Exploring the Impact of Learning Paradigms on Network Generalization: A Multi-Center IMT Study

Francesco Marzola, Kristen M. Meiburger*, Filippo Molinari, Massimo Salvi
Biolab, PoliTo[BIO]Med Lab, Department of Electronics and Telecommunications
Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Turin, Italy

***Abstract*— The intima-media thickness (IMT) is an important parameter for evaluating cardiovascular disease risk and progression and can be extracted from B-mode longitudinal ultrasound images of the carotid artery. Despite its clinical significance, inter- and intra-operator variability in IMT measurement is a challenge due to subjective factors. Therefore, automatic and semi-automatic approaches based on heuristic methods and deep neural networks have been proposed to reduce the variability in IMT measurement. However, the inter- and intra- operator variability still remains an issue as it affects the quality and diversity of ground truth (GT) data used for training deep learning models. In this study, the authors evaluate the performance of different learning paradigms using different GTs on a multi-center IMT dataset. A recent segmentation network, ConvNeXt, is trained on a dataset of 2576 B-mode longitudinal ultrasound images of the carotid artery, using different GT annotations and learning paradigms. The method is then tested on an external dataset of 448 images from four different centers for which three manual segmentations were available. The results show how the use of different GT annotations and learning paradigms can enhance the generalization ability of deep learning models, demonstrating the importance of selecting appropriate GT data and learning strategies in achieving robust and reliable solutions. The study highlights the significance of incorporating heuristic methods in the training process of deep learning models to enhance the accuracy and consistency of IMT measurement, thus enabling more precise cardiovascular disease risk assessment.**

Keywords*— *Artificial intelligence; intima-media thickness; segmentation networks; supervised learning; ultrasound;

I. Introduction

The intima-media thickness (IMT) measurement in B-mode ultrasound images is a non-invasive method used to assess the thickness of the intima and media layers of arterial walls [1], [2]. This technique has become increasingly important in clinical practice as it has been shown to provide valuable information about cardiovascular disease risk and progression. However, despite its clinical significance, there is a considerable amount of inter- and intra-operator variability in IMT measurement, which can be influenced by a variety of factors such as ultrasound device settings, operator experience, patient characteristics, and more, and numerous studies have also focused on formal methods for evaluating operator variability [3]–[6].

Completely automatic and semi-automatic approaches based on heuristic methods (e.g., snakes) have been proposed to reduce inter- and intra-operator variability in IMT measurement [7]–[13]. These methods extract objective features and improve the accuracy and consistency of measurement, minimizing the influence of subjective factors. Moreover, deep neural networks have recently shown significant promise in this domain, with state-of-the-art segmentation networks achieving high levels of accuracy and efficiency [14]–[16]. Despite significant advances in computer science and Artificial Intelligence (AI), the issue of inter-operator variability in medical image segmentation still remains a challenge. This variability results in discrepancies between different ground truth (GT) annotations, affecting the quality and diversity of the GT data used for training deep learning models. Therefore, selecting appropriate GT data for training segmentation networks is critical to achieving accurate and reliable medical image analysis [17].

In this study, we evaluate the performance of various learning paradigms using different ground truths (GTs). The GTs were generated using a range of techniques, including manual segmentation by different operators, a semi-automated segmentation algorithm, and two consensus methods, STAPLE [18] and one based on a similarity coefficient [17].

The main contributions of this work are the following:

- We trained a recent segmentation network using different ground truths (GTs) on a multi-center IMT dataset. The selected network, ConvNeXt, features a hierarchical design that shares similarities with vision transformers but relies exclusively on convolutional layers. This architecture is composed of multiple convolutional blocks, where each block employs grouped convolutions and a series of split-transform-merge operations to enhance feature interactions across channels and spatial dimensions.
- Our study provides a quantitative comparison that demonstrates how the use of different ground truth (GT) annotations and learning paradigms can enhance the generalization ability of deep learning models. Our findings highlight the importance of selecting appropriate GT data and learning strategies in achieving robust and reliable medical image analysis.

