbeyle W. (March 2022). Deep Neural Networks for Medical Image Segmentation. *J Healthc Eng*.;2022:9580991. doi: 10.1155/2022/9580991. PMID: 35310182; PMCID: PMC8930223. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8930223/

[13] Zeng W, Luo J, Cheng J, Lu Y. (August 2022). Efficient fetal ultrasound image segmentation for automatic head circumference measurement using a lightweight deep convolutional neural network. Med Phys;49(8):5081-5092. doi: 10.1002/mp.15700. PMID: 35536111. Retrieved from https://pubmed.ncbi.nlm.nih.gov/35536111/

[14] Image Caption: Baeldung. (September 2022). Intersection over Union. Retrieved from https://www.baeldung.com/cs/object-detection-intersection-vs-union

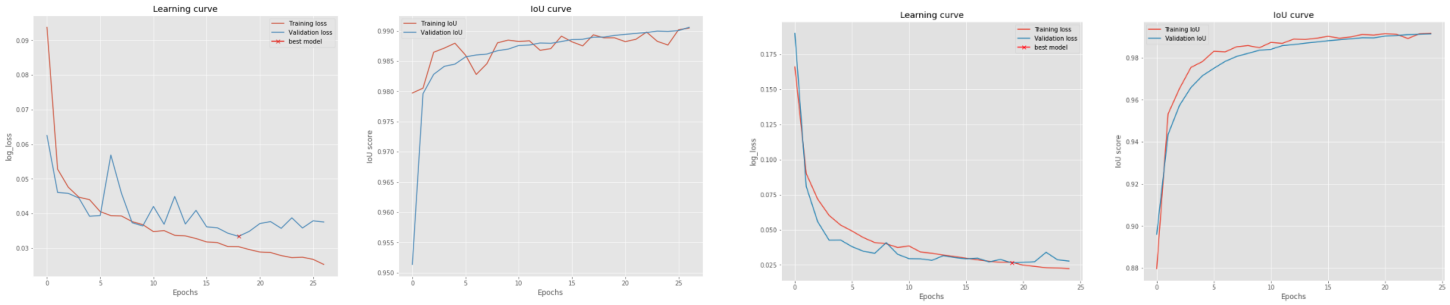
Appendix

1. Binary cross entropy loss formula:

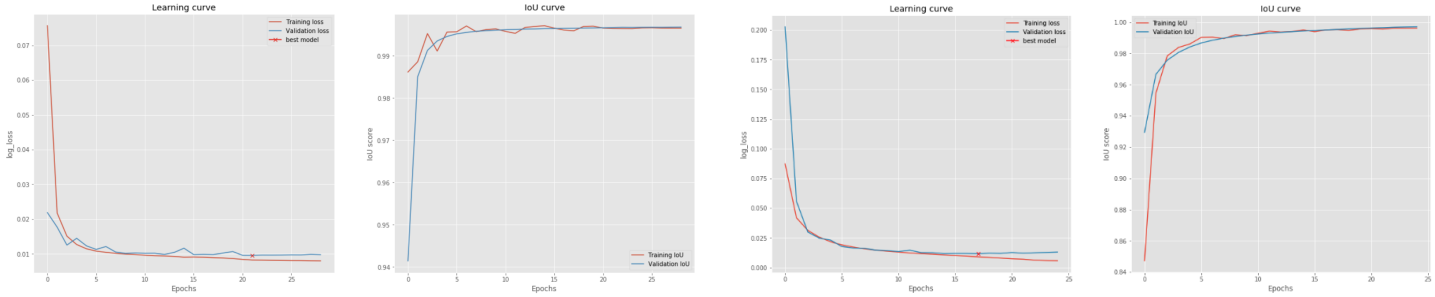
$$L_{BCE} = -\frac{1}{n}\sum_{i=1}^{n}(Y_i \cdot \log \hat{Y}_i + (1 - Y_i) \cdot \log (1 - \hat{Y}_i))$$

2. Best U-Net + VGG-16 Model: Plots of Training

Below are results for the baseline U-Net model alongside the results for the proposed U-Net + VGG-16 architecture with optimal learning rate (LR = 0.0001 for aponeurosis and fascicles).



