Comparative assessment of uncertainty-aware deep learning methods for atherosclerosis risk stratification from carotid ultrasound imaging

Kalliopi Sarafi
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
kellysaraf2000@gmail.com

Theofanis Ganitidis
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
orcid.org/0009-0006-7794-9793

Maria Athanasiou
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
orcid.org/0000-0003-1575-9100

Konstantina S. Nikita
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
orcid.org/0000-0001-8255-4354

2025 IEEE 25th International Conference on Bioinformatics and Bioengineering (BIBE) | 979-8-3315-5899-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/BIBE66822.2025.00047

***Abstract*— Carotid atherosclerosis represents a major risk factor for ischemic stroke, requiring accurate risk stratification for effective clinical intervention. While deep learning models demonstrate excellent performance in medical image analysis, their lack of uncertainty quantification limits deployment in clinical environments. This study investigates the use of two uncertainty estimation techniques, Monte Carlo Dropout (MCD) and Deep Ensembles (DE), in cardiovascular risk prediction from B-mode carotid ultrasound images. Using the CUBS dataset for training and two datasets (ATTIKON and BUSI) for external evaluation under distributional shift and out-of-distribution (OOD) conditions, the model's performance, calibration, and robustness are assessed. Results demonstrate that MCD provides superior generalization and OOD detection capabilities (AUC = 0.9589), while DE excels in graceful degradation through rejecting high uncertainty samples, achieving 16.49% accuracy improvement on low uncertainty samples. These findings highlight the complementary strengths of both methods and underscore the critical importance of uncertainty aware AI in clinical decision support systems.**

Keywords*— *uncertainty estimation, deep learning, carotid atherosclerosis, medical imaging, Monte Carlo dropout, Deep Ensembles

I. Introduction

Cardiovascular diseases (CVDs) are the leading cause of mortality worldwide, accounting for approximately 17.9 million deaths annually [1]. Among these, ischemic strokes constitute a major contributor, often resulting in severe long-term disability. In Greece, the annual incidence of stroke is estimated at over 35,000 cases, many of which result in permanent impairment or death [2], [3].

Atherosclerosis of the carotid arteries — characterized by plaque formation and luminal narrowing — is one of the most prevalent causes of ischemic stroke. Notably, the disease frequently progresses asymptomatically until a critical vascular event occurs. Therefore, early risk assessment is crucial for prevention and clinical intervention. B-mode carotid ultrasound imaging offers a non-invasive, widely accessible, and cost-effective tool for visualizing vascular morphology and assessing the degree of stenosis. However, analysis of ultrasound images remains challenging, often subject to inter-operator variability and diagnostic subjectivity, especially in asymptomatic or borderline-risk cases.

In this context, deep learning (DL) models, and specifically convolutional neural networks (CNNs), have demonstrated state-of-the-art performance in medical image analysis tasks, including cardiovascular risk stratification [4], [5]. Yet, despite their high predictive capabilities, conventional DL models lack mechanisms for quantifying uncertainty, a critical limitation in safety-sensitive domains such as clinical diagnostics. Overconfident but incorrect predictions can misguide physicians, reducing trust in automated tools. To overcome this limitation, uncertainty-aware DL methods have been proposed, enabling models to output not only predictions but also calibrated confidence estimates. Techniques such as Monte Carlo Dropout (MCD) [6] and Deep Ensembles (DE) [7] provide probabilistic approximations of epistemic uncertainty, enhancing model trustworthiness, robustness to distributional shifts, and overall reliability in real-world settings.

Another major challenge in clinical deployment is domain shift, which is associated with the performance degradation of models when these are applied to data from different acquisition settings, devices, or patient cohorts. This issue highlights the importance of developing models that are both uncertainty-informed and generalizable across heterogeneous clinical data.

