Accepted Manuscript

Deep learning strategy for accurate carotid intima-media thickness measurement: An ultrasound study on Japanese diabetic cohort

Mainak Biswas, Venkatanareshbabu Kuppili, Tadashi Araki, Damodar Reddy Edla, Elisa Cuadrado Godia, Luca Saba, Harman S. Suri, Tomaž Omerzu, John R. Laird, Narendra N. Khanna, Andrew Nicolaides, Jasjit S. Suri

PII: S0010-4825(18)30127-6

DOI: 10.1016/j.compbiomed.2018.05.014

Reference: CBM 2965

To appear in: *Computers in Biology and Medicine*

Received Date: 4 May 2018

Revised Date: 10 May 2018

Accepted Date: 10 May 2018

Please cite this article as: M. Biswas, V. Kuppili, T. Araki, D.R. Edla, E.C. Godia, L. Saba, H.S. Suri, Tomaž. Omerzu, J.R. Laird, N.N. Khanna, A. Nicolaides, J.S. Suri, Deep learning strategy for accurate carotid intima-media thickness measurement: An ultrasound study on Japanese diabetic cohort, *Computers in Biology and Medicine* (2018), doi: 10.1016/j.compbiomed.2018.05.014.

Deep Learning Strategy for Accurate Carotid Intima-Media Thickness measurement: an Ultrasound Study on Japanese Diabetic Cohort

Mainak Biswas[a], MTech., Venkatanareshbabu Kuppili[a], PhD., Tadashi Araki[b], MD., Damodar Reddy Edla[a], PhD., Elisa Cuadrado Godia[c], MD., Luca Saba[d], MD., Harman S. Suri[e], Tomaž Omerzu[f], MD., John R. Laird[g], MD., Narendra N Khanna[h], MD, DM, FACC, Andrew Nicolaides[i,j], PhD., Jasjit S. Suri[k,*] PhD., MBA, Fellow AIMBE

[a] National Institute of Technology Goa, India
[b] Toho University Ohashi Medical Center, Tokyo, Japan
[c] IMIM – Hospital del Mar, Passeig Marítim 25-29, Barcelona, Spain
[d] Department of Radiology, Policlinico Universitario, Cagliari, Italy
[e] Brown University, Providence, RI, USA
[f] Department of Neurology, University Medical Centre Maribor, Slovenia
[g] Cardiology Department, St. Helena Hospital, St. Helena, CA, USA
[h] Cardiology Department, Apollo Hospitals, New Delhi, India
[i] Vascular Screening and Diagnostic Centre, London, UK
[j] Department of Biological Sciences, University of Cyprus, Nicosia, Cyprus
[k] Stroke Monitoring and Diagnostic Division, AtheroPoint™, Roseville, CA, USA

***Corresponding author:**
Jasjit S. Suri, PhD., MBA, Fellow AIMBE
Stroke Monitoring and Diagnostic Division, AtheroPoint™, Roseville, CA, USA.
email: jasjit.suri@atheropoint.com
(916)-749 5628

Deep Learning Strategy for Accurate Carotid Intima-Media Thickness measurement: an Ultrasound Study on Japanese Diabetic Cohort

Abstract

Motivation: The carotid intima-media thickness (cIMT) is an important biomarker for cardiovascular diseases and stroke monitoring. This study presents an intelligence-based, novel, robust, and clinically-strong strategy that uses a combination of deep-learning (DL) and machine-learning (ML) paradigms.

Methodology: A two-stage DL-based system (a class of AtheroEdge™ systems) was proposed for cIMT measurements. Stage I consisted of a convolution layer-based encoder for feature extraction and a fully convolutional network-based decoder for image segmentation. This stage generated the raw inner lumen borders and raw outer interadventitial borders. To smooth these borders, the DL system used a cascaded stage II that consisted of ML-based regression. The final outputs were the far wall lumen-intima (LI) and media-adventitia (MA) borders which were used for cIMT measurements. There were two sets of gold standards during the DL design, therefore two sets of DL systems (DL1 and DL2) were derived.

Results: A total of 396 B-mode ultrasound images of the right and left common carotid artery were used from 203 patients (Institutional Review Board approved, Toho University, Japan). For the test set, the cIMT error for the DL1 and DL2 systems with respect to the gold standard was 0.126 ± 0.134 and 0.124 ± 0.100 mm, respectively. The corresponding LI error for the DL1 and DL2 systems was 0.077 ± 0.057 and 0.077 ± 0.049 mm, respectively, while the corresponding MA error for DL1 and DL2 was 0.113 ± 0.105 and 0.109 ± 0.088 mm, respectively. The results showed up to 20% improvement in cIMT readings for the DL system compared to the sonographer's readings. Four statistical tests were conducted to evaluate reliability, stability, and statistical significance.

Conclusion: The results showed that the performance of the DL-based approach was superior to the nonintelligence-based conventional methods that use spatial intensities alone. The DL system can be used for stroke risk assessment during routine or clinical trial modes.

Keywords: cardiovascular diseases; stroke; ultrasound scans; carotid intima-media thickness; intelligence; deep learning; machine learning; segmentation; accurate; reproducible.

1. Introduction

Stroke due to cardiovascular disease (CVD) causes the death of approximately five million people and disability among another five million people around the world [1]. In the USA, 795,000 people suffered from a stroke in 2010, causing direct medical care costs of approximately USD 33 billion and indirect costs of around USD 20.6 billion [2]. Stroke is generally caused by the blockage or rupturing of the common carotid artery (CCA) or internal carotid artery (ICA) that supply blood to the brain. This blockage or rupture is triggered by the formation of plaque along the arterial walls. Plaque is usually composed of cholesterol, fatty substances, cellular waste products, calcium, and fibrin and is generally formed between the lumen-intima (LI) and media-adventitia (MA) interfaces [3].

Carotid intima-media thickness (cIMT) is the mean perpendicular distance between the LI and MA interfaces and is an important biomarker for CVDs. A comprehensive risk analysis study on 5858 subjects revealed that cIMT values >1.18 mm led to an increased stroke rate [4]. In 2006, the findings of Bots et al. [5] showed that cIMT was related to the presence of atherosclerosis in the coronary artery. Risk prediction models developed by Nambi et al. [6] showed an increase in the CVD risk when cIMT and plaque information was added. The study of Meuwese et al. [7] suggested that an increase in cardiovascular risk was related to an increase in mean cIMT. Ikeda et al. [8] also confirmed the significant association between cIMT and CVDs. All of the abovementioned studies indicated that an increase in cardiovascular events (myocardial infarction) was correlated to an increase in the mean cIMT.

Although Suri et al. have diligently worked to standardize cIMT measurements [9], there are still challenges regarding accuracy and reproducibility when it comes to the CCA, ICA, and bulb regions. Several reasons contribute to this, including the variability in the studies with regards to nationality, ethnicity, disease, age groups, etc. In this regard, a concerted effort was made to construct a multiinstitutional dataset using multiple ethnicities and varying age groups [10]. There are other technical challenges associated with the cIMT measurement. For example, the images are obtained through a B-mode ultrasound (US) using a linear probe that is manually operated. The CCA extends from the jaw to the shoulder bone; however, the linear probe is unable to cover the entire carotid artery length and has to be performed in sections (i.e., distal, mid, and proximal). The CCA image quality also depends on external factors such as speckle noise, probe position, neck position, probe orientation (i.e., anterior, posterior, or posterior lateral), probe contact with the skin, linear frequency usage, gain

control, dynamic range, and features such as harmonic and compound imaging [11,12]. The traditional manual segmentation of US images is slow, error-prone, and subject to intra- and interobserver variability. Therefore the measurement of cIMT through automated methods is a growing field of interest. The previously-described completely-automated methods are briefly discussed here.

Molinari et al. [9] compared four automated techniques for cIMT measurement. The first method was Completely Automated Layers Extraction (CALEX), which is based on the integration of three approaches: feature extraction, line fitting, and classification. The second was Completely Automated Robust Edge Snapper (CARES) [13], which is based on the combination of feature extraction and edge detection. The third methodology was Completely Automated Multiresolution Edge Snapper (CAMES), which is based on a multiresolution approach and uses the concept of scale–space (SS) [14]. Finally, the fourth methodology was the Carotid Automated Double-Line Extraction System, which is based on edge flow (an edge-detection technique based on US texture and edge energies) [15]. Saba et al. [16] proposed a fully-automated system (AtheroEdge™) for cIMT measurements, while Ikeda et al. [17] proposed a cIMT measurement system with a classification paradigm that used a combination of global and local strategies involving texture-based entropy and morphology. Saba et al. [18] later developed a fully-automated cloud-based solution called AtheroCloud™ for cIMT measurements. Ikeda et al. [19] recently proposed an automated segmental cIMT measurement technique that used an automated bulb-edge point as a reference marker. The abovementioned methodologies use various features such as greyscale median, pixel classification, gradient edges, SS, or a combination of these features to predict the cIMT risk assessment. Despite their strong contributions, these external factors make the spatial-based methods prone to variability and a lack of robustness when it comes to completely automated designs.

Another challenge in the segmentation of wall interfaces is the presence of shadows on the far wall due to calcium in the near wall. This causes border position errors in the detection of LI and MA, even though the average cIMT error is well below the acceptable level. Previous methods took advantage of multiresolution approaches to increase the processing speed; however, the feature extraction at multiple levels was not derived, thus lacking a comprehensive spatial deck of information. Another important point to note is that carotid US cohorts contain shape information that can be learned via neural networks, the intelligence power of which is unsurpassable. This current deep learning (DL)-based study removes all of

the abovementioned challenges to provide reliability and robustness. The spirit of this study was motivated by the work of Suri et al., who applied machine intelligence in different fields of medicine including gynecology, urology, dermatology, neurology [20–23], and recently in endocrinology [24].

This study proposed the same intelligence-based paradigm [25,26] for cIMT measurements. It was hypothesized that by training deep layers of neural networks, the DL-based system could produce more reliable and accurate results when compared to previous methods. Unlike machine learning (ML), DL can generate its own features and thus eliminate the need for less accurate feature-extraction algorithms. The high-level features of DL are more distinctive than the features of conventional methods, thus resulting in a more accurate output. The superior training of the deep layers within the DL system allows the further provision of better regional segmentation output compared to the conventional methods. The proposed DL-based system is implemented in four phases as shown in Fig. 1.

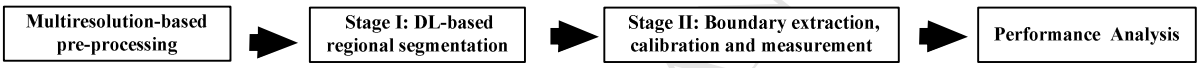


Figure 1. Overall concept of the DL-based cIMT measurement system.

Phase I is primarily adapted for data preparation. It removes the nontissue region [27] and prepares the image data using a multiresolution approach to speed up the DL paradigm. This phase is also responsible for generating the cross-validation protocol that splits the cohort into sets of training and testing carotid scans. Phase II is the heart of the DL system that performs the number crunching and consists of encoder and decoder neural networks [28,29]. In this phase, deep intelligence is derived by externally controlling the number of loops (up to 20,000). The training system uses two kinds of gold standards, namely, lumen regional information and interadventitial regional information, which leads to the design of two DL systems: DL1 and DL2. Phase III performs the boundary extraction that changes regional information to vertex point information (i.e., LI and MA boundaries or the so-called raw DL borders). This phase also ensures smooth boundaries, which attempts to get closer to the ground truth (GT) using a ML-based system, which in turn increases the overall accuracy of the system. The cIMT values are computed from the LI and MA far walls using the standardized polyline distance metric method (PDM) (discussed in Appendix A). The last phase implements performance analysis alongside risk stratification. Four statistical tests were used to assess the statistical significance: paired t-test, Mann–Whitney test, Wilcoxon test, and the Kruskal–Wallis test.

In this paper, Section 2 discusses the data collection and patient demographics. Sections 3 and 4 present the methodology and results, respectively. Section 5 shows the performance evaluation, and Section 6 discusses the statistical tests and risk analysis. The discussion and benchmarking are presented in Section 7, and, finally, the conclusions are presented in Section 8.

2. Data Demographics and US Acquisition

In this study, 204 patients (157 male and 47 female) with a mean age of 69 ± 11 years were selected. One left carotid image was not available out of the 408 images, therefore the database initially contained 407 images. Eight left carotid and three right carotid US images were rejected due to a lack of greyscale tissue information (including one patient whose left and right carotid images were removed). Thus the final dataset consisted of 396 carotid scans (left and right) from 203 patients. The sonographer's far wall cIMT readings were also available for 193 patients (346 US scans).

Informed consent was obtained from all patients and the Institutional Review Board (IRB), and ethical approval was granted by Toho University, Japan. The mean hemoglobin, glucose, low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, and total cholesterol values were 5.8 ± 1.0, 108 ± 31, 99.80 ± 31.30, 50.40 ± 15.40, and 174.6 ± 37.7 mg/dL, respectively. Of the pool of 203 patients, 92 were regular smokers. Hypertensive and high-cholesterol patients were receiving adequate medication; for example, 93 patients were taking statins to lower their cholesterol levels and 84 were receiving renin–angiotensin system antagonists. Blood pressure statistics for the patients were not available.

A sonographic scanner (Aplio XV, Aplio XG, Xario; Toshiba, Inc., Tokyo, Japan) equipped with a 7.5 MHz linear array transducer was used to examine the left and right carotid arteries. All scans were performed under the supervision of an experienced sonographer (15 years of experience). High-resolution images were acquired as per the recommendations of the American Society of Echocardiography Carotid Intima-Media Thickness Task Force. The mean pixel resolution in the database was 0.05 ± 0.01 mm/pixel.

Manual tracing of the lumen and adventitia borders was performed using ImgTracer™ (AtheroPoint™, Roseville, CA, USA), which is a user-friendly commercial software [30]. The number of points varied with the length of the carotid artery. The software zooms into the image for better visualization of the wall and provides a set of traced (x,y) coordinates.

3. Methodology

The heart of the system is an intelligence-based DL platform that supports the extraction of deep features and thereby eliminates the need for algorithms that perform poorly for feature extraction. The platform consists of two DL networks: encoder, which is used for feature extraction [28], and decoder, which is used for regional segmentation of the lumen region (LR) or interadventitial region (IAR) [29]. The DL system design allows the LR segmentation to be run in parallel with the IAR segmentation. This is called the regional segmentation block, which is the second phase of the system. Before feeding the binary training images for a LR and an IAR into the DL block, the system design expects the input data to be prepared accordingly for the DL block (the so-called multiresolution block or phase I as shown in Fig. 2). The encoder–decoder is phase II of the DL system. The image processing pipeline is always cascaded with a fine tuner to smooth or refine the outputs, therefore a ML-based system is used to extract LI-far and MA-far borders as part of the phase III subsystem. Finally, performance evaluation is implemented to benchmark the results. This is phase IV of the entire pipeline where the cIMT is measured and undergoes statistical testing. A detailed description of the system is shown in Fig. 2, and the details of these phases and their mathematical representations are discussed below.