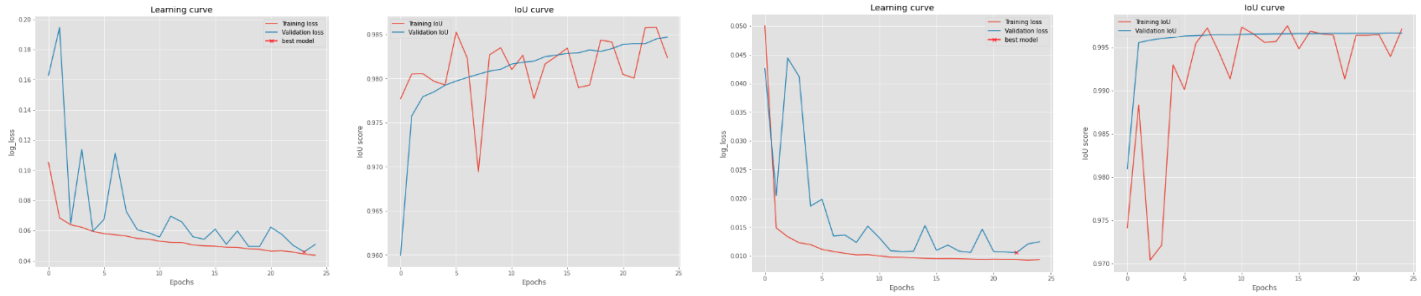
Aponeurosis Training Results. Plots 1 & 2: U-Net model. Plots 3 & 4: U-Net +VGG-16 model.



Fascicle Training Results. Plots 1 & 2: U-Net model. Plots 3 & 4: U-Net +VGG-16 model.

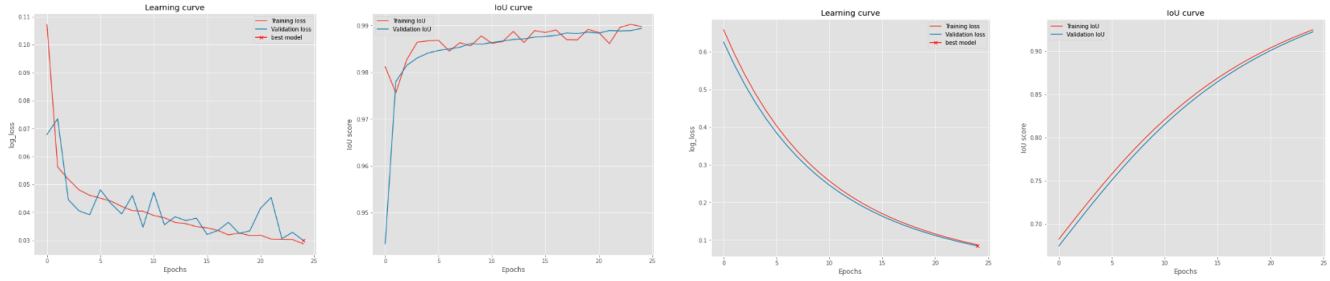
3. Learning Rate Experiments: Graphs showing resulting loss and IOU for various learning rates tested

Learning Rate = 0.01:



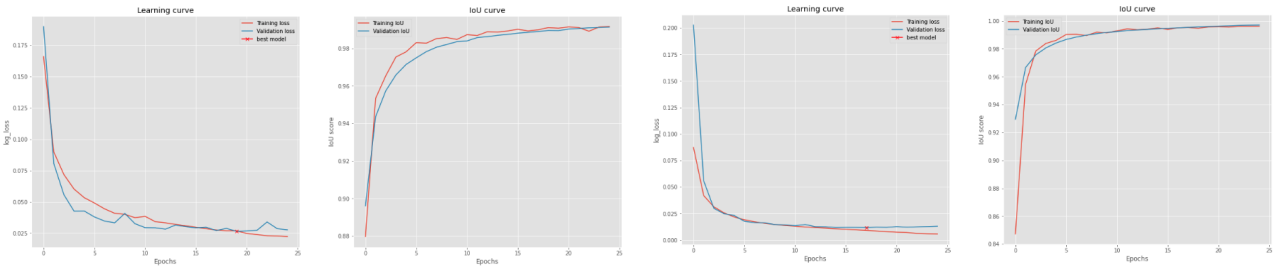
Aponeurosis Model Fascicle Model

Learning Rate = 0.001:



