A Domain Enriched Deep Learning Approach to Classify Atherosclerosis using Intravascular Ultrasound Imaging
Supplemental Material

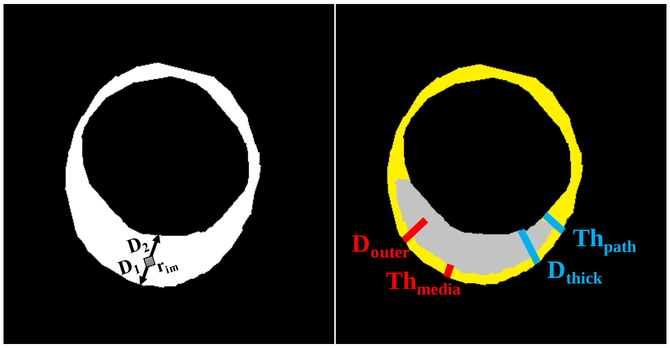
I. Materials and Methods

A. *Pathological Tissue Detection*

To determine the normal wall and the media layer locations and dimensions, two geometrical parameters, D_{thick} and D_{outer}, were computed for each ROI pixel as functions of D_1 and D_2 (Eqs. 1 and 2). D_1 is the Euclidian distances of the pixel $(i_{r_{im}}, j_{r_{im}})$ from the media-adventitia border b_{ma} and D_2 is the Euclidian distances of the pixel $(i_{r_{im}}, j_{r_{im}})$ from the lumen border b_l (Fig. 3 and Fig. S1). Threshold values for D_{thick} and D_{outer} were calculated to determine whether a pixel was in a section of sufficient thickness to be considered pathological or sufficiently close to the media-adventitia border to lie within the media.

B. *Classification*

To classify the pixels corresponding to pathological tissue, a sequence of convolutions, activations, and pooling operations were executed. The network found to perform best, and utilized in this work, is shown in Fig. S2 and Fig. 4.



(a) (b)

Fig. S1. Schematic presentation of the pathological tissue detection. (a) The Euclidean distances from a pixel ($r_{im} \in$ ROI) to the media-adventitia border (D_1) and the lumen border (D_2) were calculated. (b) Pixels within the ROI for which $D_{outer} \geq Th_{media}$ and $D_{thick} \geq Th_{path}$ correspond to pathological tissue (gray); other pixels within the ROI correspond to non-pathological tissue or media (yellow).

