Chapter 4

Automated Deep Learning-Based Paradigm for High-Risk Plaque Detection in B-mode Common Carotid Ultrasound Scans: An Asymptomatic Japanese Cohort Study

Summary

Backgrounds: The death due to stroke is caused by embolism of the arteries which is due to the rupture of the atherosclerotic lesions in carotid arteries. The lesion formation is over time, and thus, early screening is recommended for asymptomatic and moderate-risk patients. The previous techniques adopted conventional methods or semi-automated and, more recently, machine learning solutions. A handful of studies have emerged based on solo deep learning (SDL) models such as UNet architecture.

Method: The proposed research is the first to adopt *hybrid deep learning* (HDL) artificial intelligence models such as SegNet-UNet. This model is benchmarked against UNet and advanced conventional models using scale-space such as AtheroEdge 2.0 (AtheroPoint, CA, USA). All our resultant statistics of the three systems were in the order of UNet, SegNet-UNet, and AtheroEdge 2.0.

Results: Using the database of 379 ultrasound scans from a Japanese cohort of 190 patients having moderate risk and implementing the cross-validation deep learning framework, our system performance using area-under-the-curve (AUC) for UNet, SegNet-UNet, and AtheroEdge 2.0 were **0.93**, **0.94,** and **0.95** (p<0.001), respectively. The *coefficient of correlation* between the three systems and ground truth (GT) were: **0.82**, **0.89**, and **0.85** (p<0.001 for all three), respectively. The mean absolute area error for the three systems against manual GT was **4.07±4.70 mm^2**, **3.11±3.92 mm^2**, **3.72±4.76 mm^2**, respectively, proving the superior performance SegNet-UNet against UNet and AtheroEdge 2.0, respectively. Statistical tests were also conducted for their reliability and stability.

Conclusions: The proposed study demonstrates a fast, accurate, and reliable solution for early detection and quantification of plaque lesions in common carotid artery ultrasound scans. The system runs on a test US image in < 1 second, proving overall performance to be clinically reliable.

4.1 Introduction

Cardiovascular Disease (CVD) is the leading cause of death, with a mortality rate of 18 million each year [3]. Stroke, which is classified under CVD, is the third leading cause of death globally [184]. Atherosclerosis is the leading cause of plaque build-up in the carotid arteries, which supply blood to the brain [8], and thus routine screening is vital in asymptomatic patients [185].

There are several imaging modalities for screening the plaque in carotid arteries, such as magnetic resonance imaging (MRI) [186], computed tomography (CT) [187], and Ultrasound (US) [128]. The US offers several advantages over others, such as high-resolution imaging, user-friendly gain control system, low-cost [188], and wide availability [128]. Further, it is well adapted in diagnostic centers around the globe.

Carotid imaging using the US is popular in diagnosing CVD risk as a surrogate marker for coronary artery disease [189, 190]. One of the popular methods for CVD risk estimation that is well established is based on the measurement of carotid intima-media thickness (cIMT) in longitudinal scans [40, 41]. Scale-space paradigms have been previously demonstrated for the (a) segmentation of cIMT wall region or (b) lumen diameter (LD) in common carotid arteries (CCA) based on the US. However, it poses a challenge since they cannot extract the features automatically [38, 41, 191]. Plaque area was proposed as an alternative measurement of CVD risk. The radiologist (or sonographer) could delineate the plaque manually and quantify its area [35, 168, 170]. These semi-automated and automated delineation and area measurements were not based on the cohort's knowledge; however, they have been valuable for different CVD risk predictions [44, 192].

Artificial Intelligence (AI) recently started to penetrate the healthcare industry and more in the area of healthcare imaging such as MRI, CT, and US [48, 149, 150, 193-196], and even more aggressively during the COVID-19 pandemic [197-199]. Machine learning (ML) which is an integral part of AI that has expanded for imaging-based tissue characterization for disease segmentation and classification such as in breast [130], liver [131, 132], thyroid [136, 200], cancer [201], diabetes [182], coronary [137], and carotid [133, 134]. With the advancement in AI, the segmentation of plaque in the carotid artery has become less complicated and faster [168, 170]. In a recent study, an attempt has been made for US-based plaque segmentation and characterization [143].

These groups have used *solo* deep learning (SDL) models (such as UNet) for morphology-based tissue characterization and thus have demonstrated the superiority of SDL over conventional models. When cascaded by another DL, it would lead to hybrid deep learning (HDL) and has been recently shown to have more superior performance and stable design [97, 202, 203]. We, therefore, hypothesize that HDL (such as SegNet-UNet) can be more superior for plaque segmentation.

This study introduces the HDL-based model for plaque segmentation in longitudinal scans of CCA in a moderate-risk cohort. Thus, the objective of the proposed study is to automatically detect the high-risk plaques in the cohort based on automatic segmentation and its quantification of plaque area (PA). Since our data was low resolution and 10% of the cohort had sharp curvatures, higher sample size was required to achieve superior performance. For this reason, an augmentation protocol was used, which was benchmarked against the SDL model. Further, we illustrate the advantage of capturing the plaque

shape, which has protrusions in the lumen region, demonstrating positive remodeling scenarios [74, 204]. Figure 4.1 shows the global online system.

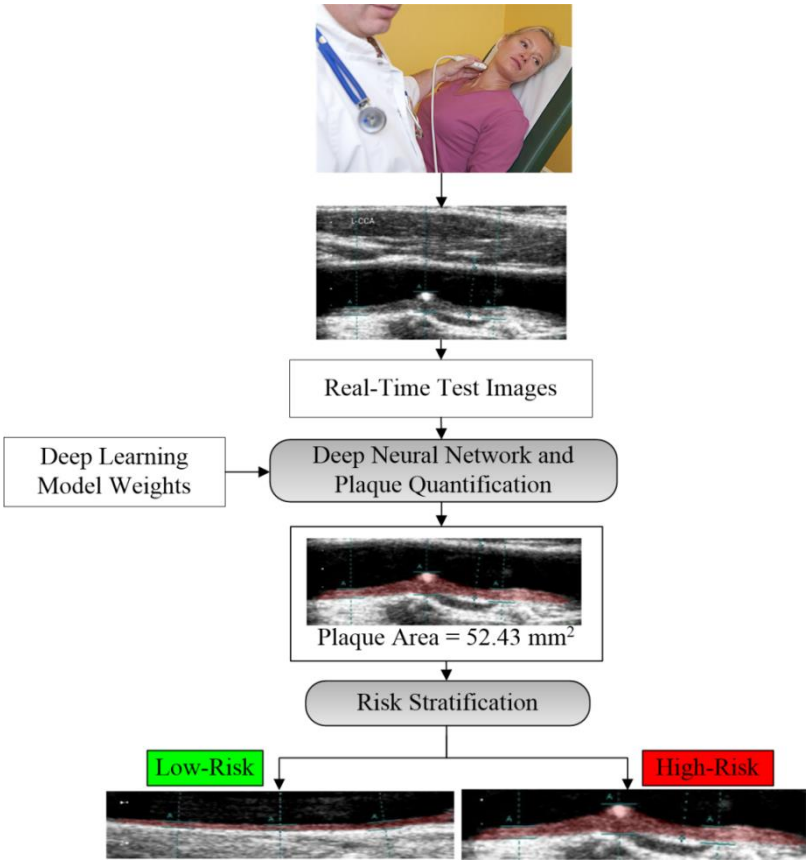

Figure 4.1 Online deep learning systems (AtheroEdge 3.0) for high-risk plaque detection, quantification, and risk stratification.

4.2 Methodology

4.2.1 Patient demographics and baseline characteristics

Our cohort consisted of 379 images of left and right common carotid arteries B-mode ultrasound images taken from 190 patients. All patients were examined retrospectively, and ethics approval was granted by an institutional review board (IRB), Toho University, Japan. All patients were informed, and written consents were obtained before the examinations. The baseline characteristics of the Japanese diabetes cohort included 147 male and 43 female patients with a mean age of 68.78±10.88 years. The mean hemoglobin (HbA1c) of all patients was 6.25±1.06 mg/dL, low density lipoprotein (LDL) cholesterol of 100.84±31.48 (mg/dL), high density lipoprotein (HDL) cholesterol of 50.60±14.73 (mg/dL), and total cholesterol (TC) of 175.27±36.55 (mg/dL). The complete baseline characteristic is shown in Table 4.1, which is divided into high-risk and low-risk categories.

