Automated CIMT measurement from ultrasound using deep learning with uncertainty estimation

Iulia-Maria Zbîrcea, Oana-Sorina Chirila, Elena-Larisa Zimbru, Frank Kramer, Florian Auer

Angaben zur Veröffentlichung / Publication details:

Zbîrcea, Iulia-Maria, Oana-Sorina Chirila, Elena-Larisa Zimbru, Frank Kramer, and Florian Auer. 2026. "Automated CIMT measurement from ultrasound using deep learning with uncertainty estimation." In *Health sciences informatics leads and empowers the digital health transformation*, edited by John Mantas, Arie Hasman, Parisis Gallos, Reinhold Haux, and Konstantinos Karitis, 739–43. Amsterdam: IOS Press. https://doi.org/10.3233/shti260943.

Health Sciences Informatics Leads and Empowers the Digital Health Transformation
J. Mantas et al. (Eds.)

doi:10.3233/SHTI260943

Automated CIMT Measurement from Ultrasound Using Deep Learning with Uncertainty Estimation

Iulia-Maria ZBÎRCEA[ab,1], Oana-Sorina CHIRILA[a], Elena-Larisa ZIMBRU[b], Frank KRAMER[c] and Florian AUER[c]
[a] *Department of Automation and Applied Informatics, "Politehnica" University of Timișoara, Timișoara, Romania*
[b] *Center of Immuno-Physiology and Biotechnologies, Department of Functional Sciences, "Victor Babeș" University of Medicine and Pharmacy, Timișoara, Romania*
[c] *IT-Infrastructure for Translational Medical Research, University of Augsburg, Augsburg, Germany*
ORCiD ID: Iulia-Maria Zbîrcea https://orcid.org/0009-0008-1374-286X, Oana-Sorina Chirila https://orcid.org/0000-0001-6165-1710, Elena-Larisa Zimbru https://orcid.org/0009-0000-0320-105X, Frank Kramer https://orcid.org/0000-0002-2857-7122, Florian Auer https://orcid.org/0000-0002-5320-8900

Abstract. Carotid intima-media thickness (CIMT) is a widely used marker for cardiovascular risk assessment, but manual measurement from ultrasound images is time-consuming and subject to substantial inter-observer variability. We propose LUCID - a single-stage deep learning pipeline combining a U-Net with a pretrained ResNet34 encoder for segmentation, sub-pixel boundary extraction for CIMT computation, and Monte Carlo Dropout with post-hoc calibration for uncertainty estimation. Trained on only 500 expert-annotated images from the Carotid Ultrasound Boundary Study (CUBS) benchmark using five-fold cross-validation, the model achieves 0.142 mm mean absolute error, matching the best traditional method by Consiglio Nazionale delle Ricerche ($CNR_{IT,}$ 0.139 mm) while requiring no task-specific preprocessing. The calibrated uncertainty estimation feeds a triage system that automatically accepts confident predictions and flags uncertain cases for clinical review. This is the first CIMT measurement method to integrate calibrated uncertainty estimation, enabling safer deployment in clinical screening workflows.

Keywords. Carotid intima-media thickness, medical image segmentation, uncertainty estimation

1. Introduction

Cardiovascular disease remains the world's primary cause of mortality [1]. Carotid intima-media thickness (CIMT), measured from B-mode ultrasound, is an established non-invasive marker for atherosclerosis [2]. However, manual CIMT measurement is time-consuming and subject to inter-observer variability on the order of clinically relevant differences. Automated CIMT measurement has been explored using both

[1] Corresponding Author: Iulia-Maria Zbîrcea, iulia.zbircea@student.upt.ro

conventional image-processing techniques and deep learning approaches. Traditional methods, such as CNR_{IT}, depend on hand-crafted features and explicit boundary modeling, whereas modern deep learning frameworks employ multi-stage pipelines trained on large annotated datasets [3,4]. Despite achieving strong segmentation performance on benchmark datasets, these methods generally lack calibrated uncertainty estimates. In clinical settings, however, the ability to identify unreliable predictions is crucial, as inaccurate measurements may result in misclassification and inappropriate patient management.

In this work, we present LUCID (**L**earning **U**ncertainty for **C**arotid **I**ntima-media thickness **D**etection), a single-stage CIMT measurement pipeline with MC-Dropout uncertainty estimation [8]. We evaluate LUCID on the publicly available Carotid Ultrasound Boundary Study (CUBS) benchmark and demonstrate the resulting uncertainty estimates enable reliable triage of difficult cases, supporting human-in-the-loop review in clinical workflows

2. Methods

2.1. The CUBS Dataset

We used the CUBS benchmark [3], which contains 500 B-mode ultrasound images acquired at five clinical centers with predefined five-fold cross-validation splits. Each image includes expert boundary annotations and a calibration factor for converting pixels to millimeters. Following benchmark convention, Manual-A1 was used as ground truth [4,5]. Binary training masks were generated by filling the region between the lumen-intima (LI) and media-adventitia (MA) boundaries.