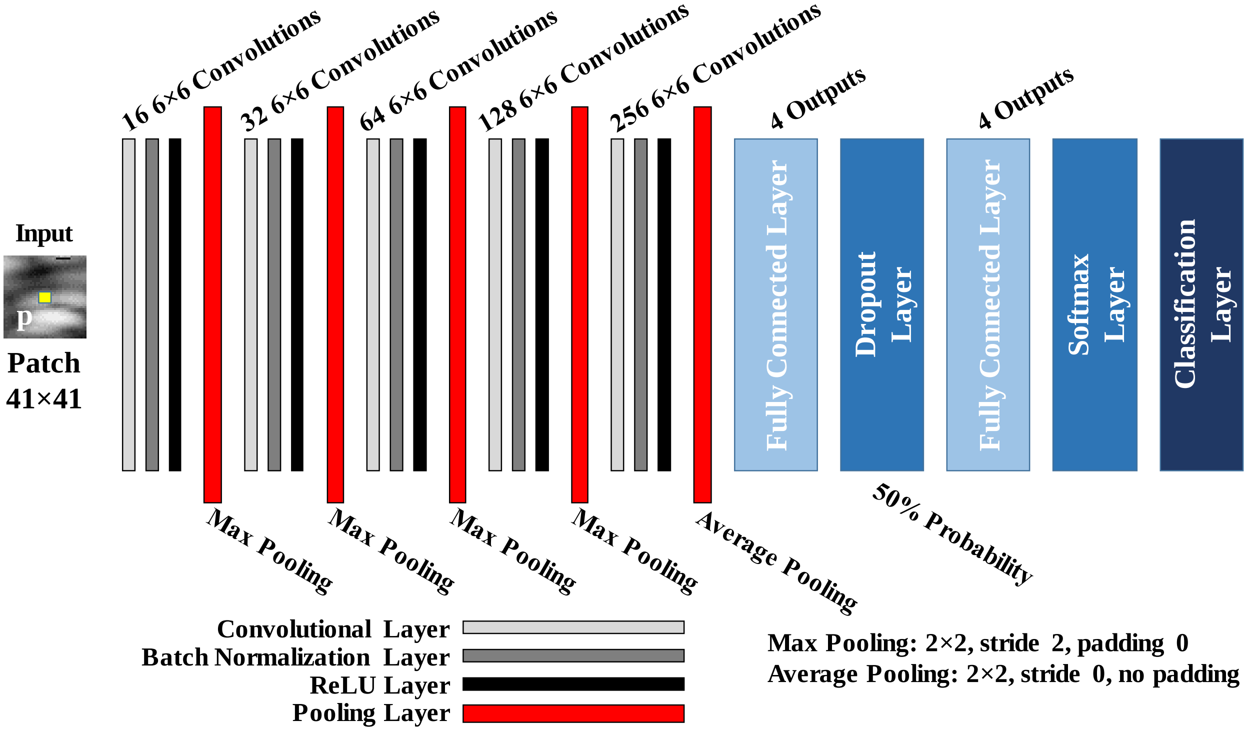


Fig. S2. Detailed architecture of the 26-layer CNN used to classify pixels within the pathological region of interest. For the "naïve" method, all architecture was the same with the exception of the final fully-connected layers, which had 5 nodes/outputs instead of 4. See Fig. 4 for representations of the intermediary data structures and activations of the hidden layers.

II. Dataset

As intravascular images were collected from a clinical population in the course of treatment, the data were inherently inhomogeneous and imbalanced. Fig. S3 shows the average quantity and relative distribution of tissue types present in the imaged vessel walls of each patient, as well as the number of VH-IVUS frames available from each patient's acquisition. Though sometimes present in small numbers, all five tissue types were detected by VH-IVUS in each patient (though notably not in each frame). In total, 553 frames were available in the dataset, from which 200 were randomly selected and set aside for exclusive use in testing.

From the 353 frames used for training and validation, 3.4×10^5 41-by-41 pixel patches of each class were extracted and augmented 6-fold (through reflection and rotation in 90° increments) for use in training the full networks. For the pixel sensitivity study (Supplemental Materials, Section III.B), a smaller sub-set of 10^4 patches of each class was selected to accelerate training; 10-fold cross-validation was performed.

From the 200 frames withheld for testing, 5×10^4 patches of each class were selected and used for evaluation of both methods. In sampling the labeled testing data, regions immediately adjacent to boundaries between tissue types and edges of tissue regions were avoided. Doing so limited uncertainties arising from VH resolution and image data degradation due to file compression; enhanced label certainty for inclusion in testing was also achieved by requiring agreement between two automated color-based methods used for determining categorical labels from colored VH images (in which color-coded labels have been integrated with the underlying grayscale image). Furthermore, avoiding transition zones and borders ensured higher certainty by targeting to a greater extent the "bulk" of a tissue region – VH-IVUS validation has been conducted by histology and atherectomy [1]–[5], which allows comparison of bulk tissue type identification rather than pixel-by-pixel comparison, thereby making such an approach to sampling data for testing against VH-IVUS prudent. A sample frame from the testing subset, alongside the corresponding "bulk" regions and actual randomly-selected pixels used in quantifying test performance, and is shown in Fig. S4. For the ablation study, a smaller sub-set of 10^4 patches of each class was randomly selected to accelerate classification and analysis.