Table 4.1 Baseline characteristics of the Japanese cohort.

Risk Type	**-**		**Low-Risk**		**High-Risk**		**-**
PA threshold	-		GTPA < 40 mm^2		GTPA ≥ 40 mm^2		-
N	Total (379)		331		48		-
-	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	***p*-values**
Gender	293 (M)		256		37		
Age	68.78	10.88	68.64	10.91	69.69	10.71	0.54
HbA1c	6.25	1.06	6.221	1.052	6.442	1.132	0.18
eGFR	45.67	20.43	45.95	19.99	43.77	23.35	0.49
LDL	100.84	31.48	100.7	31.87	101.6	29.01	0.87
HDL	50.60	14.73	51.03	14.98	47.63	12.63	0.14
TC	174.27	36.55	174.2	36.68	174.9	36.01	0.9
SBP	134.62	8.88	134.3	9.061	137.1	7.133	0.04
DBP	87.31	4.44	87.13	4.531	88.54	3.567	0.04
HT	277	-	236	-	41	-	< 0.0001
Smoking	151	-	137	-	14	-	< 0.0001
FH	47	-	40	-	7	-	<0.0001

PA: Plaque area; GTPA: Ground Truth Plaque Area.

4.2.2 Data Acquisition and Ultrasound Imaging

An Ultrasound scanner model Aplio XV, Aplio XG, Xario, Toshiba, Inc., Tokyo, Japan, equipped with a 7.5-MHz linear array transducer, was used to perform ultrasound examinations. An experienced sonographer (with 15 years of experience) performed all US scans. The duration of data collection was between July 2009 to September 2010. The recommendations of the American Society of Echocardiography Carotid Intima-Media Thickness Task Force were followed during the acquisition of these images. In this database, the mean resolution was 0.052±0.01 mm/pixel.

4.2.3 Ground Truth Data Preparation

A researcher trained by an experienced cardiologist (20 years of experience) prepared the ground truth (GT) binary masks. A state-of-the-art tracing tool developed by AtheroPoint™ LLC, Roseville, USA, was used to mark the lumen intima (LI) and media adventitia (MA) borders in CCA images. These tracings are used to generate the GT binary masks for both UNet and SegNet-UNet systems.

4.2.4 Data Augmentation

We used 379 images from the Japanese diabetes cohort and experimented with such data. However, we tested and observed that the collected data is insufficient to produce the intended segmentation results. Therefore we decided to use an augmentation method to enhance the data size. A random affine rotation transform [-15^o to 15^o] was applied to all images, thus generating 2x database.

4.2.5 The Architecture of AI models: SDL and HDL

In the present work, two segmentation models named as UNet and SegNet-UNet are proposed. Both architectures are briefly described in the following section.

4.2.5.1 Phase I - Offline System Weight Computation using UNet

The offline weight computation is the crucial step. It generates a set of numbers called "models" or "offline training model". The offline training model is based on the type of deep learning model. There are several kinds of deep learning models one can use for generating offline weight computations. We call our deep learning model as "UNet" model. The working of the UNet model is discussed below. During the training phase, one needs two kinds of inputs: (a) grayscale CCA image and (b) gold standard (binary image showing the wall region). Since we need to determine the weights of the DL model for the far wall region, we have to give the grayscale image input (a).

Further, we need some references for the training. This reference is (b), called a gold standard. The gold standard is in the form of the binary mask region of the far wall of the carotid artery.

UNet Model - Overview

The 'U' Shaped segmentation model was initially proposed by Ronneberger *et al.* [163]. Figure 4.2 shows the architecture of the UNet model with a stack of four encoders and four decoder stages on both sides of the U shape. We will discuss the encoder and decoders shortly.

The ground truth binary mask and grayscale images are given to the UNet input layer in a fixed "mini-batch size" during the training phase. In layman's, this means the number of images can be processed at one time. Typically, this is decided by the network designer. The mini-batch size is directly related to the computational ability of the hardware. Higher the number of "mini-batch size", higher is the computational requirement of the hardware computer. We took the mini-batch size=10. In layman's, thus means 10 images can be processed at one time using this UNet architecture.

Encoder Block of UNet Architecture

UNet architecture resembles a U-shape. The left half consists of stages for reducing the image size as we go along downwards. The blocks in each stage are called the layers of the deep learning architecture. These blocks are also called "Encoder blocks". The encoder stage (left half of the UNet architecture) *down samples* the images (reducing the image size), while the decoder stage *up samples* the images. Downsampling means reducing the image to half in each stage as it goes down. The purpose of downsampling the image is to keep good and relevant features. This is accomplished using a simple mathematical operation called "convolution and max-pooling".

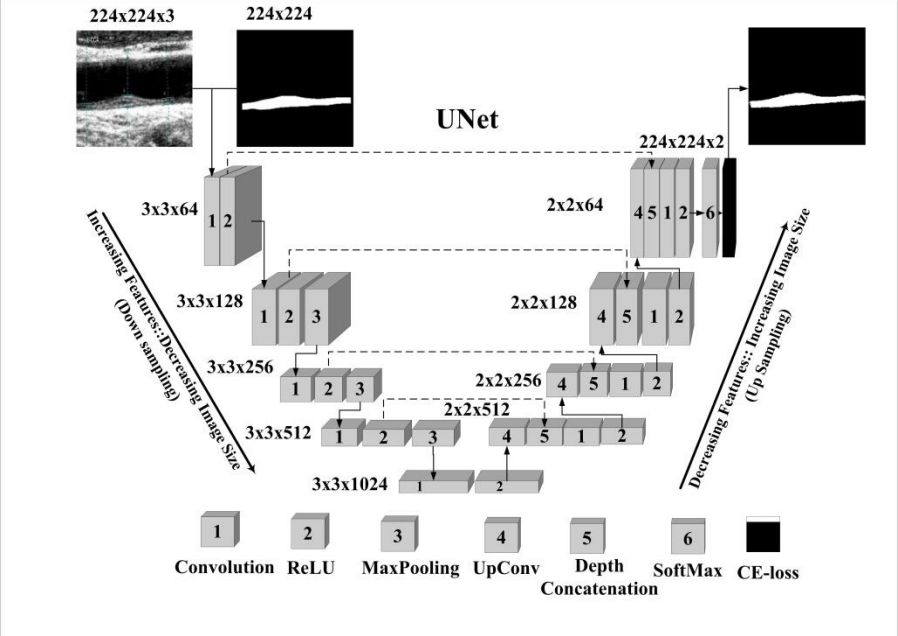

Figure 4.2 Schematic diagram of solo deep learning-based UNet architecture.

Each stage of the encoder is represented as a stack of 2D-convolution layers (block #1), ReLU activation (block #2), and max-pooling layers (block #3). The figure 2 shows that the grayscale US image is given as an input of size 224x224x3 (224 x 224 is the image size, three channels are red, blue, and green, called RGB format) from the encoder stage. Starting from the number of filters 64, the number of filters doubles in each stage of the encoder module. These numbers count to 64, 128, 256, and 512. Note that the encoder helps in extracting the features of the grayscale region. Note that the number of filters is the same as the number of features. The Figure 4.2 shows the numbers 3x3x64, 3x3x128….3x3x1024; where 3x3 is the filter size and 64…1024, are the number of filters.

Decoder Block of UNet Architecture

On the right side of Figure 4.2 are the decoder stages. It is the reverse of the encoder block. It is needed to recover the original dimensions of the training image. Counter to that, the number of filters is reduced to half in each stage of the decoder module. When we start counting from the bottom of Figure 4.2, these numbers are 512, 256, 128, and 64. The decreasing number of filters at the decoder side helps in resizing the image to its original size. After the last decoder stage, the image gains its original size with refined features. Each decoder stage consists of a stack of up-convolution-2D layers (block #4), depth-concatenation (block #5), 2D-convolution (block #1), and ReLU (block #2). The next section will discuss "Bridge Network", which links the encoder and decoder. We will also discuss the dotted line in the "skip connection".

Bridge Network of UNet Architecture

The function of the bridge network is to connect the encoder and decoder stages. Thumb rule of the concatenation is feature map should be of the exact sizes. Since the last stage of the encoder has 3x3x512 features (see Figure 4.3), and we want to concatenate 3x3x512 features of the last encoder (left side) stage with the first decoder stage (right side). The first decoder stage consists of the UpConv layer (block #4 in Figure 4.3), requiring 3x3x1024 features to convert into 3x3x512 features from the previous stage. Therefore, we need a bridge network that generates 3x3x1024 convolutional features. The bridge network consists of a 2D-convolution layer (block #1) and the ReLU layer (block #2).