2.2. Deep Learning Pipeline

Images are resized to 512 × 512 pixels and normalized to [0, 1]. A median filter with a kernel size of 3 x 3 is applied to reduce ultrasound speckle noise. Contrast Limited Adaptive Histogram Equalization (CLAHE) is additionally applied as a stochastic augmentation with a probability of 0.3, using a clip limit of 2.0 and a tile grid size of 8 × 8, to improve contrast robustness across varying conditions.

We employ a U-Net [6] with a ResNet34 [7] encoder pretrained on ImageNet. The decoder uses progressively reduced channel sizes of [256, 128, 64, 32, 16] with batch normalization. Dropout2d layers (p = 0.3) are inserted after the encoder bottleneck, at mid-level encoder features, and after the final decoder output. To address the severe class imbalance between intima-media complex (IMC) and background pixels, we minimize a composite loss:

$$L = 0.5L_{BCE} + 0.5L_{Dice} \quad (1)$$

The Binary Cross-Entropy (BCE) uses a positive class weight of 3.0 to increase the penalty for missed IMC pixels, while the Dice loss directly optimizes region overlap, which is especially effective for thin structures. Training runs for a maximum of 100 epochs with early stopping (patience of 20 epochs) based on validation loss. Mixedprecision training (FP16) is used on an NVIDIA Titan RTX GPU (24 GB). The checkpoint with the highest validation Dice score is selected for evaluation.

2.3. CIMT Computation

Boundaries are extracted from the probability map at sub-pixel resolution using a $\tau = 0.5$ threshold. Column-wise thickness is computed as the distance between LI and MA boundaries, with outliers removed via z-score filtering ($|z| < 2.5$). The final CIMT is the median of the remaining values, converted to millimeters using the calibration factor.

2.4. Uncertainty Estimation and Calibration

We adopt MC-Dropout [8] over deep ensembles [9] due to its single-model formulation, compatibility with existing dropout, and widespread use in medical image segmentation [10]. The calibration gap relative to ensembles is addressed via post-hoc calibration rather than ~5× additional compute. At inference, dropout ($p = 0.3$, selected over {0.1–0.5} on the validation set) remains active while all other layers operate in evaluation mode. Each image undergoes T stochastic forward passes, producing T independent CIMT estimates whose mean serves as the prediction and standard deviation as the uncertainty measure:

$$\sigma_{CIMT} = \mathrm{std}(\mathrm{CIMT}_1, \mathrm{CIMT}_2, \ldots, \mathrm{CIMT}_T) \tag{2}$$

$T = 20$ was chosen as the point of diminishing returns ($\Delta < 1\%$ in predictive standard deviation at $T = 50$).

2.5. Triage System

We define a three-tier triage system to classify each measurement based on the calibrated uncertainty. The CI half-width *h* is computed as:

$$h = z \times s \times \sigma_{CIMT} \tag{3}$$

where $z = 1.96$ is the critical value of the standard normal distribution for a 95% confidence level. The threshold δ represents a clinical choice: if the model's uncertainty is no larger than the disagreement between two expert sonographers measuring the same image, the automated measurement is at least as trustworthy as a second opinion. On the CUBS dataset, the inter-observer disagreement between expert A2 and A1 is $\sigma_{\mathrm{inter}} = 0.259$ mm, resulting in $\delta = 0.508$ mm.

Based on h and δ, each measurement is assigned to one of three tiers:

- **Auto-accept** ($h < \delta$): model precision is within inter-observer agreement.
- **Review** ($h \geq \delta$): the model is not confident enough for unsupervised use.
- **Reject:** the measurement is discarded if any of the following hold:
  - (i) predicted CIMT falls outside the physiologically plausible range [0.2, 2.5] mm,
  - (ii) the valid segmentation region spans fewer than 50 px, or
  - (iii) fewer than half of the MC-Dropout samples yield a valid measurement.

3. Results

3.1. Segmentation and CIMT Accuracy

Across the five folds, the model achieved 0.142 mm mean absolute error, 0.776 mean Dice score, and 0.644 mean IoU. Performance was comparable to the traditional CNR_{IT} method (0.139 mm) [4], despite using a single-stage pipeline without task-specific preprocessing. The highest errors occur in images with plaque, very thick or thin walls, or poor acquisition quality.

Table 1. Comparison of CIMT measurement methods on the CUBS benchmark.

Method	Type	Design	# Training Images	MAE (mm)
CaroSegDeep	DL	two-stage	2.676 (CUBS1+2)	0.106 ± 0.089
CNR_{IT}	Traditional	-	-	0.139 ± 0.119
LUCID	**DL**	**single-stage**	**500**	**0.142 ± 0.133**
POLITO_UNET	DL	single-stage	500	0.174 ± 0.120
Expert disagreement	Manual	-	-	0.192 ± 0.177

Table 1 places the results in the context of published methods. Our single-stage model achieves comparable performance to the best traditional method (CNR_{IT}) and narrows the gap to CaroSegDeep, which relies on a more complex two-stage architecture trained on more than five times as many images. Notably, all automated methods outperform the inter-expert disagreement between Manual-A1 and Manual-A2.