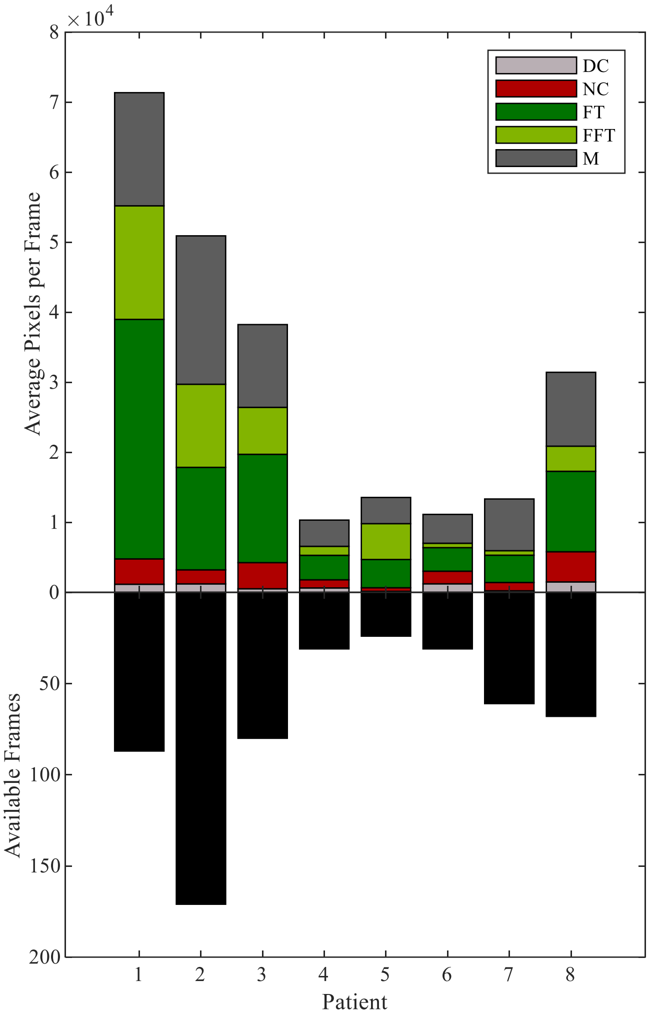


Fig. S3. Plaque distribution of patients comprising dataset. *Top*: Prominence of plaque and composition for each patient, reported in average number of pixels per frame corresponding to each characterized tissue type as determined by VH-IVUS. *Bottom*: Number of VH-IVUS frames available for each patient.

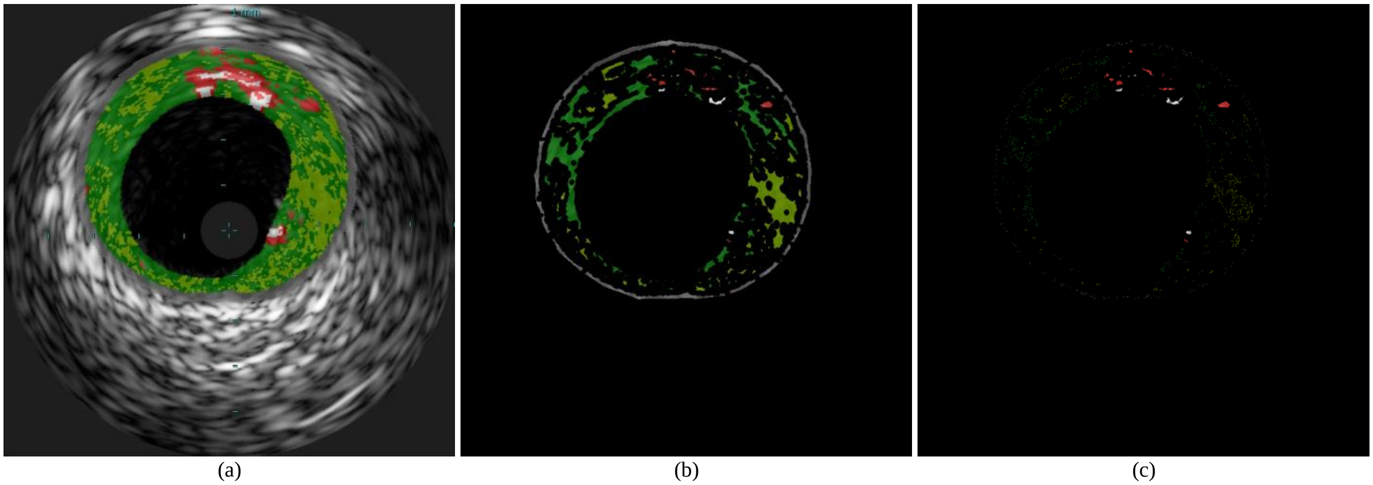


Fig. S4. Selection of testing data from a typical frame withheld from training. (a) Sample VH-IVUS frame from the testing subset. (b) Masked region showing regions of bulk tissue (avoiding boundaries between tissue types and edges of tissue regions). (c) Specific pixels included in the balanced testing dataset (selected from the bulk tissue region shown in (b)). Note how, due to the relative scarcity of dense calcium and necrotic core, a greater portion of each calcified and necrotic region is sampled in constructing the balanced set used to quantify the test performance of the networks.

III. Results

Performance of both the naïve and enriched deep learning methods described here is compared with that reported for existing methods in Table III, which also includes validation metrics reported for VH-IVUS. Included methods include those addressed and described briefly in the introduction: Zhang et al. [6], Brunenberg et al. [7], Athanasiou et al. [8], Taki et al. (2010) [9], Taki et al. (2013) [10], Athanasiou et al. [11], Hwang et al. [12], and Kim et al. [13]. Performance of VH-IVUS is reported for results from an *ex vivo* validation study carried out by Nair et al. [5].

A. Enriched Method CNN Performance

The overall performance metrics for the domain enriched method (Table I) depend both on (four-class) classifier performance and reliability of pathological tissue detection, which together share responsibility for the full segmentation procedure. The CNN classifier achieved generally high precision and recall. Tables SI and SII show the error matrix and mean predicted class scores for the enriched method's four-class CNN classifier when presented with the testing set. The model trained on just pathological tissue achieved an accuracy of 92.3% (cf. naïve five-class classifier accuracy of 90.8% for those four tissue classes). Progression of training for this network is shown in Fig. S5.

As noted above, results of the CNN for the 5-class model are the same as those for its corresponding complete segmentation method (Table II).