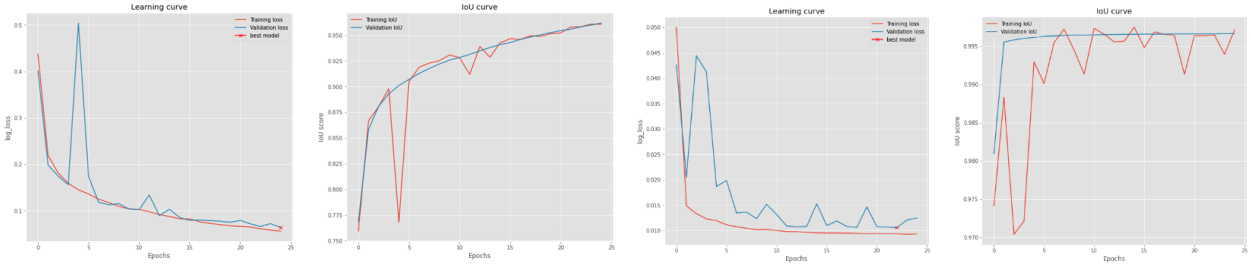
Aponeurosis Model Fascicle Model

Learning Rate = 0.0001:



Aponeurosis Model Fascicle Model

Learning Rate = 0.00001:



Aponeurosis Model Fascicle Model

Table of values:

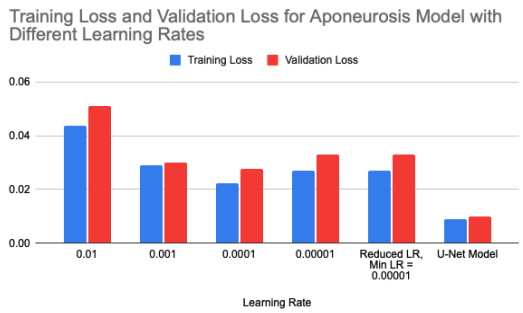
APONEUROSIS MODEL PERFORMANCE WITH CHANGING LEARNING RATE

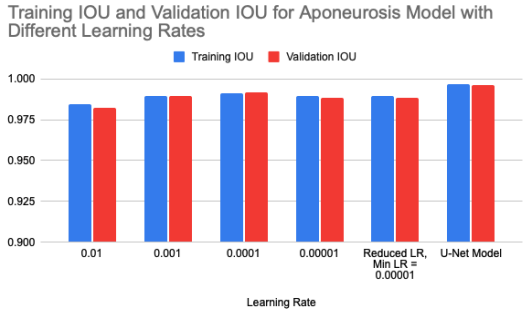
Learning Rate	Loss	Accuracy	IOU	Validation Loss	Validation Accuracy	Validation IOU
0.01	.0436	0.9810	0.9847	.0509	0.9792	0.9824
0.001	.0288	0.9850	0.9894	0.03	0.9850	0.9898
0.0001	.0222	0.9874	0.9914	0.0276	0.9864	0.9917
0.00001	.0562	0.9858	0.9617	0.0639	0.9827	0.9606
Reduce LR on Plateau: Min LR = 0.00001	.0271	0.9854	0.9898	.0331	0.9848	0.9885
U-Net Model, Reduce LR on Plateau	0.0253	0.9861	0.9906	.0375	0.9841	0.9905