979-8-3503-3959-8/23/$31.00

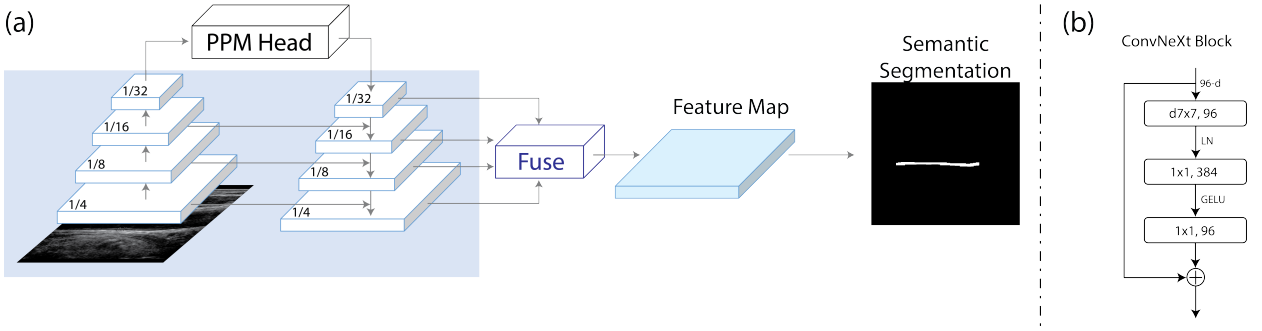

Fig. 1. Segmentation network used in this work. (a) The PPM (Pyramid Pooling Module) allow to exploit global context information by fusing features into four different pyramid scales. The feature map is then used to perform the segmentation of the intima media complex. (b) Block design of the ConvNeXt. The GELU (Gaussian Error Linear Unit) replace the traditional RELU (Rectified Linear Unit) in the convolutional block.

II. Materials and Methods

A. Dataset and Ground Truth Definition

To train and validate our model, we utilized two previously published datasets [1], [14], which are freely available for download. The resulting dataset comprised 2576 B-mode longitudinal ultrasound images of the carotid artery, with acquisition details provided in the original publications. To test different learning paradigms, we employed two manual annotations of the LI and MA profiles from the same expert analyst at two time points (A1 and A1s), along with one computerized measurement based on dynamic programming [19], which is also freely available for download [1], [14]. The dynamic programming method was developed by researchers from Technische Universität München and is referred to here as TUM for simplicity. Then, two consensus methods were employed: the well-established STAPLE method [18] and a recently proposed method based on the computation of a similarity index [17].

To evaluate the effectiveness of our proposed method, we tested it on an external dataset comprising 448 images from four different centers. For this dataset, we compared the automatic segmentation with three manual tracings, reported here as GT1, GT2 and GT3 [20].

B. Segmentation Network

The ConvNeXt network architecture (Fig. 1) is a state-of-the-art deep learning model that features a hierarchical design inspired by vision transformers. However, unlike traditional transformers, ConvNeXt exclusively utilizes convolutional layers, making it highly suitable for medical image analysis tasks. At its core, ConvNeXt consists of multiple convolutional blocks, each of which comprises multiple convolutional layers, grouped convolutions, and a series of split-transform-merge operations. These blocks are designed to capture and process high-level features from medical images with complex structures and variations, enabling the network to achieve superior performance in various medical imaging tasks.

In this segmentation network the traditional ReLU activation function is replaced with the GeLU (Gaussian Error Linear Unit) activation function. GeLU has shown promise in improving the performance of deep learning models in various image processing tasks by incorporating a non-monotonic behavior that captures both positive and negative input values. By leveraging this activation function in our segmentation network, we aim to enhance its accuracy and robustness in handling medical image data with complex structures and varying intensities.