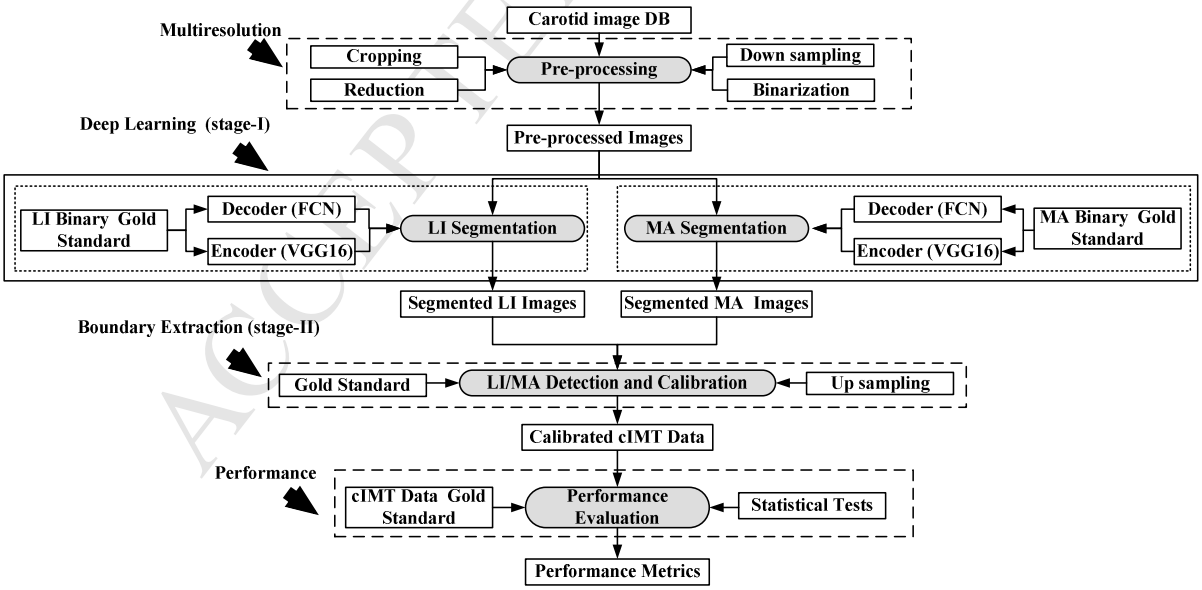


Figure 2. The four phases of a DL-based system (a class of AtheroEdge™ system, AtheroPoint™) shown in arrows. Phase I, multiresolution; phase II, the DL-based system; phase III, boundary extraction and calibration, and phase IV, performance analysis.

Multiresolution as phase I

The objective of phase I was to prepare the data for adaptability to the DL system, which required the greyscale training cohort to be cropped to remove the nontissue information [27]. This automated cropping ensured that the tissue region was retained. The greyscale images were reduced by a further 10% to ensure that very low contrast regions around the edges of the image were eliminated. These greyscale images were downsampled to improve the processing speed of the DL system under the multiresolution paradigm. In the data preparation block, the binary mapped images were also created which were mapped on a one-to-one basis with the greyscale downsampled carotid US scans. If the DL system was prepared for LR extraction, then binary maps corresponded to the LR. Conversely, if the DL system was prepared for IAR extraction, then the binary maps corresponded to the IAR. These LR and IAR binary maps were considered as the gold standard, as their borders were manually traced by experts.

DL as phase II

The DL-based system consisted of two subsystems: encoder and decoder. The encoder extracted features from the images while the decoder created segmented images from the features. The encoder consisted of 13 convolution layers and five max-pooling layers of the VGG16 network [31]. Details of the decoder network are given in Appendix B. The weights were initialized using pretrained Visual Geometry Group (VGG) weights. The convolution layers generated high-level features from the input data, and the max-pooling layers downsampled the input feature values.

The decoder consisted of three upsampling layers of the fully convolutional network (FCN) [29]. The upsampling layers upsampled the input features but with a twist. It employed two skip operations that helped recover spatial information resulting in highly accurate and crisp segmentation images. Additional information about the skip operation is presented in the discussion section. The upsampled layers were initialized using VGG weights. The cross-entropy loss function employed for segmentation was:

$$\theta_{class}(\beta_1, \beta_2) = \frac{1}{|N|} \sum_{n \in N} \sum_{l \in L} \beta_{2_n}(l) \log \beta_{1_n}(l) \quad (1)$$

where β_1 is the prediction, β_2 is the gold standard or GT, L is the total number of classes, and N is the total number of images. The loss function was defined as the difference between true

and predicted probability distributions. The DL-based system ran for 20K iterations, and intermediate outputs were collected for 4K, 8K, 12K, and 16K iterations (K = 1000). The segmented images were fed into phase III of the system for LI and MA interface extraction and calibration.

Boundary extraction as phase III

This stage extracted the information that helped further quantify the plaque burden or cIMT. Thus from the binary region, the LI-far and MA-far borders were extracted using the LR and interadventitial segmented regions. This required refinement by following the plaque morphology whilst smoothing the borders and improving the accuracy of the DL system. The refinement used a ML-based approach that adapted the cross-validation protocol to determine accuracy. It should be noted that LI far walls and MA far walls were independent of the ML-based system and can be mathematically expressed as a regression or least squares model if GT (or ideal) boundaries are given as:

$$\boldsymbol{I}\,[2N \times P]: [x_1 \quad y_1 \quad \ldots \quad x_N \quad y_N]^T \tag{2}$$

and the raw DL borders extracted using the DL-based method are given as:

$$\boldsymbol{D}\,[2N \times P]: [a_1 \quad b_1 \quad \ldots \quad a_N \quad b_N]^T \tag{3}$$

where N represents the total number of patients and P represents the total points on the border. In the adaptation of the cross-validation protocol, the DL boundaries were divided into two sets: a training set ($\boldsymbol{D}_{tr}$) and a test set ($\boldsymbol{D}_{te}$). Correspondingly, GT boundaries were also divided into training sets ($\boldsymbol{I}_{tr}$) and test sets ($\boldsymbol{I}_{te}$). Using the linear model of least squares presented in [30], one can mathematically express this as a norm equation given as $\|\boldsymbol{I} - \boldsymbol{D}\boldsymbol{\varphi}\|^2$. Letting $\hat{\boldsymbol{\varphi}}_{tr}$ be the unknown training coefficient matrix of size $[P \times P]$, one can compute it as:

$$\hat{\boldsymbol{\varphi}}_{tr} = (\boldsymbol{D}_{tr}^T.\boldsymbol{D}_{tr})^{-1}.\boldsymbol{D}_{tr}^T.\boldsymbol{I}_{tr} \tag{4}$$

where "." represents the multiplicative product. These training coefficients were used to estimate the test boundaries ($\hat{\boldsymbol{I}}_{te}$) as the product of training coefficients and raw test DL borders using:

$$\hat{\boldsymbol{I}}_{te} = \hat{\boldsymbol{\varphi}}_{tr}.\boldsymbol{D}_{te} \tag{5}$$

Finally, the DL borders underwent cIMT measurement as presented in Appendix A. The last stage (phase IV) consisted of performance evaluation as shown in Fig. 2.

Performance evaluation as phase IV

The performance of the DL system required computation of the LI and MA far wall position errors. These values were compared against the GT to estimate the precision of merit (PoM). These calculations are shown in Appendix C. These performance metrics were then compared against other systems for benchmarking (resented in the performance evaluation section).

4. Experimental Protocol and Results

The experimental protocol primarily consisted of the optimization of DL with respect to a number of iterations independent of LI and MA wall interfaces. As there were two DL systems corresponding to two GTs, the results are presented with respect to GT1 and GT2.

4.1 Experimental protocol

In this study, K10 cross-validation (i.e., 90% training dataset and 10% testing dataset) was used for training and testing. In this cross-validation, the dataset was randomly divided into 10 parts and 10 combinations were formed from these parts. Each combination contained nine parts for training and one part for testing.

The optimization protocol was implemented for 4K, 8K, 12K, 16K, and 20K iterations (K = 1000). The iterations were evaluated for LI, MA, and cIMT errors to study their effects on the encoder and decoder (shown in Fig. 3) and their ability to smooth out the glitches against the gold standard. The LI, MA, and cIMT error values after ML-based calibration were further recorded to show the least error value that smoothed the output borders and improved the accuracy of the entire DL system.

A sample visual output of the DL-based system from phase III is shown in Fig. 4. In addition to LI, MA, and cIMT error evaluations, comprehensive clinical data analyses were also performed (i.e., correlation of age vs. cIMT, risk stratification based on the cIMT threshold, and receiver operating characteristic (ROC) analysis).

4.2 Results

The results were computed for 4K, 8K, 12K, 16K, and 20K iterations (K = 1000). The plots for error versus iteration with respect to GT1 and GT2 are shown in Fig. 5(a) and (b), respectively. The LI, MA, and cIMT error values for all iterations, including fusion and calibration, corresponding to GT1 and GT2 are presented in Tables 1 and 2, respectively. The term fusion refers to the best result among all iterations. The cIMT values in the fusion rows of Tables 1 and 2 refer to the values obtained from the best optimized LI and MA wall interfaces among all iterations. All values in the calibration rows of Tables 1 and 2 indicate the final values after ML-based calibration was applied (phase II, Fig. 2). The results indicate that the optimized result for LI error with respect to GT1 was obtained at 16K iterations (i.e., 0.135 ± 0.076 mm, which later increased marginally). The optimized result for MA error with respect to GT1 was obtained at 20K iterations (i.e., 0.171 ± 0.153 mm). The best cIMT error with respect to GT1 was computed from the fusion of 16K iterations of LI interface optimization and 20K iterations of MA interface optimization (i.e., 0.128 ± 0.124 mm.) After ML-based calibration, the LI, MA, and cIMT errors were further reduced to 0.077 ± 0.057, 0.113 ± 0.105, and 0.126 ± 0.134 mm, respectively.

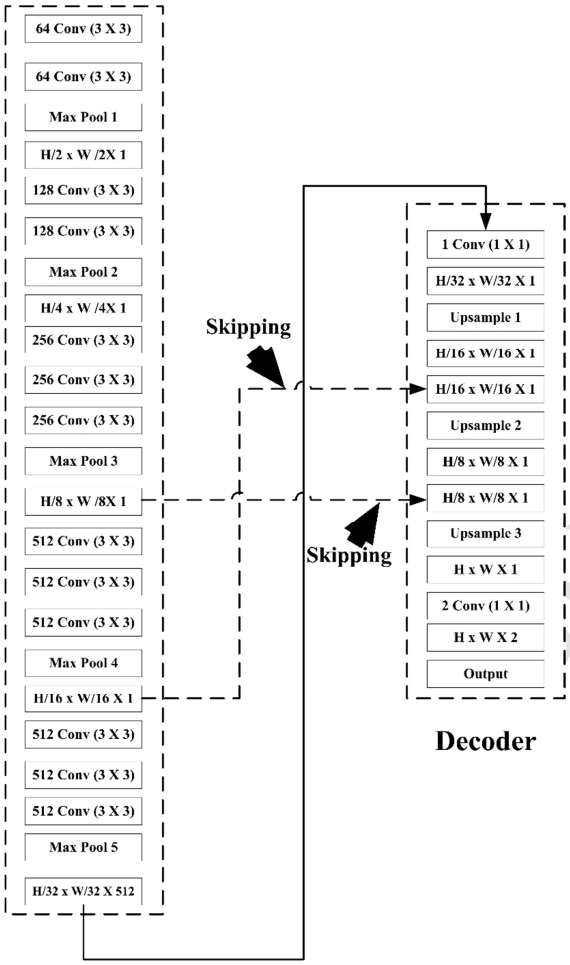


Figure 3. The combination of encoder–decoder blocks in the central DL system (a class of AtheroEdge™ system, AtheroPoint).

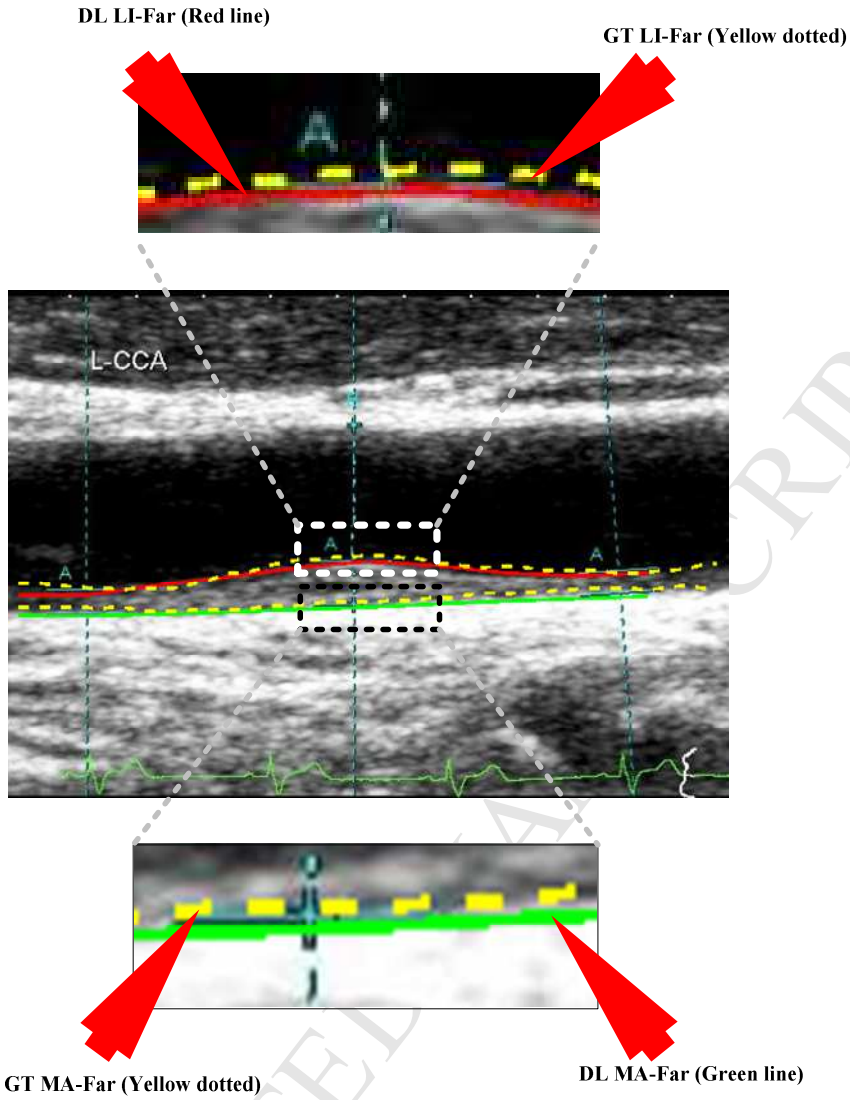


Figure 4. The DL-based system showing GT and DL outputs.