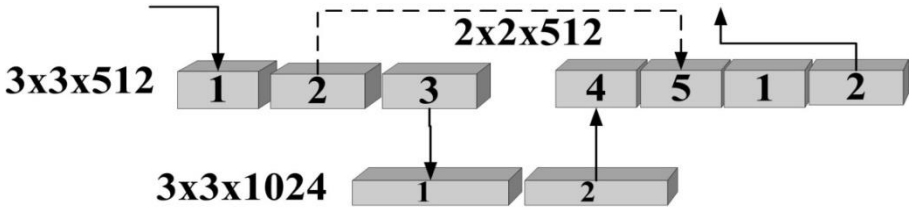

Figure 4.3 Bridge network of the UNet architecture.

Skip connection in UNet Architecture

From each encoder stage, spatial features are extracted and transferred to the *depth concatenation layer* (block #5) *of the decoder stage* via a skip connection. The features come from the corresponding encoder stage (block #2) and the output of the UpConv block of the decoder stage. The features are taken from the ReLU outputs of the encoder stage and the previous decoder stage and added via a depth concatenation block in the second decoder stage. This process is repeated for all the decoder stages.

Classification stage

Finally, after the last decoder stage, the image features are classified into two classes, i.e., plaque area (white pixels) and the background (black pixels) using the softmax classifier layer (block#6). This classified image is then subtracted from the ground truth mask, and the difference in mask pixels results in segmentation loss (or cross-entropy loss).

Loss Function and Generalization during Training

An efficient ADAM optimizer further optimizes the cross-entropy loss. Once all images from the database pass through the network, the network completes one epoch. Generally, in initial epochs network gives a large segmentation loss and low accuracy, which is further improved in subsequent epochs. After each epoch, the weights of the classification model are saved. After a certain number of epochs, the model is ready to test unknown data when the network is generalized. An efficient ADAM optimizer is used to

reduce the cross-entropy loss in plaque segmentation. If y_i represents the GT label and a_i represents the softmax classifier probability, then the cross-entropy (CE) loss is described by Equation 4.1 as follows:

$$L_{CE} = -[(y_i \times \log a_i) + (1-y_i) \times \log(1-a_i)] \quad (4.1)$$

At the end of the training process, the model is generated.

SegNet-UNet Model

Szegedy *et al.* [166] Introduced the concept of the hybrid model by fusing Inception with ResNet AI architectures. In this study, we have presented a hybrid model SegNet-UNet that fuses two SDL models such as SegNet and UNet. Figure 4.4 shows a schematic slab diagram of SegNet-UNet. This hybrid model is a parallel arrangement of SegNet and UNet models, which shares common input, softmax classification, and ADAM optimization layers. SegNet part of the hybrid model also carries an encoder-decoder-based architecture; however, the stack of layers is different from the UNet model. Each encoder stage in the SegNet part consists of a stack of 2D-convolution (block #1), batch normalization (block #7), ReLU (block #2), and max-pooling layers (block #3). Also, each decoder stage in same consists of max-unpooling (block #8), 2D-convolutional, and ReLU.

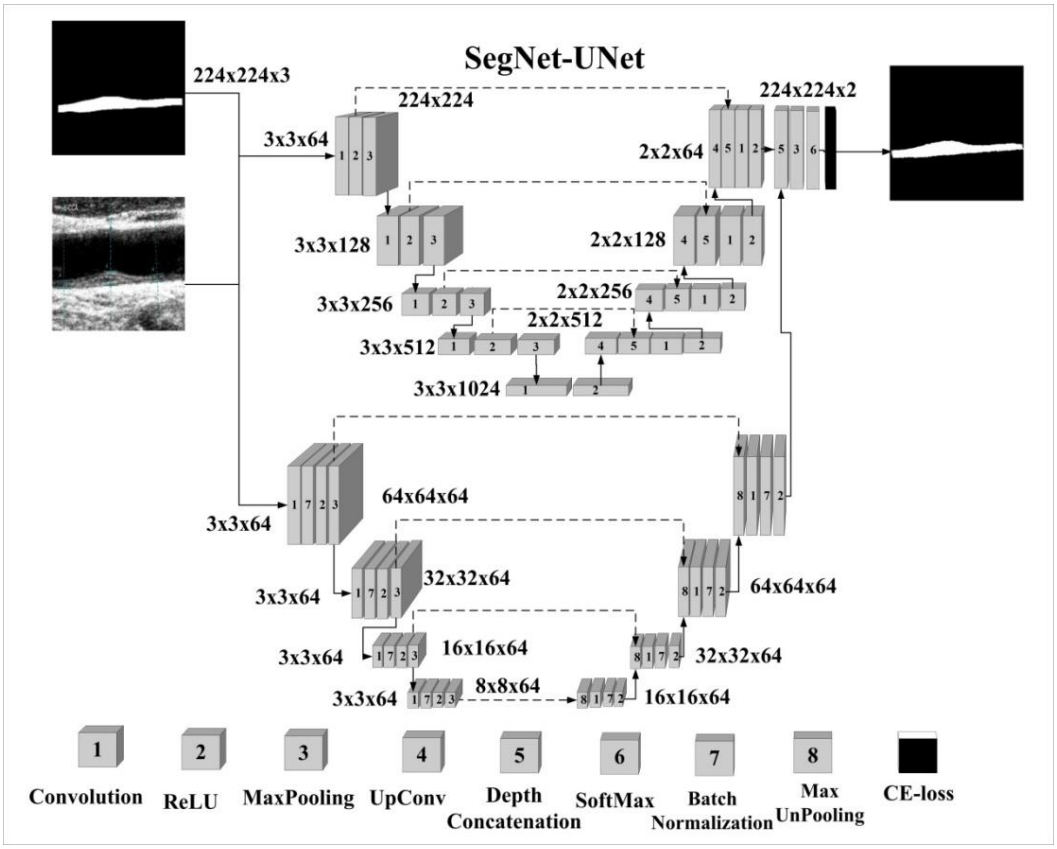

Figure 4.4 Schematic diagram of hybrid deep learning-based SegNet-UNet architecture.

4.2.5.2 Phase II - Online Prediction System for Clinicians

In the testing phase, the complete process of passing the image from the encoder and decoder stages is similar to the training phase. The difference in the training and testing phase is the testing phase uses completely unknown grayscale CCA images which were not used previously for training. Also, instead of the ground truth binary mask, test images generate the output binary mask by using the weights of the training model. This mask is compared against the model weights, and segmentation loss is calculated. No epochs are used in the testing phase (prediction phase). There is no repetition phase (unlike training).

4.2.6 Experimental Protocol using Cross-Validation

After augmentation, we have access to 758 images from the same database. We applied a popular data partitioning method (K10 fold) and used 90% images (682 images) for training and 10% images (76 images) for testing. To make our system more reliable and accurate, we used cross-validation in a cyclic order, 10 times, ensuring the training and test sets to be unique. We switched back the 10% test images into the training pool in each repetition and fetched new 10% images for testing. Thus in each repetition, we experimented with a new set of test images.

4.2.7 Plaque Quantification

An automated plaque area can be used as a CVD biomarker to classify the high-risk US image. Also, the same automatic area and ground truth (GT)-area can be compared and used in multiple ways to evaluate the system's performance. The AI models generate the binary segmentation mask corresponding to the image provided. The number of pixels in the AI-generated mask is counted and multiplied by a mean resolution factor of 0.052 mm to measure the plaque area. If n and N represent the image number and the total number of images in the cohort, respectively, and $\bar{\alpha}_{ai}(m)$ and $\bar{\alpha}_{gt}$ represent the mean plaque area for "ai" model "m", and GT binary mask, respectively, mathematically given in Eq. 4.2 as:

$$\bar{\alpha}_{ai}(m) = \frac{\sum_{n=1}^{N} \alpha_{ai}(m,n)}{N} \;\&\; \bar{\alpha}_{gt} = \frac{\sum_{n=1}^{N} \alpha_{gt}(n)}{N} \qquad (4.2),$$

4.3 Experimental Results

4.3.1 Hyper-parameters optimization and system performance

In proposed work, various hyper parameters conditions were experimented and optimized to facilitate high performance and low margin of error (MoE). Each DL experiment consisted of learning rate = 0.0001, minibatch size = 10, # epoch = 100, and CE-loss optimized by ADAM optimizer. The experiment was conducted on a Workstation with Intel Core i7-8th Generation CPU, 16 GB DDR4 RAM,

and NVIDIA Quadro P4000 8 GB DDR5 RAM with MATLAB 2020a software. We adopted the K10 protocol that consisted of 90% training and 10% testing data. We recorded training accuracy, sensitivity, specificity, dice-similarity coefficient, precision, and Jaccard index for the experiment.