To train the deep learning model for semantic segmentation, we randomly divided the 2576 images of the dataset into training and validation sets, containing 2311 and 265 images, respectively. The ConvNeXt was trained for 30 epochs, with an early stopping equal to 5 epochs. We used focal loss as the loss function and employed AdamW optimization algorithm with an initial learning rate of 10^{-4}, and a batch size of 8. On the training set we applied on-the-fly data augmentation with the following transformation: horizontal flips (with probability 0.25), blurring (with probability 0.25), and photometric distortions that changed the relative contrast (between 0.90 and 1.10) and saturation range (between 0.90 and 1.10). We selected the best model based on the Intersection over Union (IoU) metric on the validation set. Our approach was implemented using Pytorch and the mmsegmentation library [21]. Each segmentation model was trained on an RTX 3090 GPU with 24 GB of VRAM, taking approximately 3 hours of training time.

C. Validation metrics

To validate the results obtained, four different metrics were employed. First of all, the segmentation masks were compared using the Dice coefficient to determine how similar the manual and automatic masks were. Then, the absolute IMT error was computed as follows:

$$Abs.\ IMT\ Bias = |IMT_{method} - IMT_{Ground\ truth}| \quad (1)$$

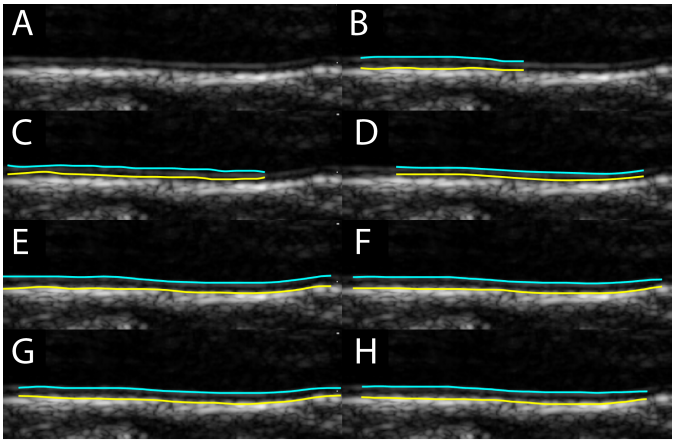

Fig. 2. Segmentation results. The LI profile is shown in cyan, and the MA profile in yellow. A) original image; B) GT1; C) GT2; D) GT3; E) $ConvNeXt_{TUM}$; F) $ConvNeXt_{STAPLE}$; G) $ConvNeXt_{A1}$; h) $ConvNeXt_{A1s}$.

where *method* refers to the automatic segmentation and *Ground truth* is the IMT measured from the three manual operators (GT1, GT2 or GT3, respectively). Finally, the Hausdorff Distance (HD) between the manual and automatic profiles were compared, considering separately the LI and MA profile (HD LI and HD MA, respectively).

III. Results

Fig. 2 shows some qualitative segmentation results obtained by the different networks. Table 1 instead reports the obtained quantitative results comparing the three manual segmentations GT1, GT2 and GT3 on the test set. The most significant metrics reveal that the network trained on TUM segmentations ($ConvNeXt_{TUM}$) consistently performs within the inter-operator variability range, even if it is not the top-performing network overall. This is in contrast to the other methods, which often perform outside of the inter-operator variability range.

It is important to note how, overall, the systems that performed the best on the external test set were those that included in some manner the TUM semi-automatic algorithm, either entirely (i.e., $ConvNeXt_{TUM}$) or through a consensus method (i.e., $ConvNeXt_{STAPLE}$ and $ConvNeXt_{HYBRID}$). In general, these three networks performed similarly when considering the Dice coefficient, but show an increase in performance when considering the absolute IMT error and the Hausdorff distance, especially considering the MA profile. This hints at the fact that the TUM semi-automatic algorithm provides essential information for accurately segmenting the MA border.

Fig. 3 presents bar plots of the results obtained by the network trained on the TUM semi-automatic algorithm compared with inter- and intra- operator variability on the test set, considering the GT1, GT2, and GT3 manual segmentations.