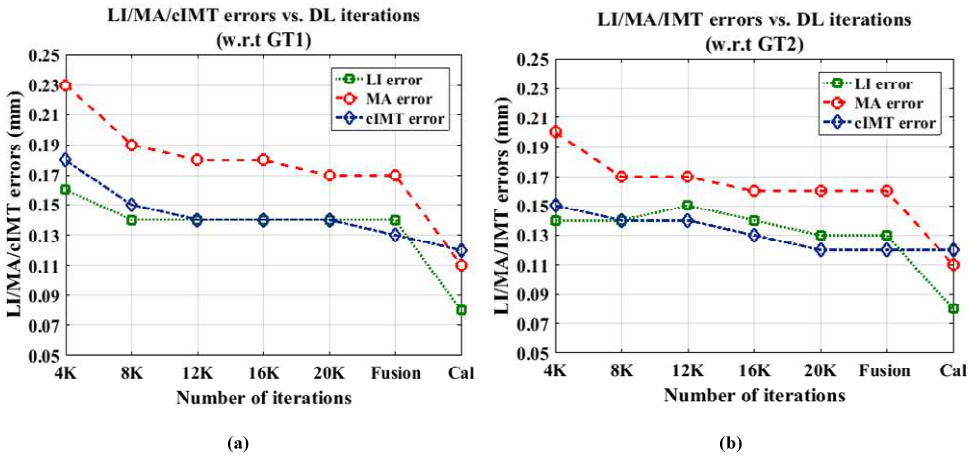


Figure 5. Plots for errors versus iterations against (a) GT1 and (b) GT2. Cal, calibration output; cIMT error (blue); LI error (green); MA error (red).

Table 1. LI, MA, and cIMT error values against GT1. Grey boxes show the optimized results for the corresponding LI, MA, and cIMT errors. *K = 1000.

DL Iterations	LI error w.r.t GT1 (mm)	MA error w.r.t GT1 (mm)	cIMT error w.r.t GT1 (mm)
4K*	0.161±0.090	0.230±0.197	0.177±0.179
8K*	0.138±0.078	0.187±0.149	0.146±0.13
12K*	0.135±0.061	0.177±0.122	0.142±0.124
16K*	0.135±0.076	0.178±0.153	0.142±0.132
20K*	0.135±0.078	0.171±0.153	0.140±0.149
Fusion	0.135±0.076	0.171±0.153	0.128±0.124
Calibrated	**0.077±0.057**	**0.113±0.105**	**0.126±0.134**

Table 2. LI, MA, and cIMT error values against GT2. Grey boxes show the optimized results for the corresponding LI, MA, and cIMT errors. *K = 1000.

DL Iterations	LI error w.r.t GT2 (mm)	MA error w.r.t GT2 (mm)	cIMT error w.r.t GT2 (mm)
4K*	0.143±0.073	0.198±0.149	0.148±0.134
8K*	0.144±0.088	0.168±0.150	0.136±0.123
12K*	0.149±0.082	0.164±0.137	0.136±0.123
16K*	0.135±0.073	0.164±0.132	0.131±0.121
20K*	0.131±0.062	0.164±0.127	0.124±0.11
Fusion	0.131±0.073	0.163±0.132	0.124±0.11
Calibrated	**0.077±0.049**	**0.109±0.088**	**0.124±0.10**

Similarly, the best results for LI and MA error optimization using the DL-based system with respect to GT2 were obtained at 20K iterations and were 0.131 ± 0.073 and 0.163 ± 0.132 mm, respectively. The cIMT error for the LI and MA interfaces was 0.124 ± 0.11 mm. After calibration, the LI, MA, and cIMT error values were further reduced to 0.077 ± 0.049, 0.109 ± 0.088, and 0.124 ± 0.10 mm, respectively.

The correlation coefficient (CC) for DL1 with respect to GT1 was 0.96 ($P < 0.0001$) and for DL2 with respect to GT2 was 0.95 ($P < 0.0001$). Therefore the CC results show a high degree of correlation between the DL outputs and the corresponding GTs. The correlation plot for DL-based system (DL1 and DL2) output with respect to GT1 and GT2 is shown in Fig. 6. The *P*-value for both plots was <0.0001, thus showing a high correlation and significance that satisfies the null hypothesis. These results prove that the DL-based system is accurate and efficient. The performance of the DL-based system is evaluated in the next section.

The results of the DL-based system with respect to GT1 and GT2 were analyzed using Bland–Altman plots. The corresponding figures with reference to GT1 and GT2 are shown in Fig. 7.

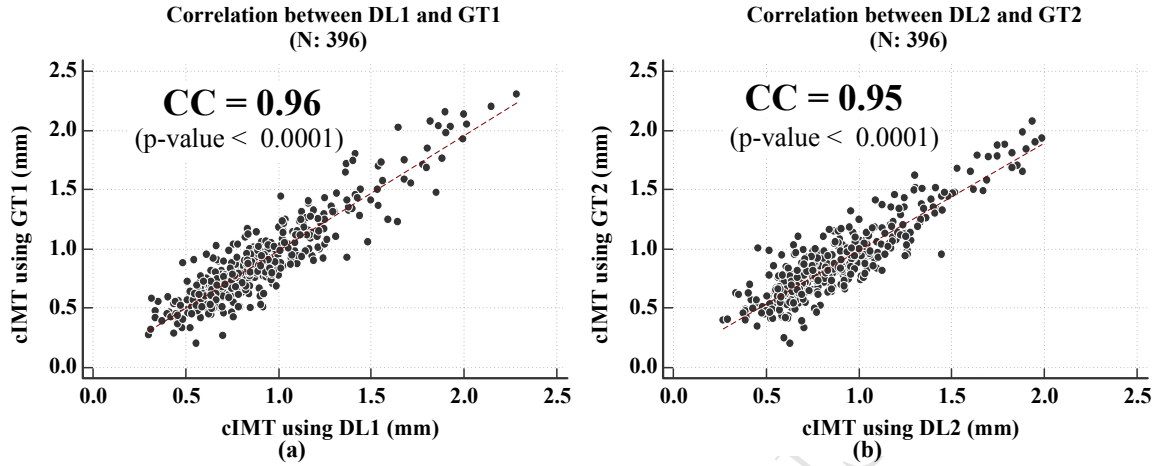


Figure 6. Correlation plots of DL-based systems against (a) against GT1 and (b) GT2.

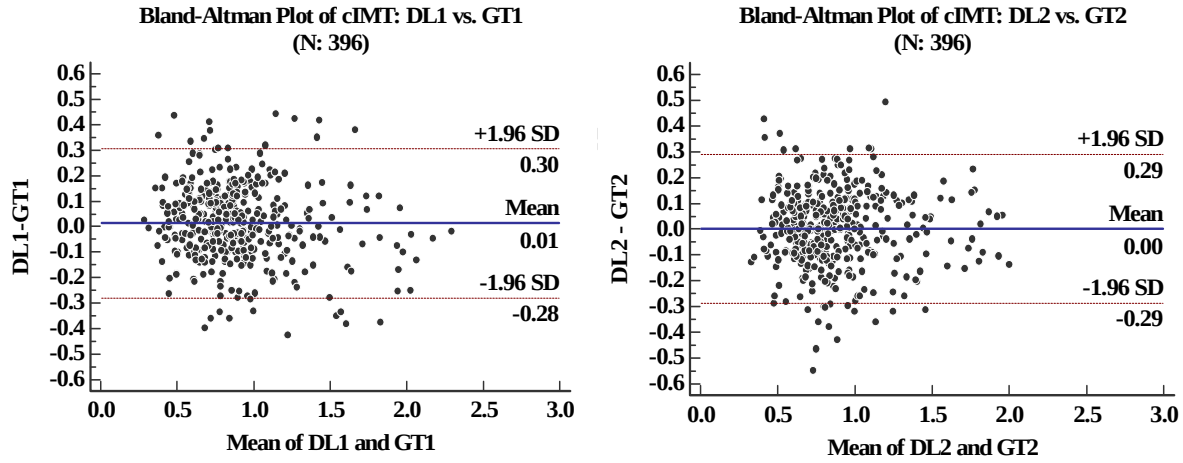


Figure 7. Bland–Altman plots of the DL-based system with reference to (a) GT1 and (b) GT2.

5. Performance of the DL Systems and Variability Analysis

Two sets of analyses were performed to evaluate the performance of the DL systems. The first set had four parts: part (i) focused on evaluating DL against manual expert tracers, part (ii) was against the sonographer's readings which were taken in real time in the US vascular laboratory, part (iii) evaluated signed and unsigned cIMT errors of the DL1 and DL2 systems, and part (iv) compared the DL system against previously-developed methods [33]. The

second set had two parts: part (a) consisted of the interoperator variability between the two DL systems (DL1 and DL2), and part (b) consisted of interobserver variability between the two GT systems (GT1 and GT2).

5.1 Comparison of DL against expert manual tracing

The cross-validation study was performed to check the effectiveness of the DL-based system when compared with other gold standards or ground truths (GTs). The correlation curves showing DL1 with respect to GT2 and DL2 with respect to GT1 are presented in Fig. 8(a) and (b), respectively. The CC values between DL1 and GT2 and DL2 and GT1 were 0.94 and 0.93, respectively, thus showing the strong interrelationship between the DL and GT. The *P*-value for both was <0.0001, which satisfies the null hypothesis. This also shows the strong statistical significance and stability of the proposed DL-based system.

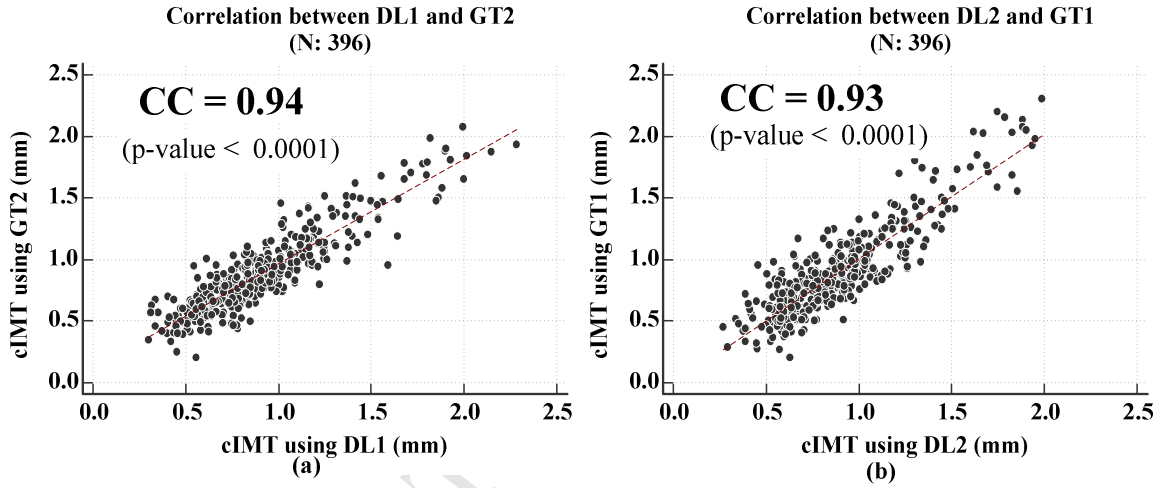


Figure 8. Correlation plots of (a) DL1 versus GT2 and (b) DL2 versus GT1.

5.2 Comparison of the DL against the sonographer's readings

This study also provided the sonographer's reading of cIMT [symbolized as Sono cIMT (ave.)]. The sonographer's reading was taken at one sample point (or one location) or two sample points (two locations) along the CCA. This reading typically consisted of the highest two plaque readings above the baseline but took into consideration the distance between LI and MA. The mean value from the two locations was computed for each image. As discussed in Section 2, of the 203 patients (396 images) in the original database, sonographer far-wall cIMT readings were only available for 193 patients (346 images). Therefore the comparison was conducted for the 346 available images. The improvements (%) in the DL results compared to the sonographer's readings are shown in Table 3. Row one (R1: CC) shows the CC between (i) the sonographer's reading and the GT reading (0.80) and (ii) DL1 and GT1

(0.96), showing an improvement of 20%. Row 2 (R2: CC) shows the CC between (i) the sonographer's reading and the GT reading (0.83) and (ii) DL2 and GT2 readings (0.95), showing an improvement of 14.5%. The correlation plot for the sonographer's cIMT readings with respect to GT1 and GT2 is shown in Fig. 9.

Table 3. Percentage improvement in DL readings compared to the sonographer's readings.