B. Patch Size Sensitivity Test

The CNN takes as input not a full image, but rather a patch, or neighborhood of pixels surrounding the pixel of interest to be classified. The size of the patch can impact the subsequent performance of the trained network – if a patch is too small, it may not capture sufficient context or information for a robust classification to be made, but if a patch is too large, additional parameters (θ) must be learned, increasing solution space complexity, training time, and data and network storage size, among other potential pitfalls. Therefore, a preliminary study of patch size was performed to assess performance achieved by networks utilizing square neighborhoods of various sizes ranging from 33 to 71 pixels in dimension (1,089 to 5,041 total pixels).

Ten-fold cross-validation was performed for each patch size using a relatively small dataset (4×10^4 patches equally distributed among the four tissue types). That is, the dataset was randomly split into ten equally-sized balanced subsets, and the training procedure was repeated ten times; each time, a different one of those subsets was withheld from training and used to evaluate the final performance of the network trained on the other nine. The same network architecture was used for all 12 patch sizes with the exception of the input layer, which was modified to accept the desired patch size; the number of inputs to the first fully-connected layer differed accordingly. Training time was recorded for a single training run executed on the same hardware in consistent conditions for all 12 patch sizes. (Due to the amount of training required, tasks were otherwise distributed across different hardware and conditions, not allowing for direct comparison. A range of times is therefore not reported.)

Results of the sensitivity study indicated that the overall domain enriched deep learning approach and method to classify atherosclerosis using IVUS isn't largely dependent on the patch size used, as shown in Fig. S6. However, a patch size of 41-by-41 was shown to offer desirable performance. In particular, this patch size had among the highest average test accuracies, lowest standard deviation in accuracy, and lowest training times of those patch sizes assessed. Though not the best performing by any of those metrics alone, taken together this patch size demonstrated the most desirable holistic performance profile. However, the study provides confidence that the quality of the data and fidelity with regards to the underlying population from which the set is drawn is more important than the specific details of patch size and similar network design features.

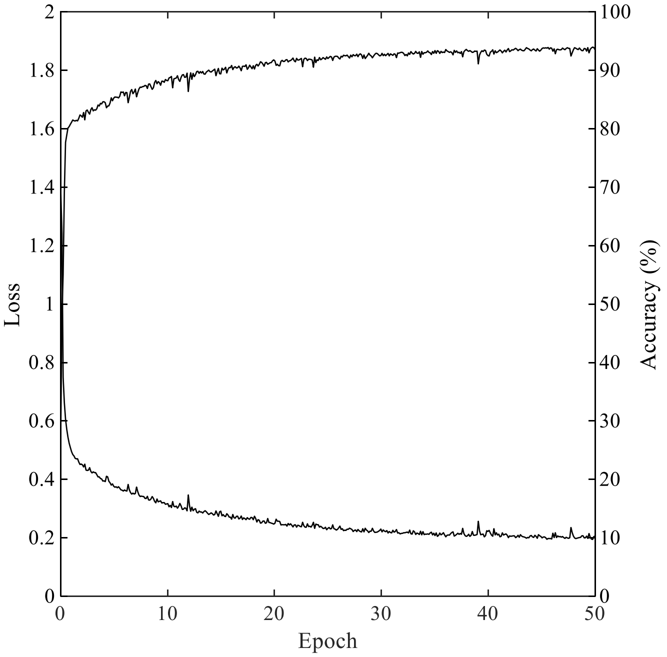


Fig. S5. CNN validation accuracy (*top*) and loss (*bottom*) through 50 epochs of training for the enriched model. Accuracy and loss plateaued by around the 50th epoch; deviation between training and validation accuracies – indicating overfitting – became apparent soon afterwards when training was allowed to proceed beyond 50 epochs.

TABLE SI
Classifier Error Matrix: 4 Classes Trained on Pathological Tissue

		Target Class					
		DC	NC	FT	FFT	***Precision***	M
Output Class	DC	49336	4664	0	0	91.4%	858
	NC	664	44340	1208	0	95.9%	12320
	FT	0	996	45657	4771	88.8%	24893
	FFT	0	0	3135	45229	93.5%	11929
	Recall	98.7%	88.7%	91.3%	90.5%	**92.3%**	-

TABLE SII
Classifier Mean Predicted Class Score

		Target Class				
		DC	NC	FT	FFT	***Mean***
Output Class	DC	**0.9855**	0.1779	0.0023	0.0000	0.2914
	NC	0.0145	**0.7969**	0.0271	0.0000	0.2096
	FT	0.0000	0.0204	**0.7158**	0.1012	0.2093
	FFT	0.0000	0.0048	0.2549	**0.8988**	0.2896
	Mean	0.2500	0.2500	0.2500	0.2500	0.2500