IV. Discussion and Conclusions

The accurate and reliable measurement of the IMT in B-mode ultrasound images is crucial in assessing cardiovascular disease risk and progression, yet the high inter- and intra-operator variability can pose significant challenges in achieving reliable and reproducible results. Despite significant advances in deep learning techniques in recent years, the issue of inter- and intra- operator variability in medical image segmentation remains an open challenge. This variability causes inconsistencies in ground truth annotations, which affects the quality and diversity of the data used for training deep learning models. Hence, choosing appropriate GT data is essential for reliable and accurate medical image analysis.

In this study, we explored the impact of various learning paradigms for deep learning networks using different GT techniques, including manual segmentations by multiple operators, semi-automated segmentation, and the consensus of multiple operators obtained using two methods. A well-established deep learning network was trained with the various learning paradigms to evaluate the different performance on an external test set that had been segmented by three different operators. The results indicate that the network trained on the semi-automatic TUM algorithm performed similarly to other

TABLE I. Performance metrics of segmentation models on the test set compared with the manual operators (GT1, GT2, and GT3). Metrics are computed on the common support between binary masks.

Metric	Segmentation models vs GT1					Between manual operators	
	ConvNeXt$_{A1}$	***ConvNeXt$_{A1s}$***	***ConvNeXt$_{TUM}$***	***ConvNeXt$_{STAPLE}$***	***ConvNeXt$_{HYBRID}$***	***GT1 vs GT2***	***GT1 vs GT3***
DSC	0.868±0.051	0.825±0.068	**0.874±0.051**	**0.876±0.045**	**0.873±0.049**	0.864±0.062	0.868±0.061
ABS IMT bias (mm)	0.121±0.104	0.261±0.126	**0.101±0.133**	0.118±0.095	0.153±0.097	0.120±0.106	0.083±0.091
HD LI (mm)	0.198±0.139	0.219±0.141	**0.186±0.145**	**0.177±0.087**	**0.179±0.109**	0.223±0.093	0.183±0.083
HD MA (mm)	0.283±0.132	0.362±0.145	**0.214±0.115**	0.256±0.119	0.268±0.126	0.195±0.139	0.200±0.108
	Segmentation models vs GT2					***GT2 vs GT1***	***GT2 vs GT3***
DSC	0.879±0.069	0.864±0.073	**0.894±0.060**	**0.896±0.062**	**0.894±0.060**	0.864±0.062	0.858±0.057
ABS IMT bias (mm)	**0.091±0.137**	0.200±0.153	**0.092±0.130**	**0.085±0.121**	0.107±0.29	0.120±0.106	0.131±0.094
HD LI (mm)	0.231±0.136	**0.185±0.131**	0.194±0.137	0.196±0.087	**0.190±0.096**	0.223±0.093	0.217±0.093
HD MA (mm)	0.283±0.198	0.362±0.199	**0.205±0.161**	0.257±0.194	0.269±0.205	0.195±0.139	0.204±0.167
	Segmentation models vs GT3					***GT3 vs GT1***	***GT3 vs GT2***
DSC	0.852±0.066	0.806±0.073	**0.870±0.052**	**0.862±0.057**	**0.858±0.059**	0.868±0.061	0.858±0.057
ABS IMT bias (mm)	0.145±0.101	0.298±0.119	**0.121±0.118**	0.142±0.090	0.185±0.099	0.083±0.091	0.131±0.094
HD LI (mm)	0.191±0.127	0.206±0.134	**0.178±0.137**	**0.169±0.076**	**0.171±0.101**	0.183±0.083	0.217±0.093
HD MA (mm)	0.301±0.136	0.382±0.148	**0.232±0.129**	0.275±0.126	0.285±0.128	0.200±0.108	0.204±0.167