Coefficient of correlation (CC) between three kinds of cIMT (ave.) readings: sonographer (Sono), deep learning (DL1 and DL2 systems) and ground truth (GT1 and GT2)			Percentage Improvement of deep learning (DL) reading over sonographer (Sono) reading
Sono cIMT (ave.) and DL1 cIMT (ave.) against GT1 cIMT (ave.)			
Attribute	Sono vs. GT1	DL1 Vs. GT1	
R1:CC	**0.80**	**0.96**	**20.0%**
Sono cIMT (ave.) and DL2 cIMT (ave.) against GT2 cIMT (ave.)			
	Sono vs. GT2	DL2 vs. GT2	
R2: CC	**0.83**	**0.95**	**14.5%**

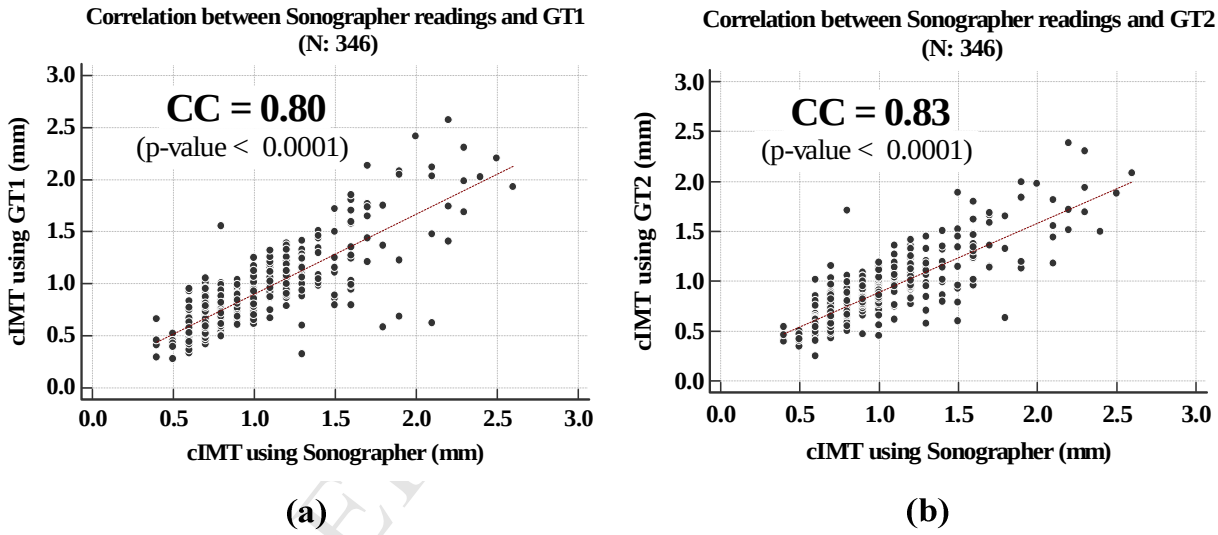


Figure 9. Correlation plots of the sonographer's cIMT readings with respect to (a) GT1 and (b) GT2.

5.3 Absolute and signed cIMT error analysis for DL1 and DL2 systems

The cumulative distribution figure plots (CDF) with respect to GT1 and GT2 are shown in Fig. 10. Fig. 10 (a) shows that 90% of patients had an absolute cIMT error <0.28 mm for GT1. The CDF plot in Fig. 10 (b) shows that 90% of patients had an absolute cIMT error <0.26 mm for GT2. The CDF plots for signed cIMT error are shown in Fig. 11. The CDF plot for signed cIMT error for GT1 indicates that 90% of patients had a signed error >−0.16 mm and 90% had a signed error <0.18 mm. Similarly for GT2, the signed cIMT error for 90% of

patients was >−0.20 mm and for 90% was <0.19 mm. This further signifies that the DL-based system performs strongly.

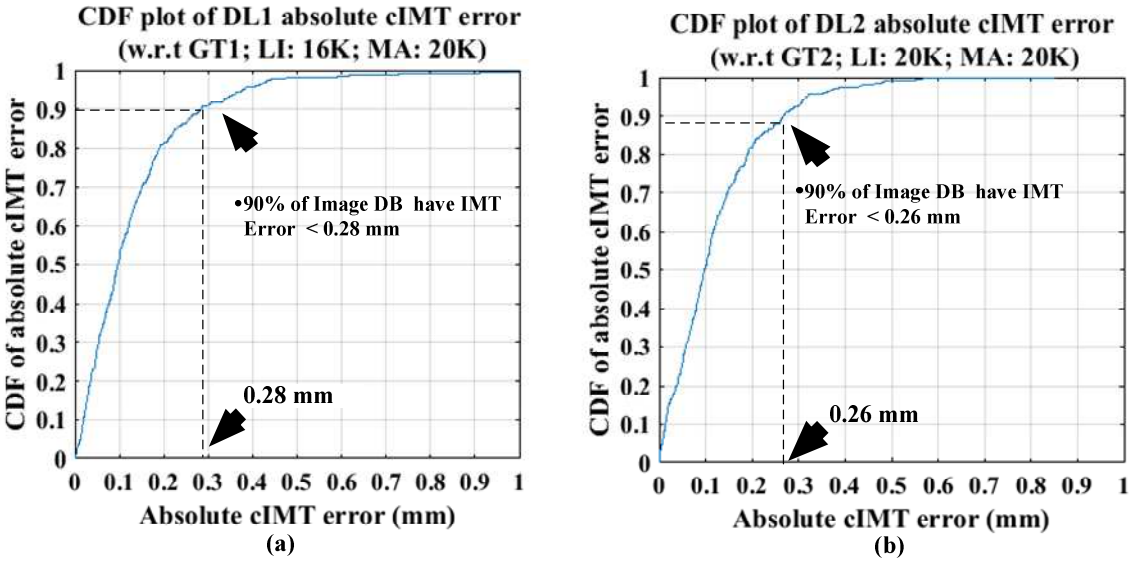


Figure 10. Absolute cIMT error for (a) DL1 and (b) DL2.

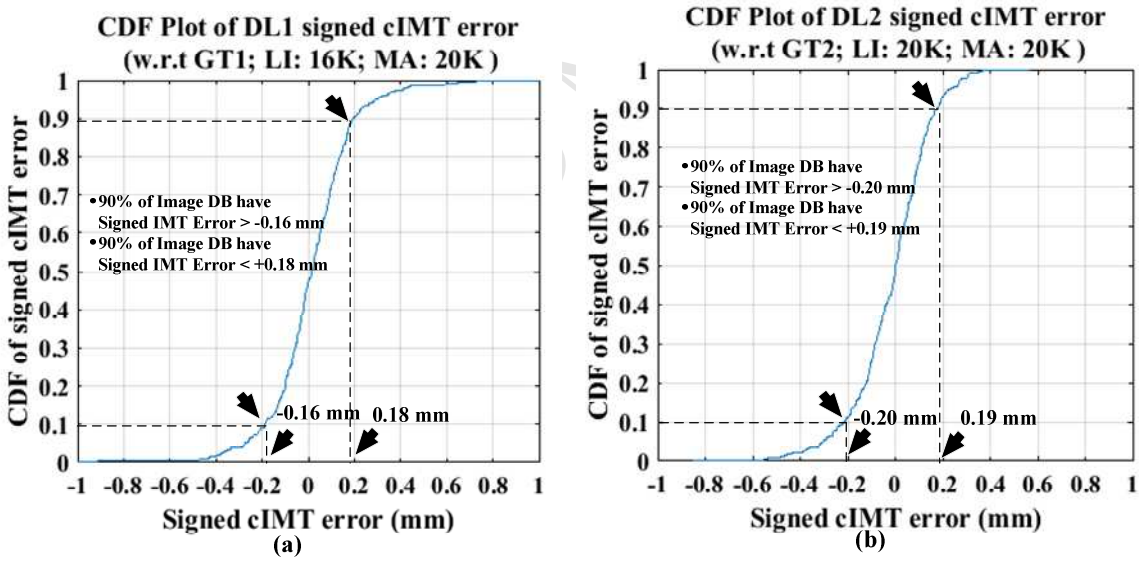


Figure 11. Signed cIMT error for (a) DL1 and (b) DL2.

5.4 DL versus previous methods

The SS method was previously implemented by Suri et al. [33]. Although the system was clinically stable, it was still compared to the DL-based strategy. A total of 360 attributes were chosen that covered the full spectrum to show the improvement of the DL strategy compared to the SS strategy. These are shown in Table 4 under column one entitled “wall characteristics,” and seven attributes were considered. The DL method used 396 images,

while the SS method used 404 images. For all attributes of the wall characteristics, the percentage improvement is shown in columns C4 and column C7 for the DL1 and DL2 systems, respectively. The lumen diameter error improvements for DL1 and DL2 were 33.2% and 39.6%, respectively. The interadventitia diameter error improvements for DL1 and DL2 were 26.7% and 28.7%, respectively. The LI-far error improvements for DL1 and DL2 were 51.9% and 63.3%, respectively. The MA-far error improvements for DL1 and DL2 were 50.9% and 58.1%, respectively. The LI-near error improvements for DL1 and DL2 were 45.5% and 52.4%, respectively. The MA-near error improvements for DL1 and DL2 were 42.6% and 38.5%, respectively. The Jaccard index (JI) for the LR improvements for DL1 and DL2 were 5.6% and 5.6%, respectively. The dice similarity (DS) for the LR improvements for DL1 and DL2 were 3.2% and 3.2%, respectively. The JI for the IAR improvements for DL1 and DL2 were 4.4% and 5.5%, respectively. Finally, the DS for the IAR improvements for DL1 and DL2 were 3.2% and 3.2%, respectively. A comparison of two images constructed using both the DL-based system and the SS system is shown in Fig. 12.

Table 4. Benchmarking of the DL-based system with regards to the SS method.** computed using >404 images.

C0	C1	C2	C3	**C4**	C5	C6	**C7**
SN	Wall Characteristics	DL1 w.r.t GT1 (mm)	SS* w.r.t GT1 (mm)	**Improv. (%)**	DL2 w.r.t GT2 (mm)	SS* w.r.t GT2 (mm)	**Improv. (%)**
1	LD error (mm)	0.167±0.181	0.25± 0.24	**33.2**	0.163±0.169	0.27±0.25	**39.6**
2	IAD error (mm)	0.176±0.167	0.24± 0.24	**26.7**	0.164±0.141	0.23 ± 0.23	**28.7**
3	LI-far error (mm)	0.077±0.057	0.16 ±0.11	**51.9**	0.077±0.049	0.21 ± 0.18	**63.3**
4	MA-far error (mm)	0.113±0.105	0.23 ±0.18	**50.9**	0.109±0.088	0.26 ± 0.15	**58.1**
5	LI-near error (mm)	0.120±0.146	0.22± 0.15	**45.5**	0.119±0.179	0.25± 0.18	**52.4**
6	MA-near error(mm)	0.132±0.147	0.23 ± 0.18	**42.6**	0.123±0.137	0.20 ± 0.17	**38.5**
7	JI (lumen region)	0.94 ± 0.03	0.89	**5.6**	0.94 ± 0.03	0.89	**5.6**
8	DS (lumen region)	0.97 ± 0.02	0.94	**3.2**	0.97 ± 0.02	0.94	**3.2**
9	JI (inter-adventitial region)	0.95 ± 0.03	0.91	**4.4**	0.96± 0.03	0.91	**5.5**
10	DS (inter-adventitial error)	0.98 ± 0.02	0.95	**3.2**	0.98±0.02	0.95	**3.2**

** computed over 404 images.

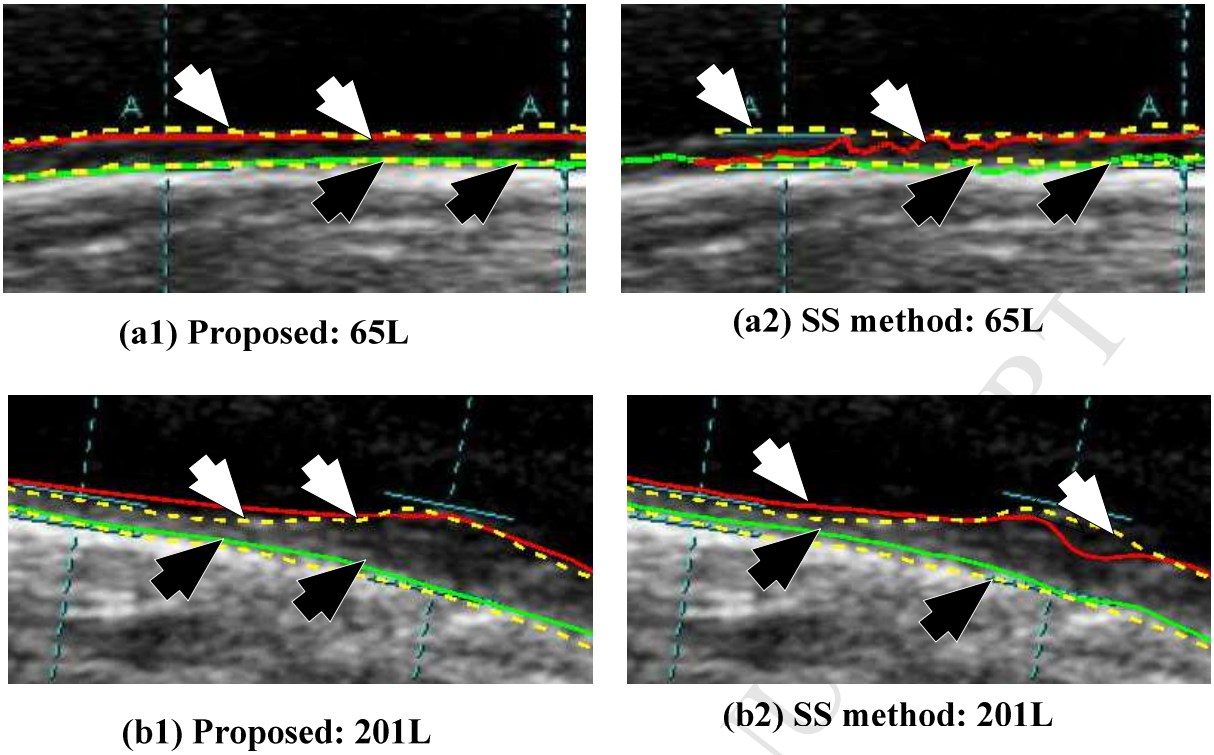


Figure 12. Application of the DL-based system and the SS system in patients 65L and 201L. The use of (a1) DL and (a2) SS in patient 65L clearly showed that the extracted borders were smoother in the former. For patient 201L, use of the DL-based method (b1) showed better accuracy than the SS system (b2).

5.5 Interoperator variability of the DL systems: DL1 and DL2

This study also compared the two DL-based systems with each other to check the reliability of the proposed DL-based system. The correlation between DL1 and DL2 is shown in Fig. 13. The correlation between DL1 and DL2 was 0.95, which indicates a strong interrelationship between DL1 and DL2. The *P*-value was <0.0001, which further satisfies the null hypothesis and implies that the DL-based system is reliable and stable.

5.6 Interobserver variability between the GT systems: GT1 and GT2

The observer readings were also compared with each other to validate that they were compatible. The correlation plot between GT1 and GT2 is shown in Fig. 14. The CC value between GT1 and GT2 was 0.97, which validates that the observer values were compatible. The *P*-value for the plot was <0.0001, which further satisfies the null hypothesis and shows that the values were highly correlated.