Further, we switched back the test data in the original pool, fetched a new set of test data, and repeated the same experiment. A series of 10 experiments using the same hyperparameters was performed with each new testing dataset corresponding to 10 combinations. Finally, the mean and standard deviation was calculated and is shown in Table 4.2 corresponding to UNet and SegNet-UNet models.

4.3.2 Plaque Segmentation and Visual Representations

Figure 4.5 shows a zoomed-in version of the AI-estimated area (black arrow) and the difference between GT and the AI-estimated area (gray arrow). Images in column one represent the ground truth (GT) area by white arrows on original images. The middle column represents the AI-estimated plaque area (PA) by black arrows with the UNet model and the difference between GT and UNet estimated areas by gray arrows. Similarly, the last column represents the PA for the SegNet-UNet model by black arrows, and gray arrows show the difference with GT.

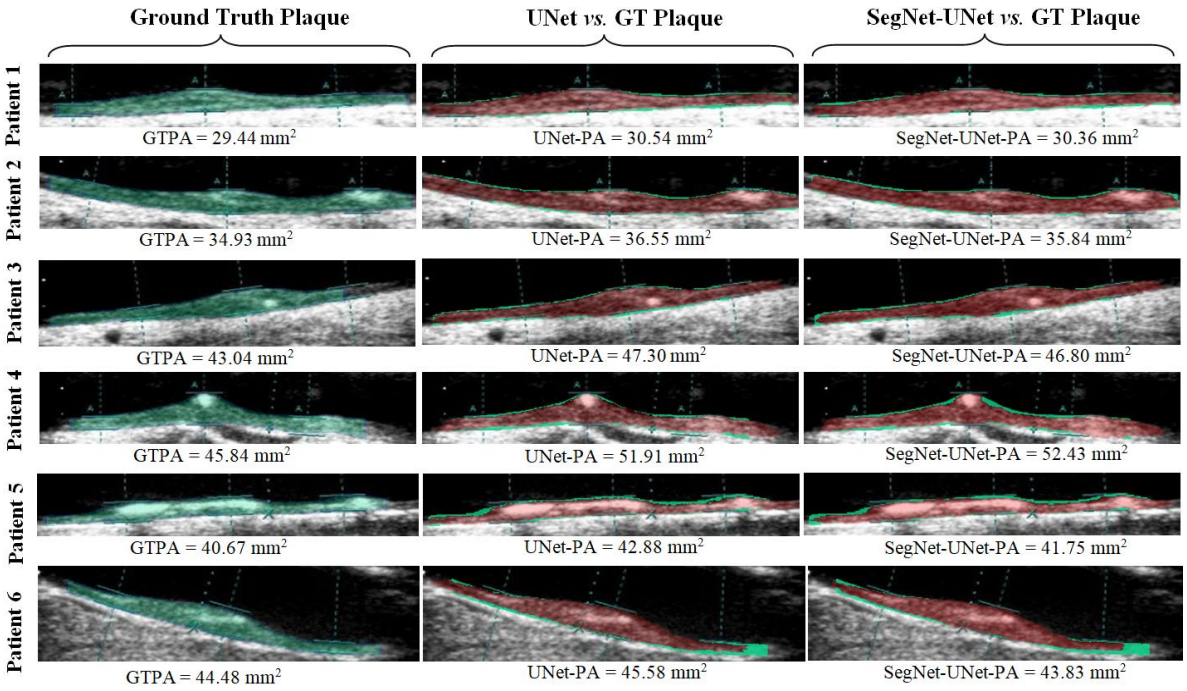

Figure 4.5 Overlays of GT (white arrow) and AI (black arrow) plaque area of the CCA US. (a) Ground truth plaque only (white arrow) (b) UNet vs GT plaque (c) SegNet-UNet vs GT plaque. Gray colour arrows show the difference between GT and AI.

4.3.3 Segmentation Performance

Using the CE-loss function, the mean accuracies for SDL/HDL models UNet, and SegNet-UNet were **98.97±0.82%, 99.08±0.66%,** respectively (Table 4.2, column 2). Other indices of segmentation show a higher value to the SegNet-UNet model. Dice-similarity-coefficient, Mathew's correlation coefficient (MCC), Jaccard index (JI), and precision for UNet and SegNet-UNet models are 86.16±9.98 and 88.23±7.75, 86.46±8.00 and 88.07±7.38, 76.72±12.10 and 79.63±10.29, and finally, 83.42±13.52 and 89.05±8.09 respectively. A bar chart showing a comparison between segmentation parameters of UNet (blue) and SegNet-UNet (red) is shown in Figure 4.6.

Table 4.2 Mean performance parameters of UNet and SegNet-UNet AI models (all in %).

Model	Accuracy	Sensitivity	Specificity	Precision	MCC	Dice	Jaccard
UNet	98.97±0.82	91.50±7.79	99.26±0.85	83.42±13.52	86.46±8.00	86.16±9.98	76.72±12.10
SegNet-UNet	99.08±0.66	88.60±10.74	99.56±0.33	89.05±8.09	88.07±7.38	88.23±7.75	79.63±10.29

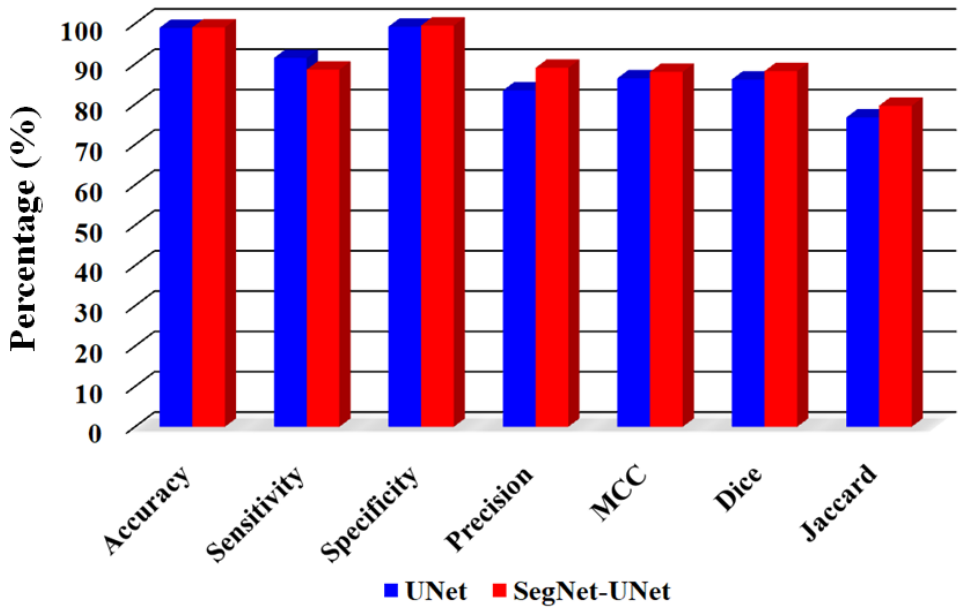

Figure 4.6 Performance parameters of the UNet and SegNet-UNet models.
MCC: Mathew's correlation coefficient.

4.4 Performance Evaluation and Statistical Tests

Various performance evaluation and statistical tests were performed on test data to validate our hypothesis. We calculated the mean of all performance and statistical parameters for both DL models and presented them in Table 4.3.

4.4.1 Correlation between AI models and Ground Truth and benchmarking against AtheroEdge 2.0

The regression analysis is an effective and straightforward tool to analyse the relationship between the two quantities. Its value lies between '0' and '1', representing a degree of match between the two quantities. CC value close to zero means a low match, and a higher value close to 1 illustrates a high match. Figure 4.7 below shows the CC values and regression curve for UNet (triangle), SegNet-UNet (empty-circle), and AtheroEdge2.0 (diamond) models against GTPA. As depicted in Figure 4.7, CC between UNet and GT is **0.82**, SegNet-UNet and GT is **0.89**, and AtheroEdge 2.0 and GT is **0.85**, all having p<0.001.