GT: Ground truth; GT1, GT2, GT3: manual operators 1, 2, and 3 that segmented the test set, respectively; DSC: Dice Similarity Coefficient; ABS IMT bias: Absolute IMT bias; HD LI: Hausdorff distance between automatic and manual LI profiles; HD MA: Hausdorff distance between automatic and manual MA profiles; *ConvNeXt$_{A1}$*: network trained with A1 manual profiles as GT; *ConvNeXt$_{A1s}$*: network trained with A1 manual profiles as GT; *ConvNeXt$_{TUM}$*: network trained with TUM semi-automatic profiles as GT; *ConvNeXt$_{STAPLE}$*: network trained with the STAPLE consensus as GT; *ConvNeXt$_{HYBRID}$*: network trained with the hybrid consensus as GT; *ConvNeXt$_{A1}$*.

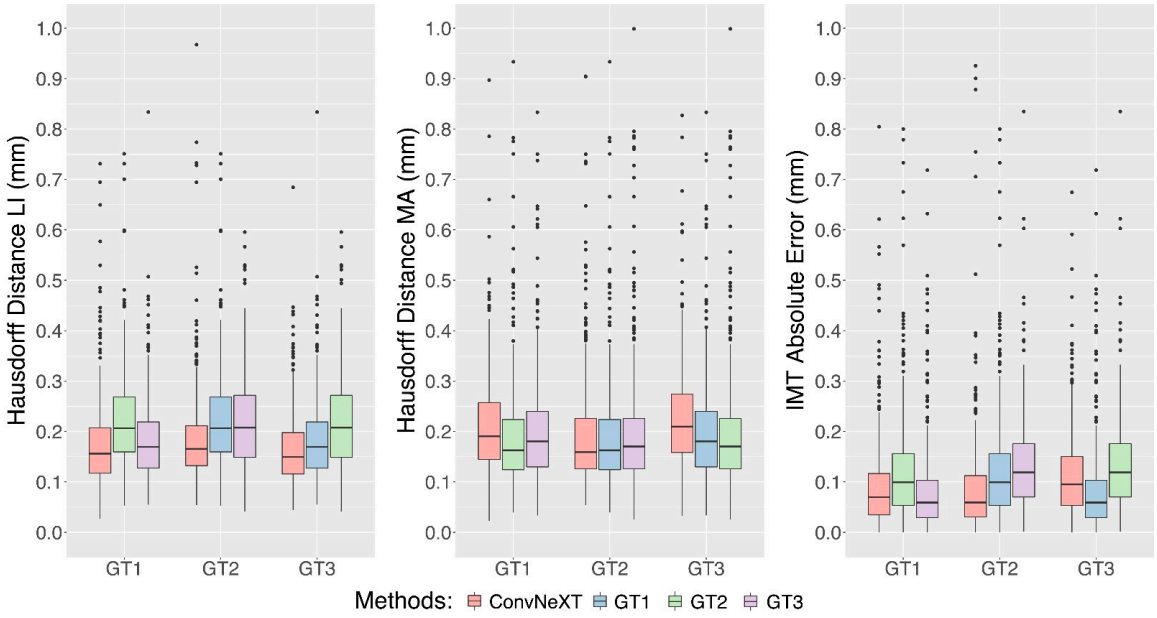

Fig. 3. Box plots comparing the quantitative results obtained by the network trained on the TUM semi-automatic segmentations (ConvNeXT) with inter-operator variability. GT1, GT2, GT3: results obtained using the manual segmentations of operators 1, 2, and 3 on the test set,

operators. While all networks showed comparable results in terms of the Dice coefficient, only some were reliable in terms of HD and absolute IMT bias, with the networks trained on STAPLE and TUM performing better. Further analysis showed that the network trained on TUM had the lowest HD value on both LI and MA profiles and, in some cases, obtained lower values than inter-operator variability (HD LI vs. GT3: 0.178±0.137 mm for TUM vs. 0.183±0.083 for GT1 and 0.217±0.093 for GT2). This indicates that the TUM-trained network was able to generalize the segmentation task better than the other networks.