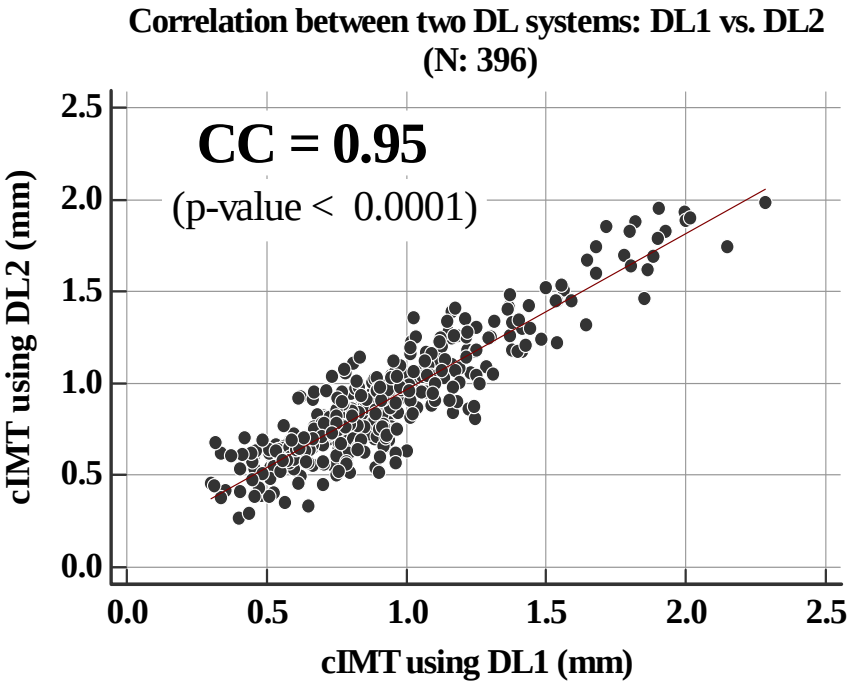


Figure 13. Correlation plot between DL1 and DL2.

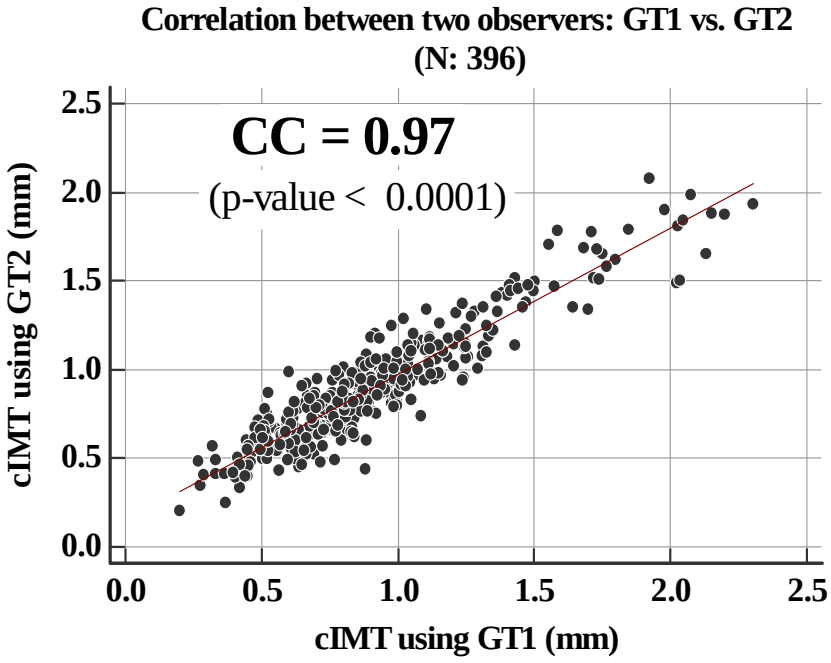


Figure 14. Correlation plot between GT1 and GT2.

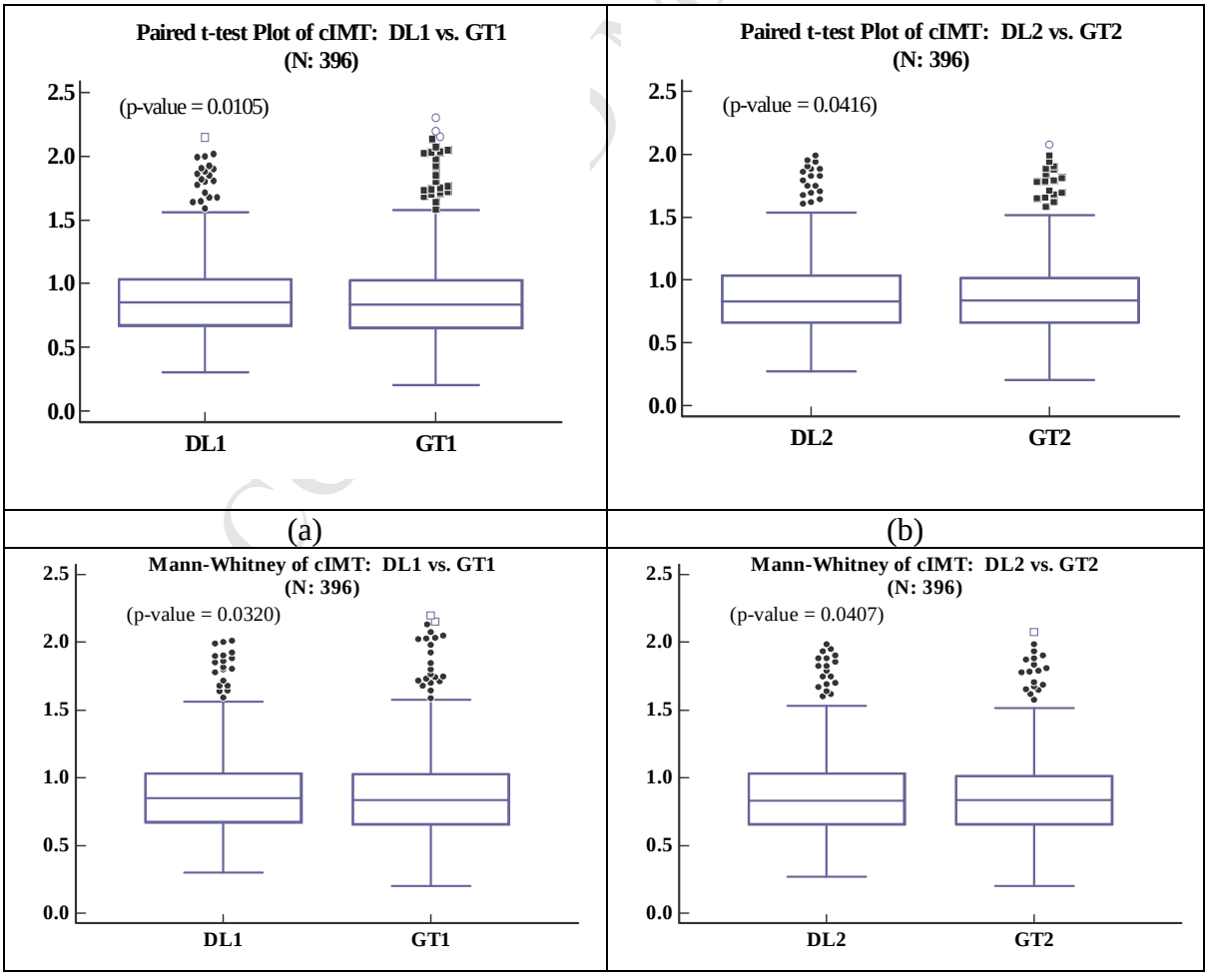
6. Statistical Tests and Risk Analysis

This section presents the four statistical tests that were used to show the significance of the proposed DL system. Risk stratification was also computed using age and risk threshold

parameters. This section also presents the ROC curves and area under the curve (AUC) analysis for the DL systems.

6.1 Four statistical tests

The outputs of the DL-based system were tested using the paired t-test, Mann–Whitney test, and Wilcoxon test, and the corresponding boxplots are shown in Fig. 15. The corresponding *P*-values for the paired t-tests of DL1 and DL2 with respect to GT1 and GT2 were 0.0105 and 0.0416, respectively. The *P*-values for the Mann–Whitney tests of DL1 and DL2 with respect to GT1 and GT2 were 0.0320 and 0.0407, respectively. Similarly, the *P*-values for the Wilcoxon test of DL1 and DL2 with respect to GT1 and GT2 were 0.0488 and 0.0348, respectively. The parameters for the paired t-test, Mann–Whitney test, and Wilcoxon test are given in Tables 5, 6, and 7, respectively. The *P*-values from all three tests were statistically significant. The Kruskal–Wallis test was also performed for DL1 and DL2, and the results are given in Table 8. The *P*-values with respect to DL1 and DL2 were 0.4905 and 0.4501, respectively. Therefore the null hypothesis that the data was taken from the same distribution was retained for DL1 and DL2.



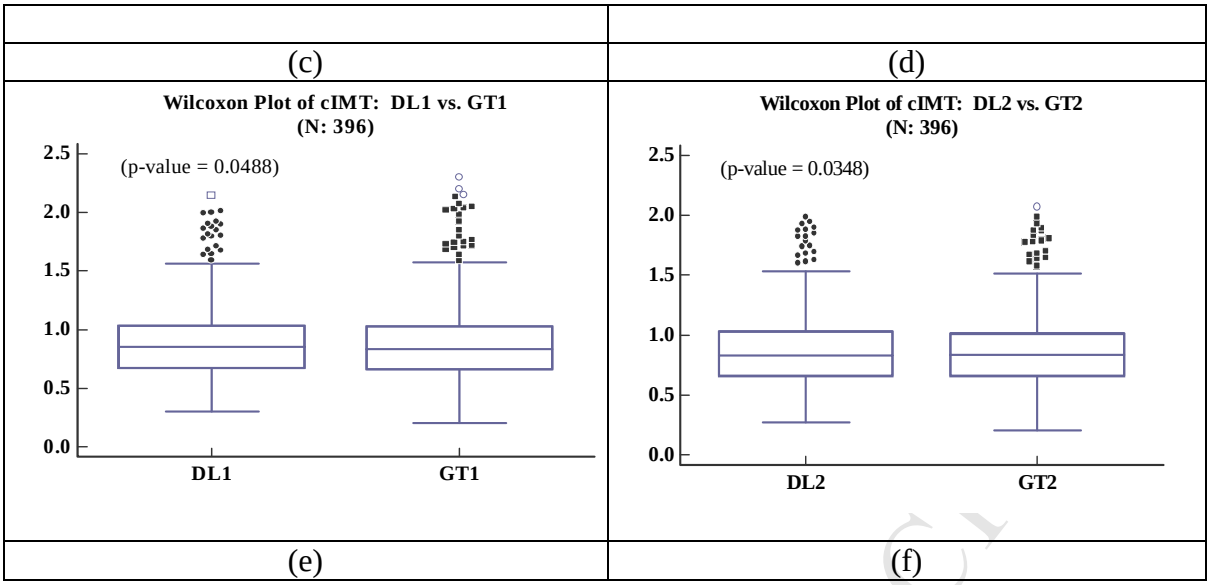


Figure 15. Statistical paired t-test with respect to (a) GT1 and (b) GT2. Mann–Whitney test with respect to (c) GT1 and (d) GT2. Wilcoxon test with respect to (e) GT1 and (f) GT2.

Table 5. Paired t-test.

Parameters	**DL1**	**DL2**
Mean difference	-0.01274	-0.001553
Standard deviation of differences	0.1490	0.1470
Standard error of mean difference	0.007489	0.007385
95% CI	-0.02747 to 0.001978	-0.01607 to 0.01297
Test statistic t	-1.702	-0.210
Degrees of Freedom (DF)	395	395
Two-tailed probability	p = 0.00105 (< 0.05)	p = 0.0416 (< 0.05)

Table 6. Mann–Whitney test.

Parameters	**DL1**	**DL2**
Average rank of first group	404.3687	404.3687
Average rank of second group	388.6313	388.6313
Mann-Whitney U	75292.00	75292.00
Large sample test statistic Z	0.968	0.968
Two-tailed probability	p= 0.03201	p = 0.0407

	(< 0.05)	(< 0.05)

Table 7. Wilcoxon test.

Parameters	DL1	DL2
Number of positive differences	186	188
Number of negative differences	210	208
Large sample test statistic Z	0.733167	1.852002
Two-tailed probability	p = 0.0488 (< 0.05)	p = 0.0348 (< 0.05)

Table 8. Kruskal–Wallis test.

Parameters	DL1 w.r.t GT1	DL2 w.r.t GT2
Test statistic	395.0000	395.0000
Corrected for ties Ht	395.0000	395.0000
Degrees of Freedom (DF)	395	395
Significance level	p = 0.490537 (> 0.05)	p = 0.4500537 (> 0.05)

6.2 Risk analysis by age

Several studies showed that cIMT increases with age [18] due to metabolic activity in the arteries [34]. The results obtained in this study were consistent with the previously-published literature. cIMT was analyzed against age (years) for the left artery, right artery, and the mean of the two carotid arteries. Table 9 shows the CC for the left, right, and combined cohort using the DL1, DL2, GT1, and GT2 systems. The number of patients in the left, right, and combined cohorts was 195, 201, and 203, respectively. Table 9 shows the positive correlation

between age and cIMT. The right carotid artery showed a higher correlation than the left; however, all patients showed a significant association between age and cIMT ($P < 0.001$).

Table 9. Comparative study of age versus cIMT for DL1 and DL2 against GT1 and GT2. The top row shows age versus DL1 and age versus GT1 for the left, right, and mean carotid arteries. The bottom row shows age versus DL2 and age versus GT2 for the left, right, and mean carotid arteries.*n = number of patients in the left, right, and combined cohorts.

Left cIMT (n=195)		**Right cIMT (n=201)**		**Mean of Left and Right cIMT (n=203)**	
Age Vs DL1 CC (p-value)	Age Vs GT1 CC (p-value)	Age Vs DL1 CC (p-value)	Age Vs GT1 CC (p-value)	Age Vs DL1 CC (p-value)	Age Vs GT1 CC (p-value)
0.20 (p<0.001)	**0.14 (p<0.001)**	**0.19 (p<0.001)**	**0.18 (p<0.001)**	**0.19 (p<0.001)**	**0.14 (p<0.001)**
Age Vs DL2 CC (p-value)	Age Vs GT2 CC (p-value)	Age Vs DL2 CC (p-value)	Age Vs GT2 CC (p-value)	Age Vs DL2 CC (p-value)	Age Vs GT2 CC (p-value)
0.18 (p<0.001)	**0.13 (p<0.001)**	**0.21 (p<0.001)**	**0.16 (p<0.001)**	**0.19 (p<0.001)**	**0.14 (p<0.001)**

*n are the number of patients for left, right and combined carotids.

6.3 Risk stratification and ROC curves

This subsection discusses the risk component of the study. Atherosclerosis screening by Bard et al. [35] suggested that patients with cIMT values >1.0 mm required more aggressive treatment; however, the population was small (95 patients) and nondiverse. A study of 7983 patients by Bots et al. [36] suggested that the risk of stroke increased when cIMT values were >0.9 mm. Other studies also stratified high-risk patients based on cIMT values > 1.0 [37] and 0.80 mm [38]. A study on 100 patients by Saba et al. [18] recommended a cIMT threshold of 0.9 mm for risk stratification.