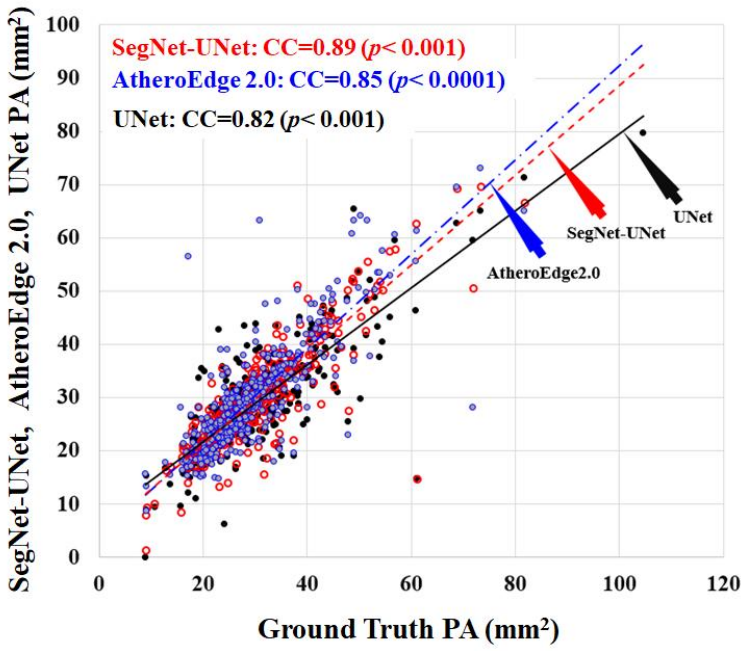

Figure 4.7 Correlation of the three models: UNet (Black), SegNet-UNet (Red), and AtheroEdge 2.0 (Blue) against GT plaque area.

Table 4.3 Performance parameters comparison of UNet, SegNet-UNet and AtheroEdge 2.0.

PE parameter	UNet	SegNet-UNet	AtheroEdge 2.0
CC	0.82	0.85	0.89
Area error*	<9.9 mm^2	<8 mm^2	<9.6 mm^2
AUC$^{\#}$	0.93	0.94	0.95
FoM	96.73%	99.07%	98.37%

* 90% image database; # PA threshold = 40 mm^2; FoM: Figure of Merit.

4.4.2 Bland-Altman Plots

Bland-Altman's plot between the UNet-GT, SegNet-UNet-GT, and AtheroEdge 2.0-GT is shown in the Figure 4.8, (a), (b), and (c). As it is clear from the plots, the SegNet-UNet AI model's output is

closely concentrated along the central lines. The other two modalities do not have such a good comparison.

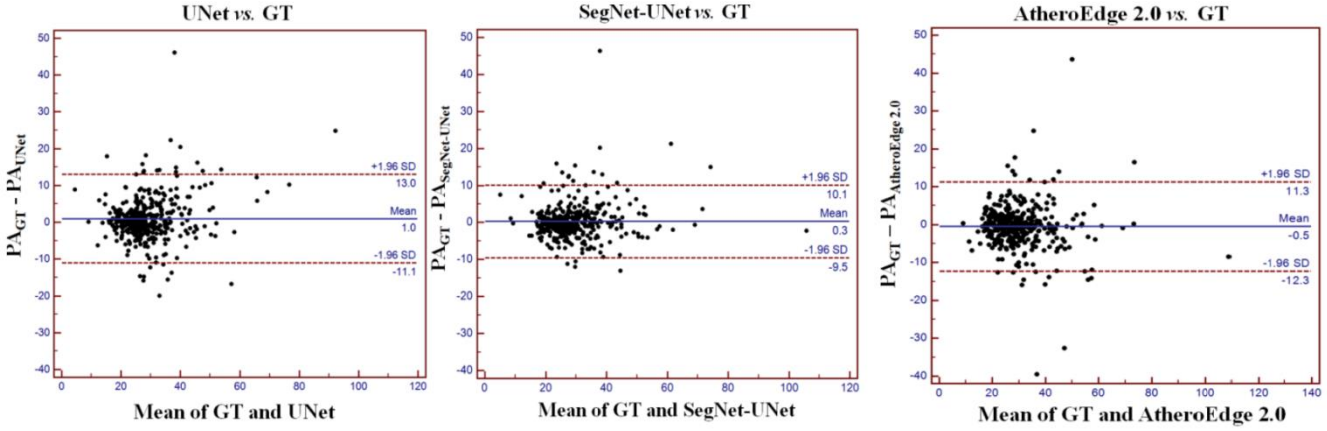

Figure 4.8 Bland-Altman plots for (a) UNet (b) SegNet-UNet (c) AtheroEdge 2.0.

4.4.3 Receiver Operating Characteristics and AUC

The receiver operating characteristics (ROC) curve and the area under the ROC curve are important performance parameters in medical image analysis. We used the ground truth plaque area (GTPA) threshold value of 40 mm^2 to generate the GT binary labels' 1' for high-risk and '0' for low-risk images. Further, the deep learning plaque area (DLPA) is used to plot the ROC curve between the GTPA labels and DLPA scores. Figure 4.9 shows ROC curves and AUC values for UNet (black), SegNet-UNet (red), and AtheroEdge 2.0 (blue) for (a) 38 mm^2 (b) 40 mm^2 and (c) 42 mm^2 as GTPA thresholds. In the discussion section, we have described the selection of the GTPA threshold in detail and the variation in the system performance (AUC) w.r.t. change in the DLPA threshold.

4.4.4 Cumulative Distribution Frequency Curve

The cumulative distribution frequency (CDF) curve is another way to measure the system performance. Figure 4.10 shows the CDF curve between area error (mm^2) on the *x*-axis and the cumulative frequency (in %) on the *y*-axis for UNet, SegNet-UNet, and AtheroEdge 2.0. As the arrow indicates, 90% of the image database have area error below 9.9 mm^2 for UNet, 8 mm^2 for SegNet-UNet, and 9.6 mm^2 for AtheroEdge 2 .0, showing comparable Cut-offs.

4.4.5 Figure of Merit

The figure of merit (FoM) is defined in terms of the central tendency of the error. Let, $\alpha_{ai}(m,n)$ and $\alpha_{gt}(n)$ represents the plaque area using AI model 'm' and GT, respectively, for the image 'n.'

Considering N as the total number of scans, the corresponding mean AI for model m and GT can be represented as $\bar{\alpha}_{ai}(m)$ and $\bar{\alpha}_{gt}$, [as defined in Eq. 4.3], then the FoM can be expressed as Eq. 4.3:

$$FoM(m) = 100 - \left[\left(\frac{|\bar{\alpha}_{ai}(m) - \bar{\alpha}_{gt}|}{\bar{\alpha}_{gt}}\right) * 100\right] \quad (4.3)$$

A FoM for UNet, SegNet-UNet, and AtheroEdge 2.0 models using the above method was then calculated which can be given as **96.73%, 99.07%,** and **98.37%,** respectively.

4.4.6 Paired sample t-Test and Wilcoxon Test

Paired Sample T-test: Figure 4.11 (a), (b), and (c) below shows the paired sample t-tests between UNet-GT, SegNet-UNet-GT, and AtheroEdge 2.0-GT using *box and whiskers*. The box and whiskers plot is a convenient way to show the data distribution. As the boxes in the plot show, the median values of SegNet-UNet and AtheroEdge 2.0 are close to the GT median values.

Wilcoxon Test: The Wilcoxon test is a non-parametric statistical hypothesis test to compare two samples to measure their mean rank. We performed Wilcoxon test between GTPA and UNet, SegNet-UNet, and AtheroEdge 2.0 area. The results of the tests are shown by box and whiskers plot in Figure 4.12 (a), (b), and (c), respectively. The GT and SegNet-UNet has a minimal difference of mean plaque area.

4.4.7 Sample Size Calculation using Power Analysis

The selection of optimal sample size is a crucial stage in SDL/HDL system design in terms of computational cost and time. Any significant variation in sample size may lead to subdued results or higher computation time. The mathematical formula to represent the sample size calculation is shown in Eq. (4.4). We used a z-score from the standard z-table [169].

$$N_p = (z)^2 \times \frac{\hat{p}(1-\hat{p})}{(MoE)^2}, \quad (4.4),$$

Where MoE is the margin of error, and $\hat{p}$ represents data proportion. Using MoE of 4.45 or confidence intervals of 95.54%, the standard z-score is 1.75. Thus, for 50% data proportion, our required sample size is 364. Thus, our CCA database pool sample size is 108.24% (almost double) higher than the number of samples needed.