In this study, we propose and evaluate two uncertainty-aware deep learning pipelines for binary risk stratification of carotid atherosclerosis using B-mode ultrasound images. Both approaches incorporate transfer learning and probabilistic modeling for uncertainty estimation. Our methodology includes in-distribution evaluation on the CUBS dataset [8], generalization testing on a domain-shifted clinical dataset from ATTIKON University Hospital [9], and robustness assessment under out-of-distribution (OOD) conditions using unrelated breast ultrasound data [10]. Through this process, we investigate the models' predictive accuracy, its ability to distinguish high-versus low-uncertainty predictions using an uncertainty threshold, and overall trustworthiness in clinically realistic deployment scenarios. By integrating uncertainty estimation into DL-based cardiovascular risk prediction, this study aims to bridge the gap between technical performance

2471-7819/25/$31.00 ©2025 IEEE
DOI 10.1109/BIBE66822.2025.00047

and human-AI interaction, paving the way for safer, more reliable, and transparent AI-assisted diagnostics.

II. Related Work

Deep learning (DL) has demonstrated remarkable success in various medical imaging tasks, particularly in diagnostic applications involving complex data such as ultrasound scans. In vascular imaging, convolutional neural networks (CNNs) have been employed to classify atherosclerotic plaques, estimate intima-media thickness, and predict cardiovascular risk with high accuracy [11], [12], [13]. Earlier studies have also explored plaque motion synchronization and morphological patterns from B-mode ultrasound, emphasizing the diagnostic potential of dynamic carotid imaging [9]. The adoption of pretrained architectures such as InceptionV3, ResNet, and VGG has further improved model performance through transfer learning, especially when training data is limited.

Despite this progress, most existing approaches operate in a deterministic context, providing point estimates without any indication of confidence. This is problematic in clinical domains, where decisions often carry life-altering consequences. To address this, uncertainty-aware DL methods have been introduced. MCD, proposed by Gal and Ghahramani [6], enables uncertainty estimation via multiple stochastic forward passes during inference. Alternatively, DE [7] combine predictions from independently trained models to capture both architectural variability and epistemic uncertainty. These methods have been applied in domains such as diabetic retinopathy detection [14], histopathology analysis [15], and brain segmentation [16], where uncertainty quantification improves interpretability and robustness in clinical decision-making. However, their application in vascular ultrasound imaging, and particularly in carotid atherosclerosis, remains limited.

Furthermore, few studies explicitly consider domain shift and out-of-distribution (OOD) inputs in medical image analysis, despite their critical relevance in real-world deployment. Clinical datasets often differ in acquisition protocols, imaging devices, or patient populations, leading to distributional shift that can severely degrade model performance [17]. While some recent works explore uncertainty as a tool for OOD detection, these are mostly confined to synthetic benchmarks or large public datasets and rarely validated under real clinical conditions [18], [19].

This work aims to bridge this gap by investigating the integration of MCD and DE into deep learning pipelines for carotid atherosclerosis risk stratification, explicitly evaluating uncertainty behavior across both domain-shifted and OOD datasets derived from real clinical settings.

III. METHODOLOGY

The study focuses on evaluating uncertainty estimation techniques in DL-based classification of carotid ultrasound images under varied distributional settings. The proposed pipeline employs three datasets: one for training, validation, and testing under standard conditions; a second to simulate domain shift; and a third to assess out-of-distribution generalization. Two representative uncertainty estimation methods are compared - MCD, using a fixed feature extractor with stochastic inference, and DE, combining multiple independently trained convolutional neural networks (CNNs). All methods are evaluated using common performance and uncertainty metrics to ensure a fair and consistent comparison across in-distribution accuracy, robustness to domain shift, and out-of-distribution detection capability. Fig. 1 provides an overview of the methodology followed.

A. Datasets

TABLE I. provides a summary of the datasets used in this study. Three datasets were utilized to train and evaluate the proposed models under in-distribution, domain-shifted, and out-of-distribution scenarios:

1) CUBS (Carotid Ultrasound B-mode Study): This dataset served as the primary in-distribution dataset [8]. It contains 1,378 labeled carotid ultrasound images (left and right) from 689 patients. Labels indicating low or high cardiovascular risk were computed based on the existence of cardiovascular events on baseline or within a three-year follow-up. Approximately 73.4% of the cases were labeled as low-risk, thus introducing class imbalance.