This dataset contained a diabetic cohort of 201 patients with moderate subclinical atherosclerosis. Although 0.9 mm is recommended as the cutoff for high-risk patients, two sets of cutoffs were actually selected: 0.85 and 0.9 mm. The corresponding ROC curves with respect to these two cutoff values for both DL systems are shown in Fig. 16 (a) and (b), respectively. The AUC values for the 0.85 mm cutoff corresponding to DL1 and DL2 were 0.88 and 0.84. When the cutoff was increased to 0.9 mm, the AUC values for DL1 and DL2

were 0.88 and 0.85, respectively. This shows that 88% of the patients were correctly identified in the low–moderate and high-risk pools.

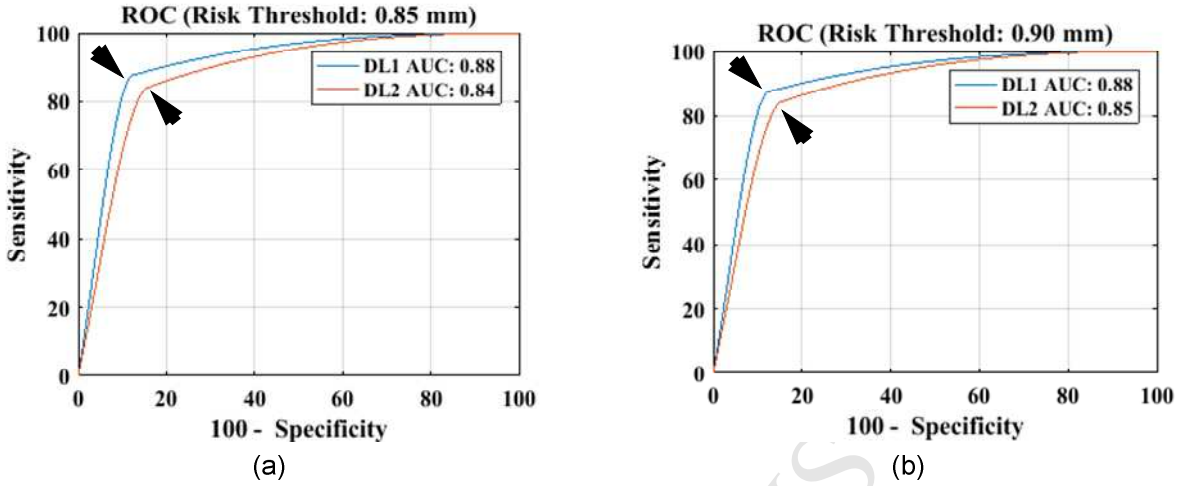


Figure 16. ROC curves for two different risk thresholds: (a) 0.85 mm (AUC values of 0.88 and 0.84 corresponding to DL1 and DL2) and (b) 0.90 mm (with AUC values of 0.88 and 0.85).

7. Discussion

This paper proposed a two-stage DL-based system implemented serially in four phases to accurately measure LI, MA, and cIMT. The DL-based system acquired preprocessed images from the first phase (i.e., multiresolution). The second phase of the entire system was stage I of the DL-based system (the heart of the DL system). The DL-based system was divided into 13 convolution layers (encoder) and three upsampling layers (decoder). These three upsampling layers belonged to the FCN. After the images were segmented, they were passed to the third phase (ML-based calibration), which represented the second stage of the DL-based system. In this phase, the LI and MA borders were extracted and calibrated using a ML-based system. The cIMT was computed from the LI and MA borders. Performance analysis was performed in phase IV. The performance results showed that the DL-based system gave better accuracy when compared to contemporary methods and was more robust and efficient. The results for different quartiles are shown in Fig. 17.

Benchmarking table

The algorithms developed for the LI, MA, and cIMT measurements are listed in the benchmarking table in Table 10. Wendelhag et al. [39] used dynamic programming for cIMT measurements. Their cIMT error was 0.030 ± 0.032 mm, which was the lowest of all the

developed techniques (Table 10; row #1); however, their dataset was limited to 69 images and the cIMT error varied widely due to different ethnicities, age groups, and nationalities. Petroudi et al. [40] used an active contour model to measure cIMT. The IMT error was 0.080 ± 0.070 mm; however, the dataset was limited to 100 patients (Table 10; row #2). Suri et al. [10] developed various techniques for IMT measurements using a larger dataset containing 344 patients. Five methods were presented, namely, CALEX 1.0, CARES, CAMES 1.0, CAUDLES, and first-order absolute moment (FOAM). FOAM showed the highest accuracy with a cIMT error of 0.150 ± 0.169 mm (Table 10; rows #3–7). Suri et al. [41] also used CALEX and CAMES for LI and MA measurements. CAMES showed the lowest LI error at 0.081 ± 0.099 mm, while the MA error was 0.082 ± 0.197 mm (Table 10; rows #8–9). The corresponding cIMT error with CALEX 2.0 and CAMES 3.0 was 0.121 ± 0.334 and 0.078 ± 0.112 mm, respectively (Table III in [41]). In 2015, Suri et al. [17] used AtheroEdge™ software for LI, MA, and cIMT measurements and achieved the lowest errors for LI, MA, and cIMT of 0.008 ± 0.099, 0.018 ± 0.013, and 0.01 ± 0.01 mm, respectively (Table 10; row #10); however, the dataset was different and contained different ethnicities. In 2016, Suri et al. [18] used AtheroCloud™ to measure LI and MA errors and achieved results of 0.065 ± 0.037 and 0.067 ± 0.036 mm, respectively (Table 10; row #11). In 2017, Suri et al. [19] used bulb-edge point detection and segmental cIMT for LI, MA, and cIMT error detection and obtained results of 0.012 ± 0.012, 0.021 ± 0.015, and 0.165 ± 0.171 mm, respectively (Table 10; row #12). This dataset also contained different ethnicities. As discussed in Subsection 5.4, Kumar et al. [33] used a diabetic cohort and achieved LI and MA errors of 0.160 ± 0.110 and 0.230 ± 0.180 mm, respectively, for GT1, and 0.210 ± 0.180 and 0.260 ± 0.150 mm, respectively, for GT2 (Table 10; rows #13–14). The same diabetic cohort was used to assess the novel DL-based system in this study, and the results showed LI and MA errors of 0.077 ± 0.057 and 0.113 ± 0.105 mm, respectively, for GT1, and 0.077 ± 0.049 and 0.109 ± 0.088 mm, respectively, for GT2. This study also reported a cIMT error of 0.126 ± 0.134 and 0.124 ± 0.10 mm for GT1 and GT2, respectively (Table 10; rows #15–28). PoM was also computed (described in Appendix C) for all experiments (Table 10; column #9, row #15–28).

Table 10. Benchmarking table.

SN	Paper	Method	#P^{+}	Data Size (N)	LI Error (mm)	MA error (mm)	cIMT Error (mm)	PoM
1	Wendelhag *et al.* [39] (1997)	*DP		69	-	-	**0.030 ± 0.032**	-
2	Petroudi *et al.* [40] (2012)	*AC	-	100	-	-	0.080 ± 0.070	-

3	Molinari *et al.* [10] (2012a)	CALEX 1.0	344	665	-	-	0.191 ± 0.217	-
4	Molinari *et al.* [10] (2012a)	CARES	344	647	-	-	0.172 ± 0.222	-
5	Molinari *et al.* [10] (2012a)	CAMES 1.0	344	657	-	-	0.154 ± 0.227	-
6	Molinari *et al.* [10] (2012a)	CAUDLES	344	630	-	-	0.224 ± 0.252	-
7	Molinari *et al.* [10] (2012a)	FOAM	344	665	-	-	0.150 ± 0.169	-
8	Molinari *et al.* [41] (2012b)	CALEX 2.0	365	365	0.088 ± 0.132	0.141 ± 0.201	0.121± 0.334	-
9	Molinari *et al.* [41] (2012b)	CAMES 3.0	365	365	0.081 ± 0.099	0.082 ± 0.197	0.078±0.112	-
10	Ikeda *et al.* [17] (2015)	AtheroEdge™	341	341	**0.008± 0.099**	**0.018± 0.013**	**0.01± 0.01**	-
11	Saba *et al.* [18] (2016)	AtheroCloud™	100	200	0.065± 0.037	0.067± 0.036	-	-
12	Ikeda *et al.* [19] (2017)	*BEP, SIMT	657	657	0.012± 0.012	0.021± 0.015	0.165 ± 0.171	-
13	Kumar *et al.* [33] (2017a)	*SS1	202	404	0.16 ±0.11	0.23 ±0.18	-	-
14	Kumar *et al.* [33] (2017a)	SS2	202	404	0.21 ± 0.18	0.26 ± 0.15	-	-
15	**Proposed**	**DL1 (4K)**	**203**	**396**	**0.161±0.090**	**0.230±0.197**	**0.177±0.179**	**94.3**
16	**Proposed**	**DL1 (8K)**	**203**	**396**	**0.138±0.078**	**0.187±0.149**	**0.146±0.13**	**94.3**
17	**Proposed**	**DL1 (12K)**	**203**	**396**	**0.135±0.061**	**0.177±0.122**	**0.142±0.124**	**92.0**
18	**Proposed**	**DL1 (16K)**	**203**	**396**	**0.135±0.076**	**0.178±0.153**	**0.142±0.132**	**99.0**
19	**Proposed**	**DL1 (20K)**	**203**	**396**	**0.135±0.078**	**0.171±0.153**	**0.140±0.149**	**98.7**
20	**Proposed**	**Fusion**	**203**	**396**	**0.135±0.076**	**0.171±0.153**	**0.128±0.124**	**97.7**
21	**Proposed**	**Calibrated**	**203**	**396**	**0.077±0.057**	**0.113±0.105**	**0.126±0.134**	**99.9**
22	**Proposed**	**DL2 (4K)**	**203**	**396**	**0.143±0.073**	**0.198±0.149**	**0.148±0.134**	**99.4**
23	**Proposed**	**DL2 (8K)**	**203**	**396**	**0.144±0.088**	**0.168±0.150**	**0.136±0.123**	**99.6**
24	**Proposed**	**DL2 (12K)**	**203**	**396**	**0.149±0.082**	**0.164±0.137**	**0.136±0.123**	**97.2**
25	**Proposed**	**DL2 (16K)**	**203**	**396**	**0.135±0.073**	**0.164±0.132**	**0.131±0.121**	**96.3**
26	**Proposed**	**DL2 (20K)**	**203**	**396**	**0.131±0.062**	**0.164±0.127**	**0.124±0.11**	**99.8**
27	**Proposed**	**Fusion**	**203**	**396**	**0.131±0.073**	**0.163±0.132**	**0.124±0.11**	**98.7**
28	**Proposed**	**Calibrated**	**203**	**396**	**0.077±0.049**	**0.109±0.088**	**0.124±0.10**	**99.9**

*AC, active contours; BEP, bulb-edge point detection; DP, dynamic programming; K, 1000 iterations; P^+, number of patients; SIMT, segmental IMT; SS, scale–space.

A short note on calibration

The ML-based calibration strategy is a regression-based method that was used to fine tune the raw DL borders to ensure smoothness. It is basically a ML-based cross-validation deformable model to regress DL-based borders from stage I closer to the actual GT borders. An independent coefficient matrix was developed from the training and GT dataset as shown in Eq. 4. A large number of patients helped to create a more generalized coefficient matrix. The predicted dataset was the product of this training-based coefficient matrix and the online test DL-based matrix. The results showed that LI, MA, and cIMT errors were reduced after the use of the ML-based calibration. The best results were obtained when this DL-based pilot

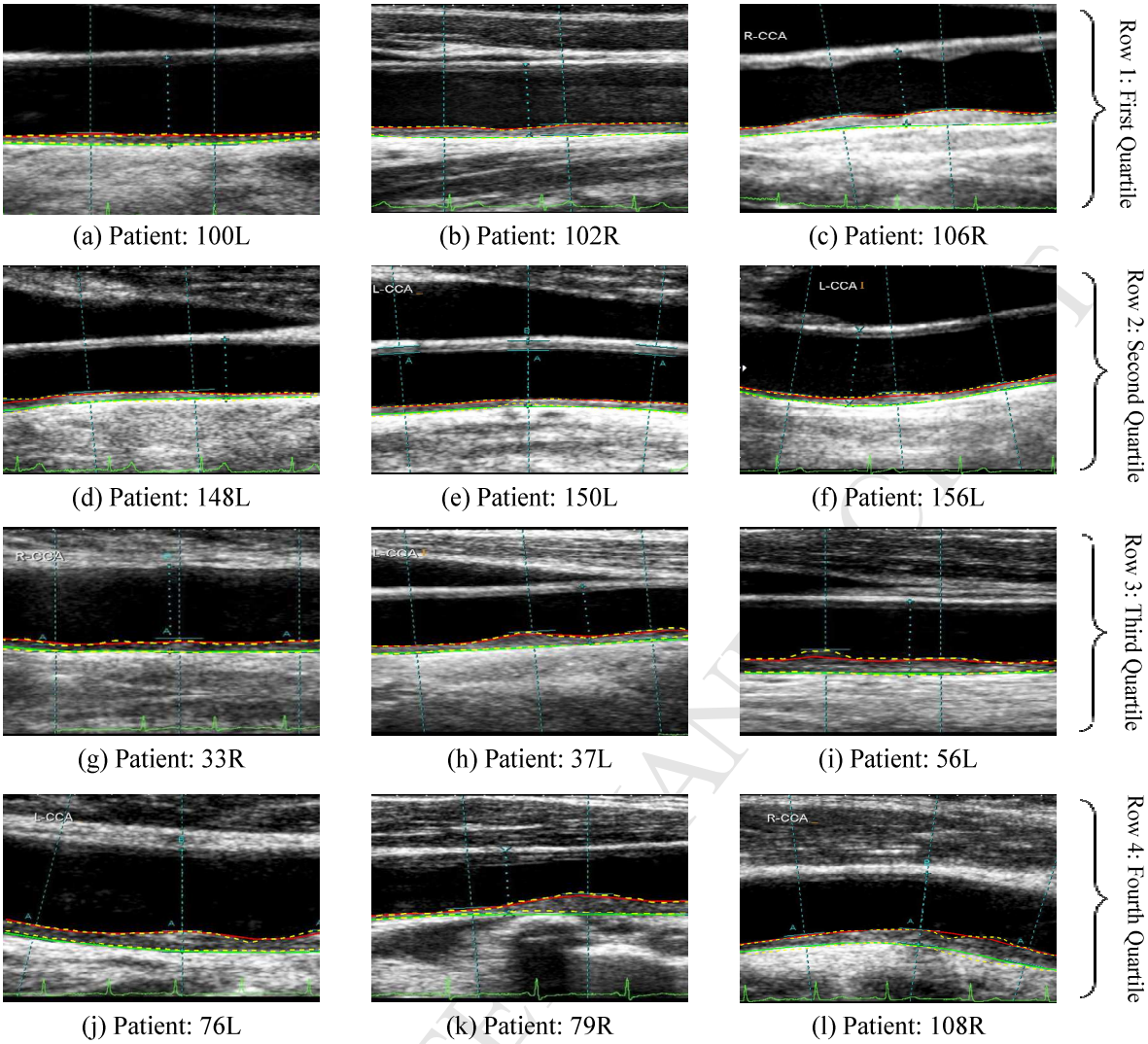
study used a jack-knifing strategy for the ML-based paradigm, where all but one instance was used for training and the remaining one was used for testing. Use of the jack-knifing strategy resulted in better accuracy for both stenotic and nonstenotic cases. Thus a strategy where ML-based calibration is cascaded with the core DL-based paradigm is stable, robust, and clinically accurate in comparison with previous methods.