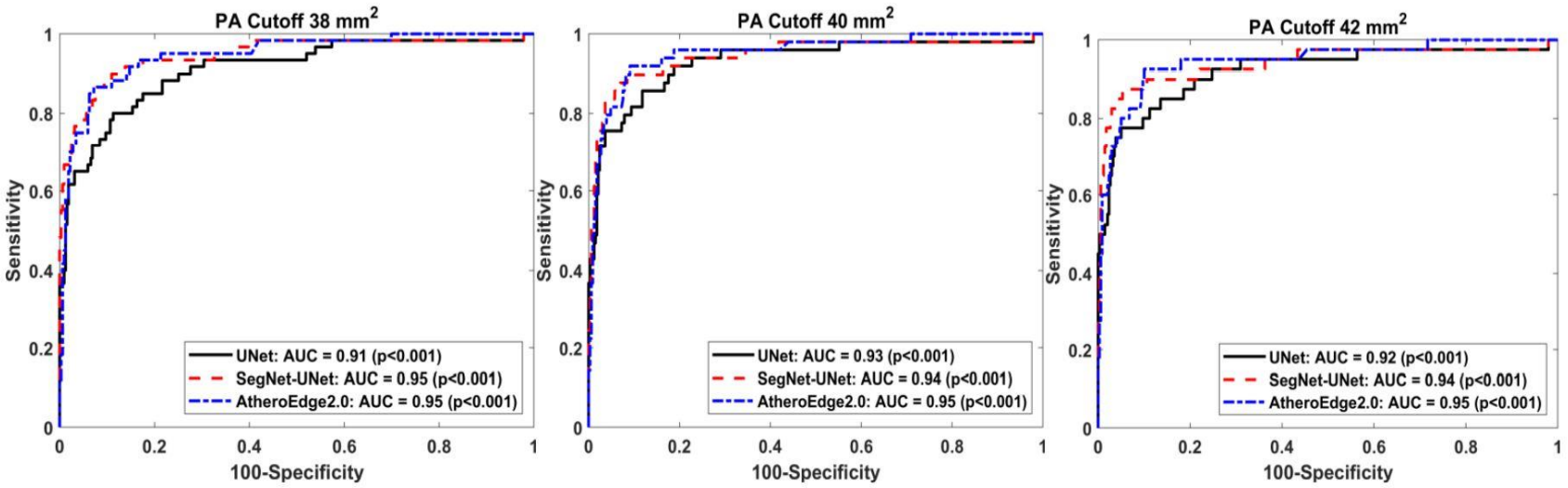

Figure 4.9 Receiver Operating Characteristics and AUC for UNet, SegNet-UNet and AtheroEdge 2.0 for three kinds of PA Cut-offs (a) 38 mm^2 (b) 40 mm^2 (c) and 42 mm^2.

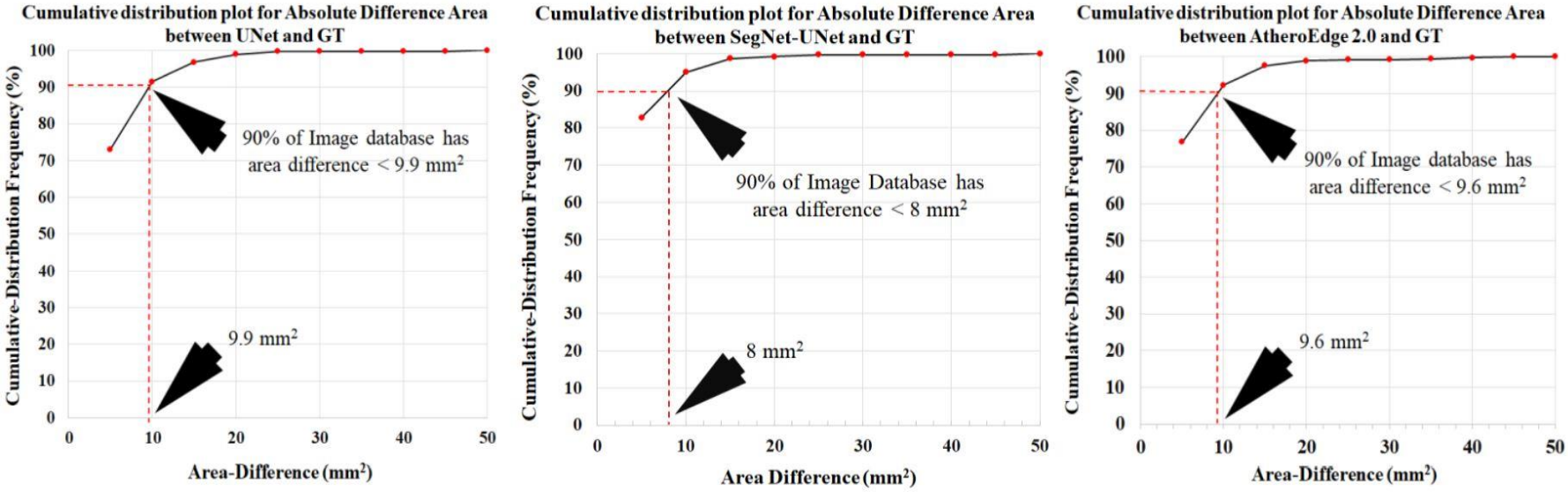

Figure 4.10. Cumulative Distribution Frequency curve and area error threshold for 90% image database. (a) UNet: **9.9 mm^2** ; (b) SegNet-UNet: **8 mm^2**; (cc) AtheroEdge 2.0: **9.6 mm^2**.

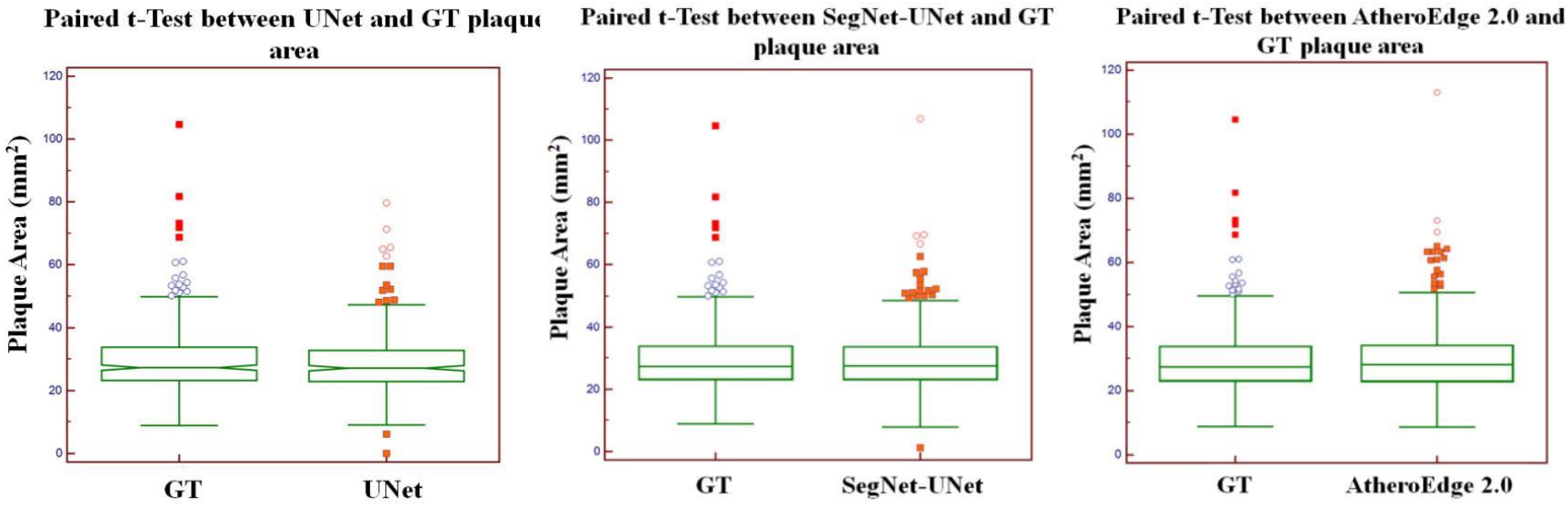

Figure 4.11 Paired sample t-Test between (a) GT and UNet (b) GT and SegNet-UNet and (c) GT and AtheroEdge 2.0.