The superior performance of the TUM network can be attributed to its ability to extract objective image features, resulting in better pixel-based performance compared to manual operator annotations, producing more consistent LI and MA profiles. The results here demonstrate the advantages of using a deep network trained on a GT obtained by a quantitative algorithm, which leads to better generalization compared to using single manual annotations or consensus methods. These findings are consistent with recent studies [22] in which deep learning generative algorithms are employed to overcome the limitations of heuristic algorithms, turning an initial semi-automatic, slow, and unstable heuristic algorithm into an automatic, near real-time, and robust solution.

Future developments could include the implementation of novel losses that focus on the common trait between manual and automatic segmentation, computed only on the common support. This would allow the network to better recognize the similarities between the two and improve segmentation accuracy. Additionally, exploring other deep network architectures, such as transformers, can provide insight into the effect of different types of ground truths. Applying this approach to other fields of medical imaging, such as nuclei instance segmentation in digital pathology or brain tumor segmentation in MRI, can lead to more reliable, generalizable, and robust algorithms. By continuing to develop and refine these techniques, we can improve medical image analysis and ultimately provide better patient care.

REFERENCES

[1] K. M. Meiburger *et al.*, "Carotid Ultrasound Boundary Study (CUBS): An Open Multicenter Analysis of Computerized Intima–Media Thickness Measurement Systems and Their Clinical Impact," *Ultrasound Med Biol*, vol. 47, no. 8, pp. 2442–2455, 2021, doi: 10.1016/j.ultrasmedbio.2021.03.022.

[2] M. W. Lorenz, H. S. Markus, M. L. Bots, M. Rosvall, and M. Sitzer, "Prediction of clinical cardiovascular events with carotid intima-media thickness: a systematic review and meta-analysis.," *Circulation*, vol. 115, no. 4, pp. 459–67, Jan. 2007, doi: 10.1161/CIRCULATIONAHA.106.628875.

[3] M. W. Lorenz *et al.*, "Predictive value for cardiovascular events of common carotid intima media thickness and its rate of change in individuals at high cardiovascular risk – Results from the PROG-IMT collaboration," *PLoS One*, vol. 13, no. 4, p. e0191172, Apr. 2018, doi: 10.1371/journal.pone.0191172.

[4] E. Bianchini *et al.*, "Functional and Structural Alterations of Large Arteries: Methodological Issues," *Curr Pharm Des*, vol. 19, no. 13, pp. 2390–2400, 2013, doi: 10.2174/1381612811319130007.

[5] J. A. C. Delaney *et al.*, "Effect of inter-reader variability on outcomes in studies using carotid intima media thickness quantified by carotid ultrasonography," *Eur J Epidemiol*, vol. 25, no. 6, pp. 385–392, Jun. 2010, doi: 10.1007/S10654-010-9442-8/TABLES/4.

[6] R. Tang *et al.*, "Baseline reproducibility of B-mode ultrasonic measurement of carotid artery intima–media thickness: the European Lacidipine Study on Atherosclerosis (ELSA)," *J Hypertens*, vol. 18, no. 2, pp. 197–201, 2000, Accessed: Apr. 28, 2023. [Online]. Available: https://journals.lww.com/jhypertension/Fulltext/2000/18020/Risk_factors_associated_with_alterations_in.00010.aspx

[7] K. M. Meiburger, U. R. Acharya, and F. Molinari, "Automated localization and segmentation techniques for B-mode ultrasound images: A review," *Comput Biol Med*, vol. 92, pp. 210–235, Jan. 2018, doi: 10.1016/j.compbiomed.2017.11.018.

[8] F. Molinari *et al.*, "Fully Automated Dual-Snake Formulation for Carotid Intima-Media Thickness Measurement," *J Ultrasound Med*, vol. 31, pp. 1123–1136, 2012.

[9] G. Zahnd, M. Orkisz, E. E. Dávila Serrano, and D. Vray, "CAROLAB – A platform to analyze carotid ultrasound data," in *IEEE International Ultrasonics Symposium (IUS), Glasgow (Scotland)*, 2019, pp. 463–466. doi: 10.1109/ULTSYM.2019.8925673.