FASCICLE MODEL PERFORMANCE WITH CHANGING LEARNING RATE

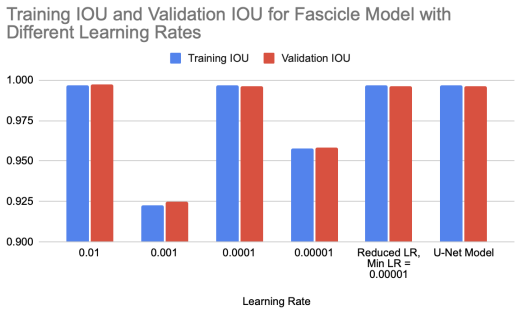
Learning Rate/Model	Loss	Accuracy	IOU	Validation Loss	Validation Accuracy	Validation IOU
0.01	0.0093	0.9922	0.9966	0.0124	0.9901	0.9971
0.001	0.0876	1.000	0.9226	0.0845	1.000	0.9250
0.0001	0.0058	0.9924	0.9969	0.0131	0.9908	0.9961

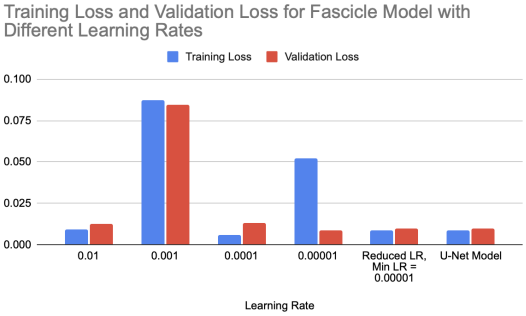
0.00001	0.0520	0.9919	0.9575	0.0511	0.9931	0.9582
Reduce LR on Plateau: Min LR = 0.00001	0.0084	0.9919	0.9960	0.0088	0.9932	0.9973
U-Net Model, Reduce LR on Plateau	0.0080	0.9936	0.9968	0.0098	0.9930	0.9966





Above: *Training Loss, Validation Loss for Aponeurosis Model with different learning rates*





Above: *Comparison of training/validation loss and IOU for different learning rates and models*

3. Learning Rate Experiments: Using Reduce LR On Plateau function, which reduces the learning rate when the model stops improving. Tried various minimum learning rates.

TABLE OF RESULTS:

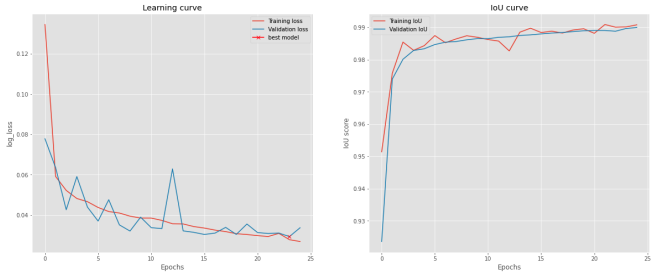
Aponeurosis Mask Model Results with Dynamic Learning Rate:

Min Learning Rate	Loss	Accuracy	Validation Loss	Validation Accuracy	Validation IOU
0.1	.0269	0.9855	0.0337	0.9847	0.9908
0.01	.0292	0.9851	0.1167	0.9581	0.9814
0.00001	.0261	0.9860	.0478	0.9817	0.9890

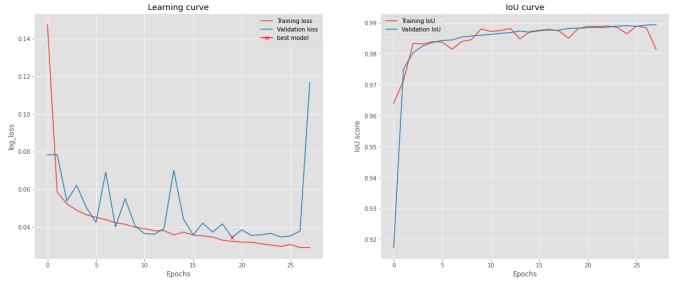
Fascicle Mask Model Results with Dynamic Learning Rate: :