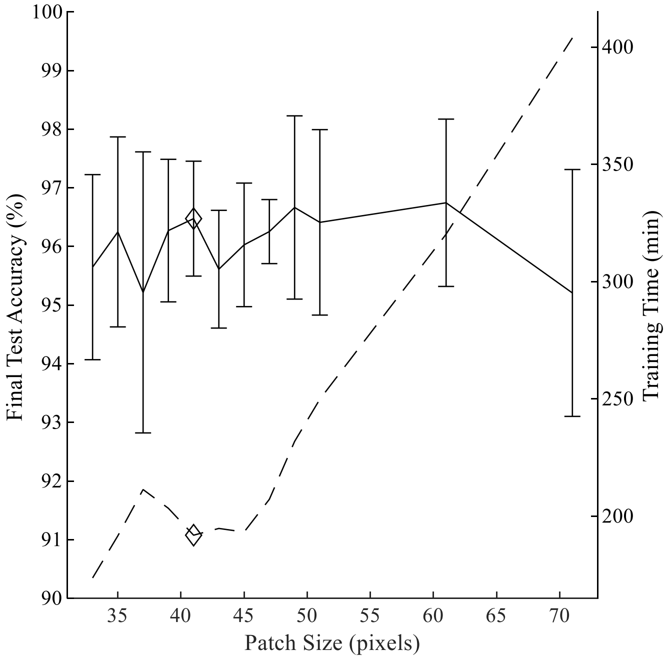


Fig. S6. Results of the sensitivity study demonstrate the effect of patch size on a small dataset. The averages and standard deviations of final test accuracies generated by ten-fold cross-validation are plotted as a solid line (left vertical axis). Time to execute the training process once is plotted as a dashed line (right vertical axis) trending upwards with patch size. Patch size as listed corresponds to a single dimension (i.e. the square root of the total number of pixels contained in a given patch of that size). The diamonds call out the performance of the networks utilizing 41-by-41 patches like those used in the remainder of the work presented herein. It can be seen that this patch size was among the best performing with regards to average final testing accuracy, standard deviation in final testing accuracy, and training time length.

C. *Ablation Study*

An ablation study was carried on the 4-class CNN classifier of the domain enriched method to interrogate the organization and relationships within the network structure. In turn, each of the 504 nodes or channels was “ablated” by negating a filter’s weights and associated bias term (in a convolutional layer) or all input connection weights and bias term (in a fully-connected layer), thereby depriving the node or channel of input and inhibiting its function by mandating an output of zero. By systematically “ablating” small segments of the network in this way, we probed where and how “knowledge” was represented in our network, how robust and redundant or centralized our network was, and how important certain parts of the network were to overall importance.

Two major analyses were performed to investigate characteristics of the network: the impact of individual channels and nodes on final class sensitivity (i.e. recall) and the relationship between changes in class sensitivity caused by the various ablations. The results of the first analysis are shown in Fig. S7, which illustrates the absolute drop in class sensitivity resulting from ablation of channels and nodes, arranged by layer. On the left side of this figure, each horizontal line represents a channel or node of the corresponding layer which was ablated, with colors indicating the extent and direction of change in sensitivity for each class resulting from its ablation. Changes are also compiled for entire layers to examine importance of network layer in the outcome for each class, as shown on the right side of the figure.

One can observe that the ablation of most functional units has a very small effects on end output, suggesting the network is fairly robust in that classification is generally distributed. A few notable exceptions exist, however. First, a few units play an outsized role in the successful classification of FFT pixels, particularly in the first convolutional layer; when these units are ablated, sensitivity for FFT drops by up to 61%p. Clearly the output of these early channels are important to the downstream processing performed in later layers to positively identify pixels showing FFT tissue. The other units with a major role are the nodes of the second fully-connected layer, the output of which enters the softmax and subsequently classification layers. As the layer whose outputs are directly compared to establish class assignment, it is expected that each node would have profound importance to the classification of a single corresponding tissue type, as observed. An interesting observation is the impact on the three other classes; we see that ablation of the node corresponding to DC results in an improvement in NC sensitivity of just over 10%p. Ablation of the nodes corresponding to FT and FFT similarly result in more modest improvements in sensitivity for the other class. One interpretation is that these units not only contribute to classification of one class, but effectively suppress the others. While most clear in the final fully-connected layer, examples of competing performance driving confusion between DC and NC and FT and FFT are visible throughout the network. For both pairs, an ablation causing a decrease in sensitivity for one class in the pair was often accompanied by an increase in the sensitivity of its counterpart. These relationships and trends emerge, and are explored more explicitly, in Fig. S8.