A special note on DL optimization

This is the first study to employ a DL strategy for cIMT measurements. Another novelty is the use of both convolution neural network (CNN) and FCN as a combination of LI and MA segmentation. This is also the first time that a ML-based system was introduced to fine tune the raw DL-based LI and MA borders. The 13 layers of CNN extract high-level features from the CCA US images. These features were upsampled using upsampling layers of FCN, and the skipping operation was performed to obtain sharp and crisp segmented images. After extracting the LI and MA borders from these images, ML-based calibration was adapted to smooth any minor glitches in the borders. Finally, the PDM method was adapted to obtain the shortest bidirectional distance.

A special note on skip operation

There are two approaches in FCN: contraction and expansion. In the contraction approach, the features were downsampled at intermediate layers using convolution and pooling operations. In the expansion approach, the inverse convolution was applied to upsample the features. Skip operations were applied to extract features (skipping features) from the contracting layers to the intermediate layers to recover spatial information lost during the downsampling in the contraction path. This was done by merging skipping features from various resolution layers in the contracting path with input features in the expansion path. In this way, a highly accurate segmentation output was obtained from the FCN. Two skipping operations were applied in the model reported here.



(a) Patient: 100L (b) Patient: 102R (c) Patient: 106R

(d) Patient: 148L (e) Patient: 150L (f) Patient: 156L

(g) Patient: 33R (h) Patient: 37L (i) Patient: 56L

(j) Patient: 76L (k) Patient: 79R (l) Patient: 108R

Figure 17. Image overlays from the first (row #1), second (row #2), third (row #3), and fourth (row #4) quartiles. Dotted yellow lines represent GT LI-far and MA-far walls, red lines represent DL LI-far wall, and green lines represent MA-far walls.

Strengths, weaknesses, and extensions

The major strength of this DL-based system is its full automation. The accuracy of the system was comparatively higher than contemporary methods and therefore it was clinically stronger. DL is an intelligence-based system that is adapted from neural connections in the brain. This is the first time that a DL-based system was used for cIMT measurements when cascaded with a ML-based calibration, and such a cascade is truly novel. Moreover, once trained, the output from the DL-based system is produced in real time and takes a few

milliseconds. However, the dataset used was limited to a Japanese diabetic cohort, and the system has not been tested on a wide variety of datasets. Therefore the system requires further analysis in a multiethnic patient population who have subclinical atherosclerosis with low, moderate, and high-risk scenarios. Further analysis also needs to be performed using a different set of original equipment manufacturer (OEM) machines as attempted by Suri et al. [10]. Finally, this DL desktop version should be extended to a web-based version (previously developed by Suri et al. [18,42]) and undergo a reproducibility analysis, which was recently attempted by the same team [43,44].

Hardware configuration

The system was implemented on central processing unit (CPU)-based hardware (i.e., Intel icore3 2.9 GHz, 8 GB RAM); however, the results were replicated on graphics processing unit (GPU)-based settings (i.e., NVIDIA GeForce GTX with 1280 cores and 5 GB memory).

8. Conclusion

This study presents a novel, robust, and clinically-viable solution to cIMT measurements using an AtheroEdge™ system from AtheroPoint™. The system uses an intelligence-based paradigm for cIMT measurement by employing the DL strategy for the segmentation of LR and IAR. To fine tune this, the system adapts a ML-based joint coefficient method for final border extraction for the far wall of the carotid artery. Data are prepared in a multiresolution paradigm which reduces the computational burden. The polyline distance method, which is a standard used in the industry, is adapted for all measurements. The system performs better than previous studies. For example, the LI position error improved by 52% and 63%, and the MA position error improved by 51% and 58%. The cIMT error for DL1 and DL2 was 0.126 ± 0.134 and 0.124 ± 0.10 mm, respectively. The CC between age and cIMT was 0.20, and the AUC had an upper bound close to 90%. The DL-based system can be adapted for clinical settings or multicentre pharmaceutical trial modes, just like the AtheroEdge™ or AtheroCloud™.

Acknowledgment

The authors at the National Institute of Technology, Goa, India, would like to acknowledge MediaLab Asia, Ministry of Electronics and Information Technology, and the Government of India for their kind support.

References

1. Lloyd-Jones, D., Adams, R.J., Brown, T.M., Carnethon, M., Dai, S., De Simone, G., Ferguson, T.B., Ford, E., Furie, K., Gillespie, C. and Go, A., Heart disease and stroke statistics-2010 update: a report from the American Heart Association. *Circulation*, 121(7) (2010): e46.
2. Available Online. *http://www.who.int/cardiovascular_diseases.*
3. Libby, Peter, YONG-JIAN GENG, Galina K. Sukhova, Daniel I. Simon, and Richard T. Lee. Molecular determinants of atherosclerotic plaque vulnerability. *Annals of the New York Academy of Sciences* 811 (1) (1997): 134-145.
4. O'Leary, Daniel H., Joseph F. Polak, Richard A. Kronmal, Teri A. Manolio, Gregory L. Burke, and Sidney K. Wolfson Jr, . Carotid-artery intima and media thickness as a risk factor for myocardial infarction and stroke in older adults. *New England Journal of Medicine*, 340 (1) (1999): 14-22.
5. Bots, M.L.,. Carotid intima-media thickness as a surrogate marker for cardiovascular disease in intervention studies. *Current medical research and opinion*, *22*(11) (2006): 2181-2190.
6. Nambi, Vijay, Lloyd Chambless, Aaron R. Folsom, Max He, Yijuan Hu, Tom Mosley, Kelly Volcik, Eric Boerwinkle, and Christie M. Ballantyne,. Carotid intima-media thickness and presence or absence of plaque improves prediction of coronary heart disease risk: the ARIC (Atherosclerosis Risk in Communities) study. *Journal of the American College of Cardiology*, 55 (15) (2010): 1600-1607.
7. Meuwese, Marijn C., Eric de Groot, Raphaël Duivenvoorden, Mieke D. Trip, Leiv Ose, Frans J. Maritz, Dick CG Basart,. ACAT inhibition and progression of carotid atherosclerosis in patients with familial hypercholesterolemia: the CAPTIVATE randomized trial. *Jama*, 301(11) (2009): 1131-1139.
8. Ikeda, N., L. Saba, Filippo Molinari, M. Piga, K. Meiburger, K. Sugi, M. Porcu *et al.* Automated carotid intima-media thickness and its link for prediction of SYNTAX score in Japanese coronary artery disease patients. *International angiology: a journal of the International Union of Angiology*, 32(3) (2013): 339-348.

9. Molinari, Filippo, Guang Zeng, and Jasjit S. Suri. Intima-media thickness: setting a standard for a completely automated method of ultrasound measurement. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 57 (5) (2010).

10. Molinari, Filippo, Kristen M. Meiburger, Luca Saba, U. Rajendra Acharya, Giuseppe Ledda, Guang Zeng, Sin Yee Stella Ho *et al*. Ultrasound IMT measurement on a multi-ethnic and multi-institutional database: our review and experience using four fully automated and one semi-automated methods. *Computer methods and programs in biomedicine*, 108 (3) (2012): 946-960.
11. Molinari, Filippo, Guang Zeng, and Jasjit S. Suri. A state of the art review on intima–media thickness (IMT) measurement and wall segmentation techniques for carotid ultrasound. *Computer methods and programs in biomedicine*, 100 (3) (2010): 201-221.
12. Molinari, F., Acharya, U.R., Zeng, G., Meiburger, K.M. and Suri, J.S., 2011. Completely automated robust edge snapper for carotid ultrasound IMT measurement on a multi-institutional database of 300 images. *Medical & biological engineering & computing*, 49 (8) (2011): 935-945.
13. Molinari, Filippo, Constantinos S. Pattichis, Guang Zeng, Luca Saba, U. Rajendra Acharya, Roberto Sanfilippo, Andrew Nicolaides, and Jasjit S. Suri. Completely automated multiresolution edge snapper—a new technique for an accurate carotid ultrasound IMT measurement: clinical validation and benchmarking on a multi-institutional database. *IEEE Transactions on image processing* 21 (3) (2012): 1211-1222.
14. Molinari, Filippo, Kristen M. Meiburger, Guang Zeng, Andrew Nicolaides, and Jasjit S. Suri. CAUDLES-EF: carotid automated ultrasound double line extraction system using edge flow. *Journal of digital imaging*, 24(6) (2011): 1059-1077.
15. Londhe, Narendra D., and Jasjit S. Suri. Superharmonic Imaging for Medical Ultrasound: a Review. *Journal of medical systems*, 40 (12) (2016): 279.
16. Saba, Luca, Filippo Molinari, Kristen M. Meiburger, U. Rajendra Acharya, Andrew Nicolaides, and Jasjit S. Suri. Inter-and intra-observer variability analysis of completely automated cIMT measurement software (AtheroEdge™) and its benchmarking against commercial ultrasound scanner and expert Readers. *Computers in biology and medicine*, 43(9) (2013): 1261-1272.
17. Ikeda, Nobutaka, Ajay Gupta, Nilanjan Dey, Soumyo Bose, Shoaib Shafique, Tadashi Arak, Elisa Cuadrado Godia *et al*. Improved correlation between carotid and coronary

atherosclerosis SYNTAX score using automated ultrasound carotid bulb plaque IMT measurement. *Ultrasound in medicine & biology*, 41(5) (2015): 1247-1262.

18. Saba, Luca, Sumit K. Banchhor, Harman S. Suri, Narendra D. Londhe, Tadashi Araki, Nobutaka Ikeda, Klaudija Viskovic *et al*. Accurate cloud-based smart IMT measurement, its validation and stroke risk stratification in carotid ultrasound: A web-based point-of-care tool for multicenter clinical trial. *Computers in biology and medicine*, 75 (2016): 217-234.
19. Ikeda, Nobutaka, Nilanjan Dey, Aditya Sharma, Ajay Gupta, Soumyo Bose, Suvojit Acharjee, Shoaib Shafique *et al*. Automated segmental-IMT measurement in thin/thick plaque with bulb presence in carotid ultrasound from multiple scanners: Stroke risk assessment. *Computer Methods and Programs in Biomedicine*, 141 (2017): 73-81.
20. Acharya, U.R., Mookiah, M.R.K., Sree, S.V., Yanti, R., Martis, R.J., Saba, L., Molinari, F., Guerriero, S. and Suri, J.S.,. Evolutionary algorithm-based classifier parameter tuning for automatic ovarian cancer tissue characterization and classification. *Ultraschall in der Medizin-European Journal of Ultrasound*, 35(03) (2014): 237-245.
21. Acharya, U.R., Sree, S.V., Kulshreshtha, S., Molinari, F., Koh, J.E.W., Saba, L. and Suri, J.S.,. GyneScan: an improved online paradigm for screening of ovarian cancer via tissue characterization. *Technology in cancer research & treatment*, *13*(6) (2014): 529-539.
22. Pareek, G., Acharya, U.R., Sree, S.V., Swapna, G., Yantri, R., Martis, R.J., Saba, L., Krishnamurthi, G., Mallarini, G., El-Baz, A. and Ekish, S.A.. Prostate tissue characterization/classification in 144 patient population using wavelet and higher order spectra features from transrectal ultrasound images. *Technology in cancer research & treatment*, *12*(6) (2013): 545-557.
23. Shrivastava, V.K., Londhe, N.D., Sonawane, R.S. and Suri, J.S.,. Reliable and accurate psoriasis disease classification in dermatology images using comprehensive feature space in machine learning paradigm. *Expert Systems with Applications*, 42(15) (2015): 6184-6195.
24. Accurate Diabetes Risk Stratification using Machine Learning: Role of Missing value and Outliers, Md. Maniruzzaman, Md. Jahanur Rahman, Md. Al-MehediHasan, Harman S. Suri4, Md. Menhazul Abedin, Ayman El-Baz, Jasjit S. Suri, *Journal of Medical Systems*, 2018.
25. LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton,. Deep learning. *Nature*, 521(7553) (2015): 436-444.

26. Teichmann, Marvin, Michael Weber, Marius Zoellner, Roberto Cipolla, and Raquel Urtasun,. MultiNet: Real-time Joint Semantic Reasoning for Autonomous Driving. *arXiv preprint arXiv:1612.07695*, (2016).
27. Molinari, Filippo, William Liboni, Pierangela Giustetto, Sergio Badalamenti, and Jasjit S. Suri. Automatic computer-based tracings (ACT) in longitudinal 2-D ultrasound images using different scanners. *Journal of Mechanics in Medicine and Biology*, 9(4) (2009): 481-505.
28. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton,. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, (2012).
29. Long, Jonathan, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2015): 3431-3440.
30. Molinari, Filippo, Kristen M. Meiburger, Luca Saba, Guang Zeng, U. Rajendra Acharya, Mario Ledda, Andrew Nicolaides, and Jasjit S. Suri. Fully Automated Dual-Snake Formulation for Carotid Intima-Media Thickness Measurement. *Journal of Ultrasound in Medicine*, 31 (7) (2012): 1123-1136.
31. Simonyan, K., & Zisserman, A.. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
32. J.S. Suri, R.M. Haralick, F.H. Sheehan. Greedy algorithm for error correction in automatically produced boundaries from low contrast ventriculograms, *Pattern Anal. Appl*., 3 (1) (2000): 39-60.
33. Kumar, P.K., Araki, T., Rajan, J., Saba, L., Lavra, F., Ikeda, N., Sharma, A.M., Shafique, S., Nicolaides, A., Laird, J.R. and Gupta, A.. Accurate lumen diameter measurement in curved vessels in carotid ultrasound: an iterative scale-space and spatial transformation approach. *Medical & biological engineering & computing*, 55 (8) (2017): 1415-1434.
34. Suri, Jasjit S., Chirinjeev Kathuria, and Filippo Molinari, eds. Atherosclerosis disease management. *Springer Science & Business Media*, 2010.
35. Bard, Robert L., Henna Kalsi, Melvyn Rubenfire, Thomas Wakefield, Beverly Fex, Sanjay Rajagopalan, and Robert D. Brook. Effect of carotid atherosclerosis screening on risk stratification during primary cardiovascular disease prevention. *American Journal of Cardiology*, 93 (8) (2004): 1030-1032.