Min Learning Rate	Loss	Accuracy	Validation Loss	Validation Accuracy	Validation IOU
0.01	.0090	.9920	0.0089	0.9921	0.9966
0.00001	0.0084	0.9919	0.0088	0.9932	0.9973

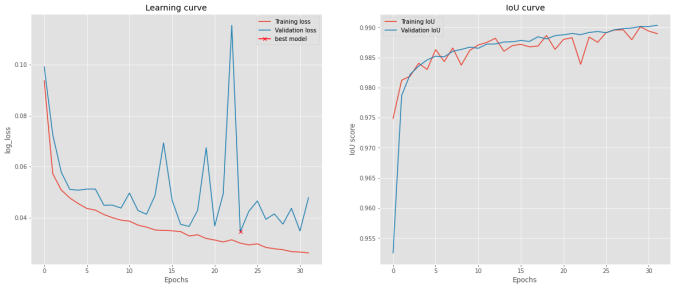
Aponeurosis Model Results:



Above: Learning Rate = 0.1



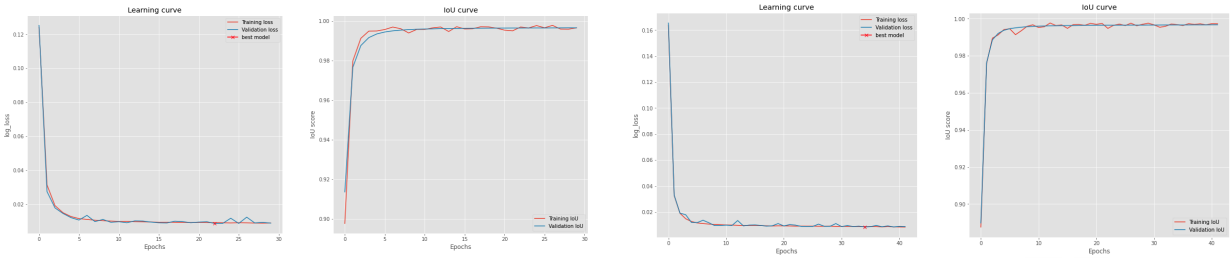
Above: Learning Rate = 0.01



Above: Learning Rate: 0.00001

Fascicle Model Results, Reduce Learning Rate Upon Plateau

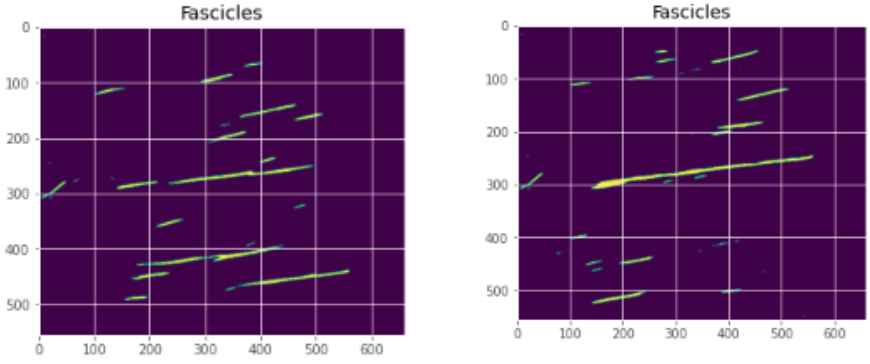
Learning Rate: 0.01



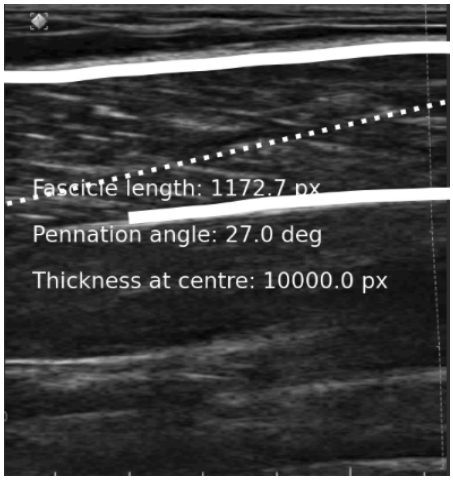
Min Learning Rate = 0.01

Min Learning Rate = 0.00001

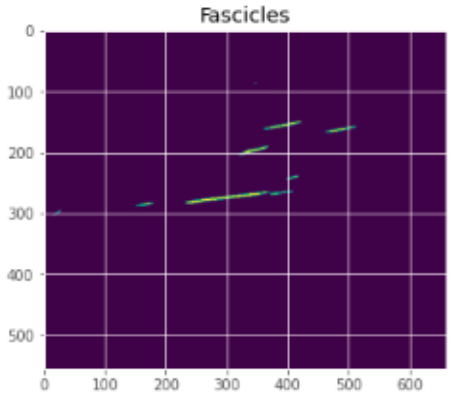