[10] C. P. Loizou, "A review of ultrasound common carotid artery image and video segmentation techniques," *Med Biol Eng Comput*, vol. 52, no. 12, pp. 1073–1093, Dec. 2014, doi: 10.1007/s11517-014-1203-5.

[11] C. P. Loizou, C. S. Pattichis, A. N. Nicolaides, and M. Pantziaris, "Manual and automated media and intima thickness measurements of the common carotid artery," *IEEE Trans Ultrason Ferroelectr Freq Control*, vol. 56, no. 5, pp. 983–994, May 2009, doi: 10.1109/TUFFC.2009.1130.

[12] C. P. Loizou, C. S. Pattichis, M. Pantziaris, T. Tyllis, and A. Nicolaides, "Snakes based segmentation of the common carotid artery intima media," *Med Biol Eng Comput*, vol. 45, no. 1, pp. 35–49, Jan. 2007, doi: 10.1007/s11517-006-0140-3.

[13] C. P. Loizou, A. Nicolaides, E. Kyriacou, N. Georghiou, M. Griffin, and C. S. Pattichis, "A Comparison of Ultrasound Intima-Media Thickness Measurements of the Left and Right Common Carotid Artery," *IEEE J Transl Eng Health Med*, vol. 3, pp. 1–10, 2015, doi: 10.1109/JTEHM.2015.2450735.

[14] K. M. Meiburger *et al.*, "Carotid Ultrasound Boundary Study (CUBS): Technical considerations on an open multi-center analysis of computerized measurement systems for intima-media thickness measurement on common carotid artery longitudinal B-mode ultrasound scans," *Comput Biol Med*, vol. 144, no. 105333, 2022, doi: 10.1016/j.compbiomed.2022.105333.

[15] L. Gago, M. del M. Vila, M. Grau, B. Remeseiro, and L. Igual, "An end-to-end framework for intima media measurement and atherosclerotic plaque detection in the carotid artery," *Comput Methods Programs Biomed*, vol. 223, p. 106954, 2022, doi: 10.1016/j.cmpb.2022.106954.

[16] N. Laine, G. Zahnd, H. Liebgott, and M. Orkisz, "Segmenting the carotid-artery wall in ultrasound image sequences with a dual-resolution U-net," *IEEE International Ultrasonics Symposium, IUS*, vol. 2022-Octob, 2022, doi: 10.1109/IUS54386.2022.9957590.

[17] F. Marzola, K. M. Meiburger, F. Molinari, and M. Salvi, "Can multiple segmentation methods enhance deep learning networks generalization? A novel hybrid learning paradigm," in *SPIE Medical Imaging*, 2023.

[18] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans Med Imaging*, vol. 23, no. 7, pp. 903–921, 2004, doi: 10.1109/TMI.2004.828354.

[19] G. Zahnd *et al.*, "A Fully-Automatic Method to Segment the Carotid Artery Layers in Ultrasound Imaging: Application to Quantify the Compression-Decompression Pattern of the Intima-Media Complex During the Cardiac Cycle," *Ultrasound Med Biol*, vol. 43, no. 1, pp. 239–257, Jan. 2017, doi: 10.1016/J.ULTRASMEDBIO.2016.08.016.

[20] F. Molinari *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," *Comput Methods Programs Biomed*, vol. 108, no. 3, pp. 946–960, 2012, doi: 10.1016/j.cmpb.2012.05.008.

[21] Mms. Contributors, "MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark." 2020. [Online]. Available: https://github.com/open-mmlab/mmsegmentation

[22] M. Salvi *et al.*, "DermoCC-GAN: A new approach for standardizing dermatological images using generative adversarial networks," *Comput Methods Programs Biomed*, vol. 225, p. 107040, 2022, doi: 10.1016/j.cmpb.2022.107040.