The results of the second analysis, shown in Fig. S8, illustrate relationships between the representation of different tissue classes. Positive correlations may suggest that the tissue classes share many features encoded by the network such that ablation of the units strongly activated by these features similarly impacts both classes. However, negative correlations may also suggest shared features; if two classes share many features, ablation of units strongly activated by distinguishing features could cause pixels of the feature-sharing class to be selected with greater frequency, thereby driving changes in sensitivity of the two classes in opposite directions. In such a case, the unit activated by the distinguishing feature could be considered to be a suppressor or inhibitor of the related class which is inactivated by the ablation. Given this understanding, the results indicate that network representations of DC and NC are closely related, as are representations and classification pathways of FT and FFT. Both pairs demonstrate negative correlations. However, the complexity of these relationships are neither fully characterized nor done justice by so simple a summary. The results of the ablation study illustrate the inextricable, distributed, and interdependent nature of the interrogated neural network.

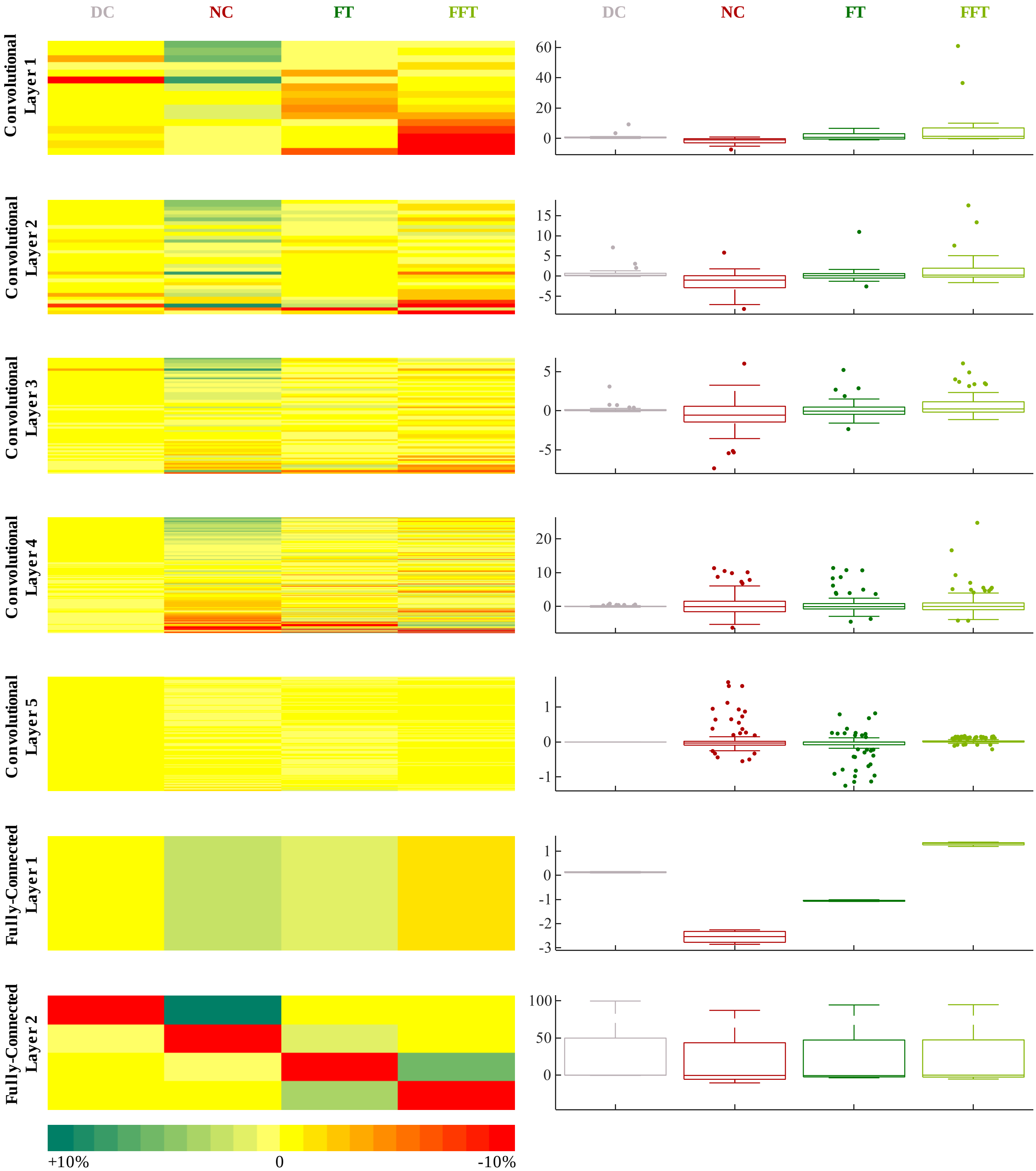


Fig. S7. Results of the neural network ablation study illustrate distributed responsibility for classification. The absolute drop in class sensitivity resulting from ablation of channels and nodes shows that most individual ablations do not drastically impact overall performance. *Left:* Each horizontal line represents a channel or node of the corresponding layer which was ablated, with colors indicating the change in sensitivity for each class resulting from its ablation. (Note that, to show detail, the color scale does not span the entire observed range of outcomes.) An inverse trend appears between the direction of change for some classes (see Fig. S8 for more detailed assessment). *Right:* Changes are also compiled for entire layers, showing importance of network layer in the outcome for each class. Boxplots illustrate drop (in percentage points) to class sensitivity. (Larger numbers indicate larger drops in sensitivity, and therefore a greater role of the functional unit in the successful identification of the associated class.)