4. ENLARGED FIGURES

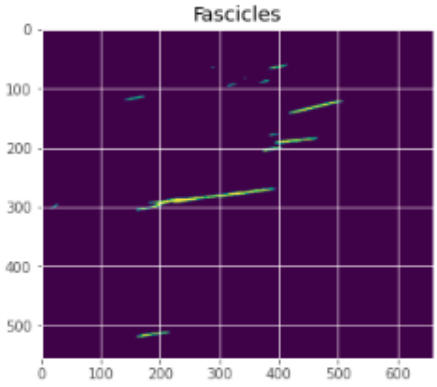


Above: Original U-Net: Masks of Image 349 and 350 respectively. Inconsistent number of fascicles detected.

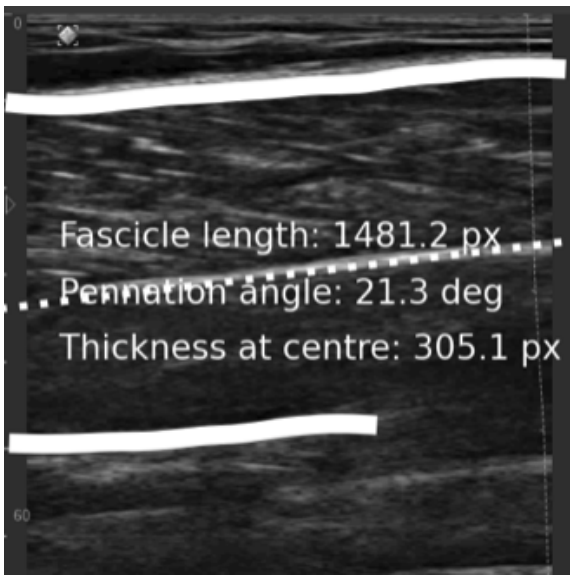


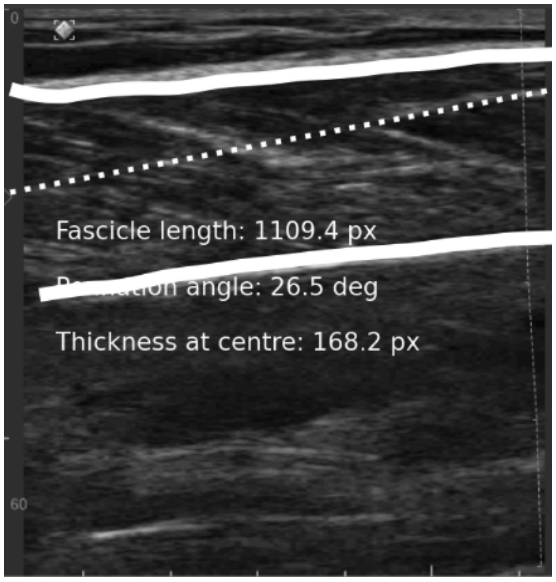
Above: U-Net model from Cronin et al. Measurements are not consistent or not extractable from one frame to the next. .





Above: U-Net + VGG16: Image 349 and 350. Same fascicles are detected in successive frames.





Above: With U-Net + VGG16, metrics are found in both images. They are relatively consistent, except for thickness.