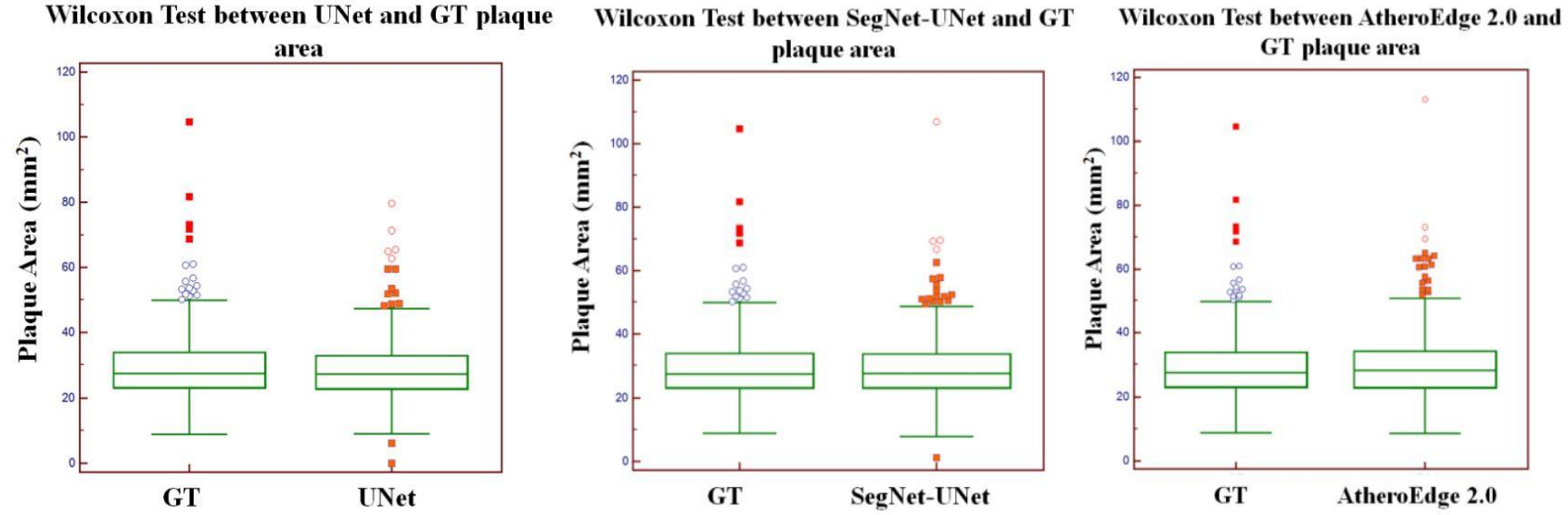

Figure 4.12 Wilcoxon test between (a) GT and UNet (b) GT and SegNet-UNet and (c) GT and AtheroEdge 2.0.

4.5 Discussion

The main contribution of our study consists of the design of automated plaque and wall segmentation using two deep learning AI methods, such as SDL (UNet) and HDL (SegNet-UNet) under the class AtheroEdge™ 3.0. This model was further benchmarked against the previously available system AtheroEdge™ 2.0 (AtheroPoint™, Roseville, CA, USA). The system performance of HDL was superior to SDL in terms of figure-of-merit, AUC, and area error against the gold standard. Finally, the UNet and SegNet-UNet were correlated with GT showing CC of **0.82** and **0.89** (both $p<0.001$), respectively, while AtheroEdge™ 2.0 was **0.85** ($p<0.001$).

4.5.1 Benchmarking

This section presents a benchmark study of all DL techniques, which involves plaque area measurement. Table 4.4 shows the studies adapted to benchmark against our proposed models for plaque area measurement. Biswas *et al.* [107] analysed the same Japanese diabetic cohort of 204 patients with mean age 69±11 years, mean HbA1c 5.8±1.0 mg/dL, mean glucose 108±31 mg/dL, mean LDL 99.80±31.30 mg/dL, mean HDL 50.40±15.40, and mean total cholesterol 174.6±37.7 mg/dL. The authors randomly selected only 250 images of right and left CCA (unlike our current study used 379 images) and converted them into 4000 patches (16 patches/image) for their DL-based system. Biswas *et al.* [107] proposed a two-stage DL system that finds the cIMT and PA in CCA images. The first stage took the input patch images and classified them into the wall and non-wall ROI. The measured PA error in 250 images is 2.7939±2.3702 mm^2. The second DL system segments the cIMT and plaque area. Even the system was DL-based; it was a concatenation of two independent DL systems. Such a system is vulnerable to instability, unlike our one fully automated system.

In their recent study, Zhou *et al.* analysed two datasets of 144 and 497 subjects named SPARC and Zhongnan datasets [169]. One hundred forty-four (144) subjects of the SPARC dataset constituted 510 plaque images, which they further divided into R1=33, R2=33, R3=34 for training, and 44 test subjects. Baseline characteristics of SPARC dataset includes mean age: 65±7 years, BMI: 29.03±4.99, M/F:97/47, SBP: 123.08±13.04, DBP: 72±7.99, TG: 1.15±0.60, LDL: 1.86±1.04 mmol/dL, 1.43±0.41 mmol/dL, mean GT area: 21.47±13.78 mm^2 . Zhou *et al.* proposed three DL models UNet, UNet++, and Deeplab v3+ [98]. The measured ΔTPA between AI and GT was 4.39±3.33, 3.34±2.31, and 6.81±7.11 mm^2, respectively for R1, R2, and R3 SPARC sub-dataset using UNet++ model. Their UNet model ΔTPA results for the same sub-datasets are -6.75±14.63, -2.19±15.75, and 0.29±12.20 mm^2, respectively, which are comparable to our proposed models. The global online system is shown in Figure 4.1, a class of AtheroEdge™ 3.0 (AtheroPoint™, Roseville, CA, USA). It consists of two main components: (i) plaque segmentation using AI models using SDL and HDL architectures and (ii) risk stratification for high-risk

patient detection based on moderate risk Cut-off of 40 mm^2 for detection of high-risk patients. Finally, we benchmark the proposed system against previously published studies, demonstrating an AUC of **0.92** for UNet and **0.94** for SegNet-UNet architectures. The system accuracy showed **98.97±0.82**% for UNet and **99.08±0.66**% for the SegNet-UNet system. The system is fast and runs in less than one second per image. Our proposed absolute ΔPA was 4.07±4.71 mm^2 and 3.12±3.93 mm^2 for UNet and SegNet-UNet DL models.

Table 4.4 Benchmarking table showing studies related to PA measurement.

Citation	AT	Techniques	CV	Epochs	# Patients	# Images	Area Error (mm^2)
Biswas *et al.* [107]	CCA	CNN+FCN	K10	100	125	250	ΔTPA=2:7939±2:3702
Zhou *et al.* [98]	CCA,ICA	UNet, UNet++, Deeplab v3+	Tr: 33, 34,33 Te:44	200	R1=33; R2=33; R3=34; Test1=44 Test2=497	DB1=510 DB2=638	ΔTPA(R1)=0.73±9.63; ΔTPA(R2)=3.91±9.46; ΔTPA(R3)=6.75±10.04
Proposed Model 1	CCA	UNet	K10	100	190	379	Absolute ΔPA=4.07±4.71
Proposed Model 2	CCA	SegNet-UNet	K10	100	190	379	Absolute ΔPA=3.12±3.93

AT: Artery Type; CCA: Common carotid artery; ICA: Internal Carotid Artery; CV: Cross-Validation; Cohort: Size of the data set. R1, R2, and R3 are number of patients; DB1, DB2 are databases; PA: plaque area.

4.5.2 The Rationale for Plaque Area Cut-off

Table 4.5 shows some benchmark studies which involve the measurement of PA measurement in CCA. Adams *et al.* [205] conducted a clinical study on 382M/82F with mean age 54±6/54±5 years and proposed granularity of risk as Type I, Type IIa, Type IIb, Type III, Type Iva, and Type IVb. All risk criteria (except Type 1) involved cIMT and age-related PA Cut-off criteria. They suggested the lowest PA Cut-off of 60 mm^2 for the age of 40 years. The purpose of their study was to find Type III and Type IVb patients in asymptomatic patients with advanced atherosclerosis in carotid arteries. In another study, Romanens *et al.* [206] suggested a PA Cut-off > 80 mm^2 as a coronary risk identifier (TPA80) in young adults in a German and Swiss cohort. Their study suggested a 20% risk of fatal and nonfatal MI in 10 years in patients with TPA80.