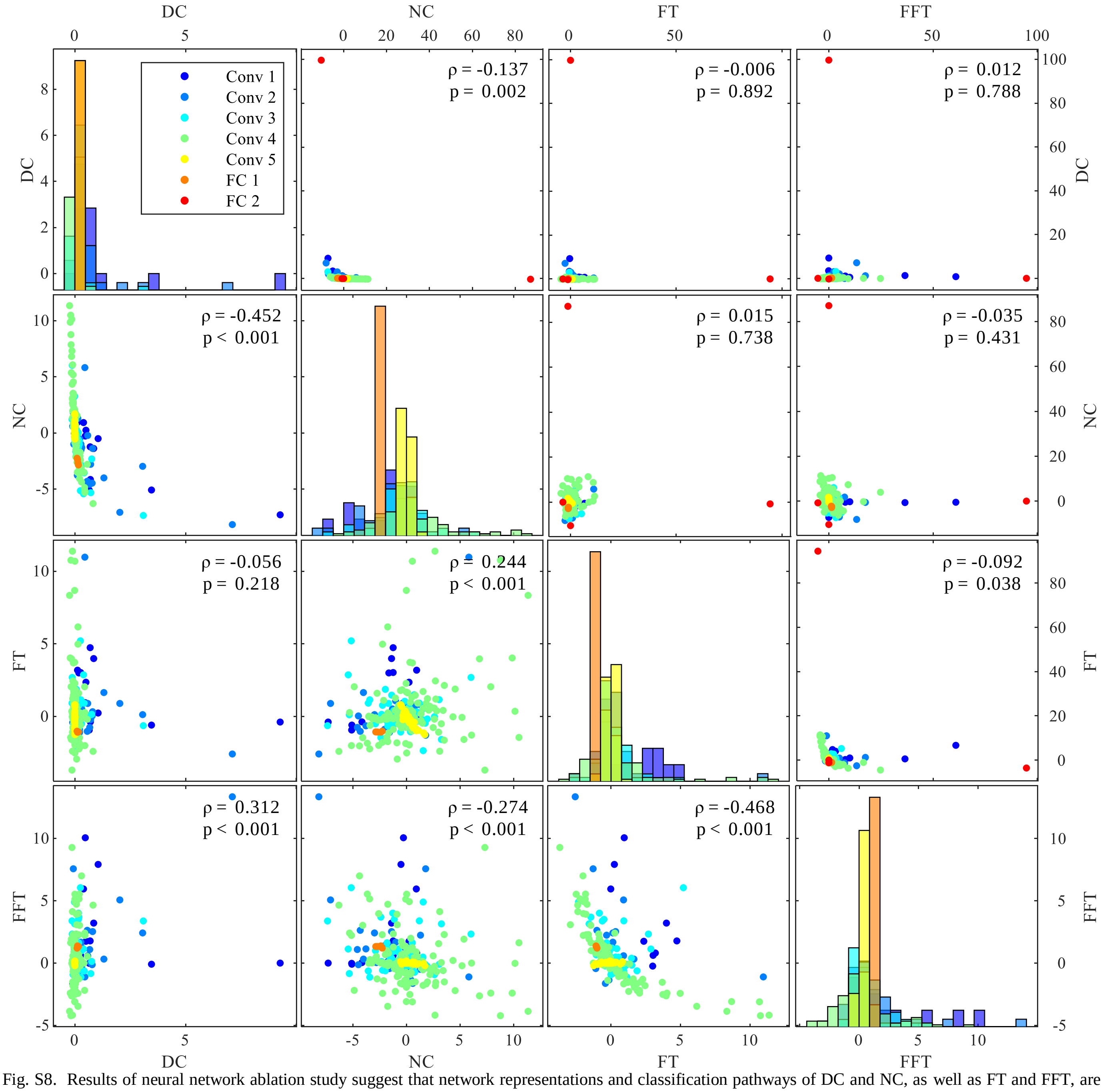


Fig. S8. Results of neural network ablation study suggest that network representations and classification pathways of DC and NC, as well as FT and FFT, are closely related. Each point represents a single unit ablation, with position indicating drop in class sensitivity (in percentage points); as such, each point corresponds to a single horizontal line in Fig. S7. Color indicates the layer to which the ablated unit belonged. Plots mirrored across the diagonal show the same information in transpose but with different axis ranges; plots on the upper right include 9 outlier points (maximum change resulting from ablation greater than 3 standard deviations from the mean) that are excluded from those on the bottom left (as well as the histograms). Note that the scales differ for the plots above and below the diagonal. Pearson's linear correlation coefficient (ρ) and p-value (p) are displayed for each set of observed changes. (Values below the diagonal are calculated excluding the 9 outliers while those above the diagonal include all 504 points.) The p-values less than the displayed precision are as follows: $p_{DC\text{-}NC} = 2.67\times10^{-26}$; $p_{DC\text{-}FFT} = 1.32\times10^{-12}$; $p_{NC\text{-}FT} = 3.85\times10^{-8}$; $p_{NC\text{-}FFT} = 5.43\times10^{-10}$; $p_{FT\text{-}FFT} = 2.68\times10^{-28}$. The strongest correlations are negative ones between DC and NC and FT and FFT, supporting observations in Fig. S7 and described in the text. Conv: Convolutional Layer; FC: Fully-Connected Layer.

References

[1] D. G. Vince *et al.*, “Automated coronary plaque characterization with intravascular ultrasound backscatter: In vivo and ex vivo validation,” *J. Acoust. Soc. Am.*, vol. 119, no. 5, pp. 3256–3256, May 2006.

[2] A. Konig and V. Klauss, “Virtual histology,” *Heart*, vol. 93, no. 8, pp. 977–982, Aug. 2007.

[3] K. Nasu *et al.*, “Accuracy of In Vivo Coronary Plaque Morphology Assessment,” *J. Am. Coll. Cardiol.*, vol. 47, no. 12, pp. 2405–2412, Jun. 2006.