36. Bots, M.L., Hoes, A.W., Koudstaal, P.J., Hofman, A. and Grobbee, D.E. Common Carotid Intima-Media Thickness and Risk of Stroke and Myocardial Infarction: The Rotterdam Study. *Circulation*, 96(5) (1997): 1432-1437.
37. Araki, Tadashi, Nobutaka Ikeda, Nilanjan Dey, Suvojit Acharjee, Filippo Molinari, Luca Saba, Elisa Cuadrado Godia, Andrew Nicolaides, and Jasjit S. Suri. Shape-Based Approach for Coronary Calcium Lesion Volume Measurement on Intravascular Ultrasound Imaging and Its Association With Carotid Intima-Media Thickness. *Journal of Ultrasound in Medicine*, 34 (3) (2015): 469-482.
38. Kao, Amy H., Apinya Lertratanakul, Jennifer R. Elliott, Abdus Sattar, Linda Santelices, Penny Shaw, Mehret Birru *et al*. Relation of carotid intima-media thickness and plaque with incident cardiovascular events in women with systemic lupus erythematosus. *American Journal of Cardiology*, 112 (7) (2013): 1025-1032.
39. Wendelhag I, Liang Q, Gustavsson T, Wikstrand J. A new automated computerized analysing system simplifies reading and reduces the variability in ultrasound measurement of intima media thickness. *Stroke*, 28 (1997):2195–2200
40. Petroudi, Styliani, Christos Loizou, Marios Pantziaris, and Constantinos Pattichis. Segmentation of the common carotid intima-media complex in ultrasound images using active contours. *IEEE transactions on biomedical engineering*, 59(11) (2012): 3060-3069.
41. Molinari, Filippo, Constantinos S. Pattichis, Guang Zeng, Luca Saba, U. Rajendra Acharya, Roberto Sanfilippo, Andrew Nicolaides, and Jasjit S. Suri. Completely automated multiresolution edge snapper—a new technique for an accurate carotid ultrasound IMT measurement: clinical validation and benchmarking on a multi-institutional database. *IEEE Transactions on image processing*, 21(3) (2012): 1211-1222.
42. Saba, Luca, Sumit K. Banchhor, Narendra D. Londhe, Tadashi Araki, John R. Laird, Ajay Gupta, Andrew Nicolaides, and Jasjit S. Suri. Web-based accurate measurements of carotid lumen diameter and stenosis severity: An ultrasound-based clinical tool for stroke risk assessment during multicenter clinical trials. *Computers in biology and medicine*, 91 (2017): 306-317.
43. Luca Saba, Sumit K Banchhor, Tadashi Araki, Harman S Suri, Narendra D Londhe, John R Laird, Klaudija Viskovic, Jasjit S Suri. Intra- and Inter-operator Reproducibility Analysis of Automated Cloud-based Carotid Intima Media Thickness Ultrasound Measurement. *Journal of Clinical and Diagnostic Research*, 12(2) (2018): KC01-KC11.
44. Saba, Luca, Sumit K. Banchhor, Tadashi Araki, Klaudija Viskovic, Narendra D. Londhe, John R. Laird, Harman S. Suri, and Jasjit S. Suri. Intra-and inter-operator reproducibility

of automated cloud-based carotid lumen diameter ultrasound measurement. *Indian Heart Journal* (2018).

Appendix A

Polyline Distance Method

Polyline distance metric

The Polyline Distance Metric (PDM) [32] is used to measure cIMT between LI- and MA interfaces, LI-error between deep learning LI-far and ground truth LI-far interfaces, and MA-error between deep learning MA-far and ground truth MA-far interfaces. The PDM computation is given as follows: Let the first and second interfaces be denoted as C_1 and C_2. Let the reference point on C_1 be vertex P_1 and the segment in C_2 be defined by vertices P_2 and P_3. Let the distance between P_1 and P_2 be d_1 and the distance between P_1 and P_3 be denoted as d_2. Let $D(P_1, L)$ be the polyline distance between vertex P_1: (x_1, y_1) on C_1 and line segment L formed by two points P_2: (x_2, y_2) and P_3: (x_3, y_3). Let delta (δ) be the distance of the reference point, P_1 towards the line segment L. The perpendicular distance between the line segment L and the reference point, P_1, is given by d_p. Then, the polyline distance $D(P_1,\ L)$ can be defined as:

$$D(P_1, L) = \begin{cases} |d_p| & 0 < \delta < 1 \\ min(d_1, d_2) & \delta < 0, \delta > 1 \end{cases} \quad \text{(A.1)}$$

where,

$$d_1 = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad \text{(A.2)}$$

$$d_2 = \sqrt{(x_1 - x_3)^2 + (y_1 - y_3)^2} \quad \text{(A.3)}$$

$$\delta = \frac{(y_3 - y_2)(y_1 - y_2) + (x_3 - x_2)(x_1 - x_2)}{(x_3 - x_2)^2 + (y_3 - y_2)^2} \quad \text{(A.4)}$$

and

$$d_p = \frac{(y_3 - y_2)(x_2 - x_1) + (x_3 - x_2)(y_1 - y_2)}{\sqrt{(x_3 - x_2)^2 + (y_3 - y_2)^2}} \quad \text{(A.5)}$$

The process to obtain $D(P_1, L)$ is repeated for the rest of the points of the contour C_j and is given by:

$$D(C_1, C_2) = \sum_{i=1}^{N} D(P_i, S_{C_2}) \quad \text{(A.6)}$$

where, N is the total number of points on C_1 and S_{C_2} is the segment on contour C_2. This algorithm is repeated in reverse, where C_2 becomes the reference contour and C_1 becomes the segment contour. The reverse is represented as $D(C_2, C_1)$. Finally, by combining both $D(C_1, C_2)$ and $D(C_2, C_1)$, we obtain the **PDM** which is given by:

$$D_{PDM}(C_1 : C_2) = \frac{D(C_1, C_2) + D(C_2, C_1)}{(\#\ points \in C_1 + \#\ points \in C_2)} \quad \text{(A.7)}$$

Appendix B

Encoder and Decoder Network

Encoder and Decoder

The Convolution Neural Networks have the ability to decompose images into feature maps generating like a deck of cards representing the feature maps which can then be fed into limited layered neural networks for training. Mathematically, a basic convolution can be represented as:

$$\mathrm{d(x,y)} = \mathrm{I(x,y)} \otimes \mathrm{w(x,y)} = \sum_{s=-\frac{m}{2}}^{\frac{m}{2}} \sum_{t=-\frac{m}{2}}^{\frac{m}{2}} \mathrm{I(x+s, y+t)} \times \mathrm{w(x,y)} \quad \text{(B.1)}$$

where the image I is convolved with kernel w, yielding an output d, $\otimes$ represents the convolution operation. The convolution is basically a sum of all products between image I and kernel w, represented by Eq. (B.1), where the kernel is represented as a vector of size $m \times m$ and is shown for the point locations (x, y), while s and t are the dummy variables. The pooling reduces the dimensionality of each feature map but retaining the most important information i.e., max pooling and average pooling. Pooling is done to simplify the output from CNN.

In the architecture given in Fig. 3, for encoder, we have used 13 convolution layers. Each convolution layer M (=64, 128, 256, 512) kernels where each kernel is represented as a vector of size 3 x3. Small kernels allow large depth without increasing memory requirement. There are intermediate five max-pool layers to downsample the feature maps which are later concatenated and fed into next stage. In the decoder, the reverse happens. The input deck is up-sampled to original size using up-sample layers with the help of skip operations to get the segmentation output.

Appendix C

LI/MA Position Errors, cIMT Errors and Precision-of-Merit

LI Error

The LI error ($\epsilon_{LI}(i)$) for patient i is computed as the PDM between the GT LI-far wall ($LI_{far}^{gt}(i)$) and DL LI-far ($LI_{far}^{dl}(i)$) wall for the patient, which is given by:

$$\epsilon_{LI}(i) = D_{PDM}(LI_{far}^{gt}(i): LI_{far}^{dl}(i)) \quad \text{(C.1)}$$

If $\epsilon_{LI}(i)$ represents the LI error for the patient i, then, the mean LI error ($\overline{\epsilon}_{LI}$) for all N patients is given by:

$$\overline{\epsilon}_{LI} = \frac{\sum_{i=1}^{N} \epsilon_{LI}(i)}{N} \quad \text{(C.2)}$$

MA Error

Similarly, the MA error ($\epsilon_{MA}(i)$) is computed as the PDM between the GT MA-far wall ($MA_{far}^{gt}(i)$) and DL MA-far ($MA_{far}^{dl}(i)$) wall for patient i is given by:

$$\epsilon_{MA}(i) = D_{PDM}(MA_{far}^{gt}(i): MA_{far}^{dl}(i)) \quad \text{(C.3)}$$

The mean MA error ($\overline{\epsilon}_{MA}$) for all N patients is given by:

$$\overline{\epsilon}_{MA} = \frac{\sum_{i=1}^{N} \epsilon_{MA}(i)}{N} \quad \text{(C.4)}$$

cIMT Error

The cIMT error ($\epsilon_{cIMT}(i)$) for patient i is computed as the PDM between the ground truth cIMT ($cIMT_{gt}(i)$) and deep learning cIMT ($cIMT_{dl}(i)$)wall for the patient. The $cIMT_{gt}(i)$ for patient i is computed as the PDM between GT LI-far wall ($LI_{far}^{gt}(i)$) and GT MA-far wall ($MA_{far}^{gt}(i)$) which is given as:

$$cIMT_{gt}(i) = D_{PDM}(LI_{far}^{gt}(i): MA_{far}^{gt}(i)) \quad \text{(C.5)}$$

Similarly, the $cIMT_{dl}(i)$ is computed as the PDM between DL LI-far wall ($LI_{far}^{dl}(i)$) and DL MA-far wall ($MA_{far}^{dl}(i)$) which is given as:

$$cIMT_{dl}(i) = D_{PDM}(LI_{far}^{dl}(i): MA_{far}^{dl}(i)) \quad \text{(C.6)}$$

Therefore, the cIMT error ($\epsilon_{cIMT}(i)$) for patient i is computed as absolute difference between $cIMT_{gt}(i)$ and $cIMT_{dl}(i)$.

$$\epsilon_{cIMT}(i) = \left|cIMT_{gt}(i) - cIMT_{dl}(i)\right| \quad \text{(C.7)}$$

If $\epsilon_{cIMT}(i)$ signifies the cIMT error for the patient i, then, the mean cIMT error ($\overline{\epsilon}_{cIMT}$) for all N patients is given by:

$$\overline{\epsilon}_{cIMT} = \frac{\sum_{i=1}^{N} \epsilon_{cIMT}(i)}{N} \quad \text{(C.8)}$$

Precision-of-Merit (PoM)

Using Equations (B.1) and (B.2), one can, therefore, define mathematically the precision-of-merit (PoM) and is given as:

$$PoM_{cIMT}(\%) = 100 - \left(\frac{\sum_{i=1}^{N} \frac{\left|cIMT_{dl(i)} - cIMT_{gt(i)}\right|}{cIMT_{gt(i)}}}{N}\right) \times 100 \quad \text{(C.9)}$$

All the symbols are discussed in Appendix D: Table D.

Appendix D

Table D: Symbol table.

SN.	Symbol	Abbreviation
1	β_1	Predicted output
2	β_2	Ground truth
3	L	Total number of classes
4	N	Total number of images
5	θ	Loss function
6	$\boldsymbol{I}$	Ground truth boundaries
7	$\boldsymbol{D}$	Predicted DL boundaries
8	m	Total number of boundary points
9	tr	Training symbol
10	te	Testing symbol
11	$\widehat{\boldsymbol{\varphi}}_{tr}$	Estimated coefficient matrix using training data
12	C_1	First interface
13	C_2	Second interface
14	P_1	Reference point on C_1

15	P_2	Reference point on C_2
16	P_3	Reference point on C_2
17	L	Line segment formed by vertex P_1 and vertex P_2 on C_2
18	d_1	Euclidean distance between vertex P_1 and vertex P_2
19	d_2	Euclidean distance between vertex P_1 and vertex P_3
20	δ	Distance of the reference point P_1 and the line segment, L
21	d_p	Perpendicular distance between L and the reference point P_1
22	$D(P_1, L)$	Polyline distance between reference point P_1 and the line segment, L
23	$D(C_1, C_2)$	Mean polyline distance between all points on contour C_1 with respect to contour C_2
24	$D(C_2, C_1)$	Mean Polyline distance between all points on contour C_2 with respect to contour C_1
25	D_{PDM}	Bidirectional polyline distance metric by combining $D(C_1, C_2)$ and $D(C_2, C_1)$
26	$LI_{far}^{gt}(i)$	LI-far interface or contour taken from ground truth for patient i
27	$LI_{far}^{dl}(i)$	LI-near interface or contour taken from deep learning for patient i
28	$MA_{far}^{gt}(i)$	MA-far interface or contour taken from ground truth for patient i
29	$MA_{far}^{dl}(i)$	MA-far interface or contour taken from deep learning for patient i
30	$\epsilon_{LI}(i)$	Absolute LI error for patient i
31	$\overline{\epsilon}_{LI}$	Mean LI error for N patients
32	$\epsilon_{MA}(i)$	Absolute MA error for patient i
33	$\overline{\epsilon}_{MA}$	Mean MA error for N patients
34	$cIMT_{gt}(i)$	PDM between GT LI-far wall and GT MA-far interfaces for patient i
35	$cIMT_{dl}(i)$	PDM between DL LI-far wall and DL MA-far interfaces for patient i
36	$\epsilon_{cIMT}(i)$	Absolute cIMT error for patient i
37	$\overline{\epsilon}_{cIMT}$	Mean absolute cIMT error for N patients
38	PoM_{cIMT}	Precision-of-Merit for cIMT