Alsulaimani *et al.* investigated a relation between elevated homocysteine levels and atherosclerotic plaque development in their North Manhattan Study (NOMAS) [207]. Their study included 1327 stroke-free subjects with a mean age of 66±9 years, 41% men, 19% Black, 62% Hispanic, and 17% White residents of Manhattan. Based on the grayscale median value (GSM) of the plaque and measured PA, they classified subjects into mild, moderate, and severe patients. Their suggested PA Cut-off was mild=5.2 mm^2, moderate=14.2 mm^2, and severe=43.2 mm^2. In their recent clinical study, Spence *et al.* presented a regression between LDL cholesterol and atherosclerotic plaque in the SPARC dataset

[208]. They observed a PA Cut-off of 113.33±121.52 mm^2 earlier than 2003 for n_1=2025 patients; however, with the new approach of treating the arteries, they evaluated the PA Cut-off of 129.56±134.32 mm^2 for n_2 = 1174 patients. Another research group Cuadrado *et al.* performed their study of Japanese diabetic data and found a mean PA of 31.27±9.75 mm^2 and 30.48±9.90 mm^2 [44]. Thus we can conclude that the selection of plaque Cut-off is based on various geometrical and clinical factors.

4.5.3 Correlation between AI and Significant Baseline Covariates

As per the baseline characteristics Table 1, certain covariates such as hypertension, family history, smoking, and cIMT values, showed strong significance (having *p*-values<0.001). We, therefore, decided to evaluate these covariates against the DLPA. Based on the DLPA classification (DLPA>40 mm^2), 33 patients (17.37%) who had hypertension are classified as high-risk. Figure A.1 shows a scatter plot between the number of patients with HT labels (HT=1, and no HT=1) and plaque area values. Similarly, 12 patients (6.32%) labelled as smokers are identified as high-risk patients using the DLPA classification. Figure A.2 shows a scatter plot between the number of patients with smoking labels (smoker = 1 and non-smoker = 0) and AI PA values. Six patients (3.16%) showed a history of cardiovascular events and were labelled as 1. These patients are classified as high-risk patients (3.16%) using the criteria of DLPA > 40 mm^2. Figure A.3 shows a scatter plot between the number of patients with a history of CVEs (FH = 1 and no FH = 0) and plaque area values.

4.5.4 Sensitivity Analysis of Plaque Area Cut-off

For the following experiment base, PA Cut-off 40 mm^2 was chosen for UNet and SegNet-UNet models, which gave AUC of 0.93 ($p<0.001$) and 0.94 ($p<0.001$), respectively (see Table 4.6). In this section, we discuss variation in PA Cut-off and its effect on the performance of the system. We took ±5% variation in PA Cut-off with respect to the base Cut-off of 40 mm^2, and this accounted for PA Cut-offs of 38 mm^2 and 42 mm^2. On decreasing the PA Cut-off from 40 mm^2 to 38 mm^2, UNet performance decreased by 2.15%, while SegNet-UNet showed an increase by 1.06%. On increasing the Cut-off from 40 mm^2 to 42 mm^2, the performance of UNet was reduced by 1.08%, and the performance of SegNet-UNet was unchanged. Thus we conclude that change in PA Cut-off has a less than 2% effect on system performance. Also, UNet is more sensitive to PA Cut-off change compared to SegNet-UNet.

Table 4.5 Benchmark studies related to TPA cut-off, granularity of risk, cohort type.

SN		Patient	Age	Cohort type	Purpose	# of Classes (Granularity of Risk)	TPA (mean)
1	Adams *et al.* [51]	M=382 F=82	M = 54±6 F = 54 ± 5	Asymptomatic patients with advanced vascular atherosclerosis of carotid artery	Finding Type III and IVb patients	Type I: TPA upto 24 mm^2 Type IIa: IMT ≤2 mm and plaque area < age criteria Type IIb: IMT ≤2 mm and plaque area > age criteria Type III: IMT ≥3.5 mm and plaque area < age criteria Type IVa: IMT > 2 mm and plaque area < age criteria Type IVb IMT > 2 mm and plaque area > age criteria	M=150 ± 63 F=120 ± 62 Age related TPA Cut-off < 60 mm^2 for age 40 yrs; <80 mm^2 for 40-49 yrs; <110 mm^2 for 50-59 yrs; 130 mm^2 for 60-64 yrs;
2	Romanens *et aL*[52]	N1=2202 N2=2942	N1=57±9 N2=46±10		Assessment of the sensitivity of different risk calculators for detection of TPA80	-	N1= 52±50 N2=36 ± 50
3	Alsulaimani *et al.* [53]	N= 1327 M:F::41:59	66±9	North Manhattan Study (NOMAS)	Linkage of elevated homocysteine (tHcy) levels to atherosclerotic carotid plaque development	Echolucent Plaque (GSM# : 17.0-80.7) Intermediate Density plaque (GSM: 81.0-103.3) Echodense plaque(GSM 103.5-180) TPA: Mild, moderate and severe plaque.	Mean = 20.3±20.6 mm^2 Mild= 5.2 mm^2 Moderate =14.2 mm^2 Severe=43.2 mm^2
4	Spence *et al.* [54]	n1=2025 (F=982) n2=2487 (F=1174)	n1= 60.95±12.65 n2= 63.59±13.41	SPARC Stroke Prevention and Atherosclerosis	To find a relation between change in LDL-C to the progression of atherosclerosis before and after 2003	Regressive: Progressive: Who shows consistent plaque growth over the year. Stable: Who shows no plaque growth.	n1= 113.33±121.52 mm^2 n2= 129.56±134.32 mm^2
6	Elisa *et al.* [19]	N=199 (398) M/F= 157/42	M=67 F=75 68.96 ± 10.98	Japanese diabetic cohort	Finding cIMT and TPA using AtheroEdge2.0	High Risk (HbA1c ≥ 6.5 mg/dl): 150 Low Risk (HbA1c < 6.5 mg/dl): 49	O1=31.27 ± 9.75 O2=30.48 ± 9.90 (AtheroEdge2.0)

GSM: Gray Scale Median

Table 4.6 Variation in AUC w.r.t. change in PA cut-off.

	AUC-based on PA Cut-off		
AI Model	Cut-off: 38 mm^2	Cut-off: 40 mm^2	Cut-off: 42 mm^2
UNet	0.91	0.93	0.92
SegNet-UNet	0.95	0.94	0.94
AtheroEdge 2.0	0.95	0.95	0.95

4.5.5 Strength, Weakness, and Extensions

Our HDL system labelled as AtheroEdge™ 3.0 was able to characterize the plaque in CCA scans successfully. It demonstrated encouraging results in terms of performance and statistical tests. Even though the system was robust and took less than one second, the system can be improved in low-contrast US scans. Further, the system can be extended toward US videos or motion imagery in real-time in a big data framework. The system can also be developed to estimate the composite CVD risk based on estimated plaque area similar to our previous methods [209, 210] or study the relationship of HDL-area against other biomarkers [44]. Lastly, the system can be extended to estimate the plaque in near-wall and stenosis estimation [38, 211].

4.6 Conclusion

We hypothesized that deep learning could be a powerful solution for risk stratification of asymptomatic patients with carotid artery disease, an established biomarker for cardiovascular disease risk. We showed that both AI-based models such as UNet and SegNet-UNet had successfully identified the high-risk patients in the database of 379 ultrasound scans from a Japanese cohort with moderate risk. Further, it showed comparable results to AtheroEdge 2.0 (AtheroPoint, Roseville, CA, USA). Both AI models were also associated with baseline patient demographics having covariates such as hypertension, family history, smoking, and cIMT values. The online AtheroEdge™ 3.0 runs on a test US image in < 1 second, proving that the system can be adapted in clinical settings.

4.7 Proposed Extension for Next Chapter

The current work demonstrates the application of SDL and HDL on low-to-moderate risk CCA plaque images. However, the current work uses database from same ethnic patients. Also, the images are scanned using one type of scanner. Thus to avoid the data selection bias, the system should use multi-ethnic, multi-centre database. Also, the system should be tested on unseen database. In the next chapter we have proposed an unseen deep learning based system which fills the gap from present chapter.