3.2. Uncertainty Estimation

The uncertainty is monotonically predictive of absolute error (Spearman $\rho = 0.276$, $p = 3.6 \times 10^{-10}$). Stratifying predictions into quartiles by confidence confirms a consistent gradient: Q1 (most confident) achieves 0.099 mm MAE, compared to 0.124 mm (Q2), 0.152 mm (Q3), and 0.195 mm (Q4). Since MC-Dropout CI are known to be systematically underestimated [9], we apply post-hoc calibration by scaling the standard deviation by a factor s, determined on the validation set such that the resulting 95% prediction interval achieves nominal coverage. Calibration is verified across multiple confidence levels.

3.3. Triage System Evaluation

Applying the defined triage system, 80% of predictions (n=400) are auto-accepted at 0.129 mm MAE. The remaining 20% are flagged for clinician review. No predictions were rejected on this dataset, indicating that the model produces anatomically plausible segmentations for all 500 benchmark images.

4. Discussion

Competitive performance on a limited dataset is facilitated by a pretrained ResNet34 encoder, whose transferable low-level features (e.g., edges and textures) generalize to ultrasound. The most confident quartile attains 0.099 mm MAE, outperforming CaroSegDeep (0.106 mm) trained on substantially larger data, indicating that uncertainty-guided selection can offset data scarcity. Unlike prior CIMT methods, the proposed pipeline provides calibrated uncertainty estimates and triages low-confidence predictions for clinician review, a prerequisite for safe clinical deployment. Although deep ensembles yield superior calibration, their cost under cross-validation is prohibitive, making MC-Dropout a tractable single-model alternative whose calibration gap is addressable through temperature scaling.

However, this approach has several limitations: evaluation on a single benchmark, in-distribution calibration, and a fixed triage threshold δ tied to the inter-observer disagreement of the CUBS dataset. Generalization to unseen scanners and populations requires external validation. Practical deployment further requires DICOM integration, management of ~2 s per-image inference latency, and effective uncertainty visualization.

5. Conclusion

We presented LUCID, a system for automated CIMT measurement from B-mode ultrasound images, combining a single U-Net with a pretrained ResNet34 encoder and MC-Dropout-based uncertainty estimation with post-hoc calibration. On the CUBS benchmark, the model achieves 0.142 mm MAE, matching the best traditional method CNR_{IT} (0.139 mm) while using only 500 training images.

The principal strength of our approach is that it combines competitive measurement accuracy with reliable uncertainty estimation in a simple, reproductible, single-stage pipeline. The integrated triage mechanism enables safe deployment in clinical workflows where automated screening must coexist with expert oversight. The source code is available at https://github.com/IuliaMZbircea/cimt-measurement.

References

[1] Lorenz MW, et al. Prediction of clinical cardiovascular events with carotid intima-media thickness: A systematic review and meta-analysis. Circulation. 2007;115(4):459-467. doi:10.1161/CIRCULATIONAHA.106.628875.
[2] Touboul PJ, et al. Mannheim carotid intima-media thickness and plaque consensus (2004-2006-2011). Cerebrovasc Dis. 2012;34(4):290-296. doi:10.1159/000343145.
[3] Meiburger KM, et al. Carotid Ultrasound Boundary Study (CUBS): Technical considerations on an open multi-center nalysis of computerized measurement systems for intima-media thickness measurement. Comput Biol Med. 2022;144:105333. doi:10.1016/j.compbiomed.2022.105333.
[4] Bianchini E, et al. Carotid intima-media thickness measurement: FOAM operator versus automatic edge detection. Ultrasound Med Biol. 2010;36(5):799-807. doi:10.1016/j.ultrasmedbio.2010.02.004.
[5] Laine N, et al. Carotid artery wall segmentation in ultrasound image sequences using a deep convolutional neural network. In: ICCVG; 2022. p. 1-12. doi:10.1007/978-3-031-22025-8_6
[6] Ronneberger O, et al. U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI; 2015. p. 234-241. doi:10.1007/978-3-319-24574-4_28.
[7] He K, et al. Deep residual learning for image recognition. In: CVPR; 2016. p. 770-778. doi:10.1109/CVPR.2016.90.
[8] Gal Y, et al. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: ICML; 2016. p. 1050-1059. doi:10.5555/3045390.3045502.
[9] Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: NeurIPS; 2017. p. 6402-6413. doi:10.48550/arXiv.1612.01474.
[10] Nair T, Precup D, Arnold DL, Arbel T. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. Med Image Anal. 2020;59:101557.