[4] J. Van Herck, G. De Meyer, G. Ennekens, P. Van Herck, A. Herman, and C. Vrints, “Validation of in vivo plaque characterisation by virtual histology in a rabbit model of atherosclerosis,” *EuroIntervention*, vol. 5, no. 1, pp. 149–156, May 2009.

[5] A. Nair, M. P. Margolis, B. D. Kuban, and D. G. Vince, “Automated coronary plaque characterisation with intravascular ultrasound backscatter: ex vivo validation,” *EuroIntervention*, vol. 3, no. 1, pp. 113–20, May 2007.

[6] X. Zhang, C. R. McKay, and M. Sonka, “Tissue characterization in intravascular ultrasound images,” *IEEE Trans. Med. Imaging*, vol. 17, no. 6, pp. 889–899, Dec. 1998.

[7] E. Brunenberg, O. Pujol, B. ter Haar Romeny, and P. Radeva, “Automatic IVUS Segmentation of Atherosclerotic Plaque with Stop & Go Snake,” in *Med. Image Comput. Comput. Assist. Interv.*, 2006, pp. 9–16.

[8] L. S. Athanasiou *et al.*, “A hybrid plaque characterization method using intravascular ultrasound images,” *Technol. Heal. Care*, vol. 21, no. 3, pp. 199–216, Jun. 2013.

[9] A. Taki *et al.*, “A New Approach for Improving Coronary Plaque Component Analysis Based on Intravascular Ultrasound Images,” *Ultrasound Med. Biol.*, vol. 36, no. 8, pp. 1245–1258, Aug. 2010.

[10] A. Taki, A. Roodaki, S. K. Setarehdan, S. Avansari, G. Unal, and N. Navab, “An IVUS image-based approach for improvement of coronary plaque characterization,” *Comput. Biol. Med.*, vol. 43, no. 4, pp. 268–280, May 2013.

[11] L. S. Athanasiou *et al.*, “A novel semiautomated atherosclerotic plaque characterization method using grayscale intravascular ultrasound images: Comparison with virtual histology,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 3, pp. 391–400, May 2012.

[12] Y. N. Hwang, J. H. Lee, G. Y. Kim, E. S. Shin, and S. M. Kim, “Characterization of coronary plaque regions in intravascular ultrasound images using a hybrid ensemble classifier,” *Comput. Methods Programs Biomed.*, vol. 153, pp. 83–92, Jan. 2018.

[13] G. Y. Kim, J. H. Lee, Y. N. Hwang, and S. M. Kim, “A novel intensity-based multi-level classification approach for coronary plaque characterization in intravascular ultrasound images,” *Biomed. Eng. Online*, vol. 17, no. S2, p. 151, Nov. 2018.