2) ATTIKON Dataset: This dataset includes 87 ultrasound images obtained from the ATTIKON University Hospital [9]. It was used for external validation and domain-shift evaluation. The high-risk class included patients exhibiting symptoms or presenting a stenosis degree

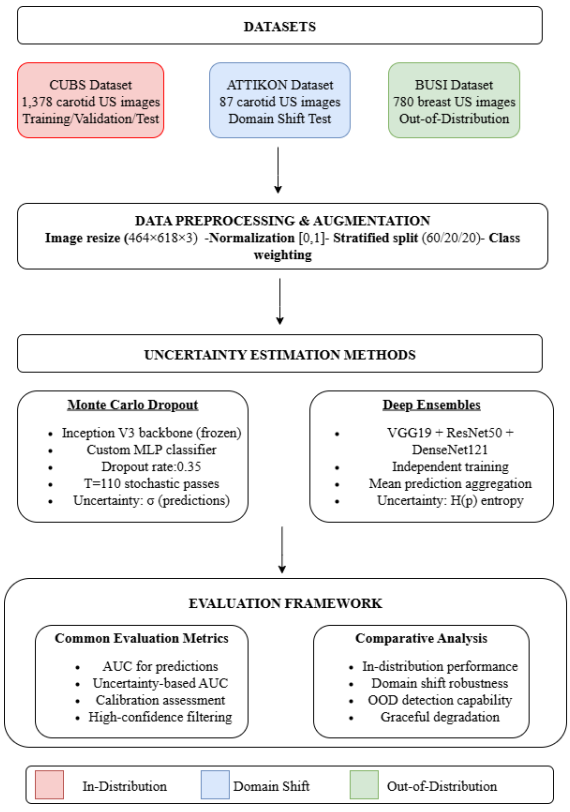

Fig. 1. Overall methodology for uncertainty estimation in carotid atherosclerosis prediction. The framework uses three datasets: CUBS for training/validation/test, ATTIKON for domain shift evaluation, and BUSI for out-of-distribution testing. Two uncertainty estimation approaches are compared: MCD with InceptionV3 backbone and DE combining three CNN architectures. Both methods are evaluated using common metrics to ensure fair comparison across in-distribution performance, domain shift robustness, and out-of-distribution detection capabilities.

TABLE I. SUMMARY OF DATASETS USED FOR TRAINING AND EVALUATION, CATEGORIZED AS IN-DISTRIBUTION (CUBS), DOMAIN-SHIFTED (ATTIKON), AND OUT-OF-DISTRIBUTION (BUSI).

	IN-DISTRIBUTION DATASET	SHIFTED DATASET	OUT-OF-DISTRIBUTION DATASET
DATASET SOURCE	Carotid Ultrasound Boundary Study (CUBS)	Vascular Surgery Department, PIGNA ATTIKON	Breast Ultrasound Dataset
TOTAL IMAGES	1378	87	780
LOW-RISK IMAGES	1012 (73.43%)	20 (22.99%)	-
HIGH-RISK IMAGES	366 (26.57%)	67 (77.01%)	-

exceeding 70%. The dataset was imbalanced, with a predominance of high-risk cases.

3) BUSI (Breast Ultrasound Images): This public dataset [10] contains 780 ultrasound images unrelated to carotid pathology. It was employed solely for out-of-distribution (OOD) evaluation, to assess the behavior of the uncertainty estimation techniques on semantically unrelated inputs.

B. Preprocessing

All images were resized to 464×618×3 and normalized to a [0,1] range. The CUBS dataset was split into training (60%), validation (20%), and testing (20%) subsets using a stratified sampling strategy. The ATTIKON and BUSI datasets were also resized to the same dimensions and used for inference only. No fine-tuning was performed on these datasets. Class weights were computed using the inverse frequency method to address label imbalance during training.

C. Monte Carlo Dropout (MCD)

The MCD approach employed a pre-trained InceptionV3 network as a frozen feature extractor. A custom multilayer perceptron (MLP) was appended on top, consisting of three dense layers with 1024, 512, and 256 units respectively, all using ReLU activations. Dropout with a rate of 0.35 was applied after each dense layer and remained active during inference.

The final output was a sigmoid-activated neuron for binary classification. L1-L2 regularization was applied to the dense layers' kernels. The model was trained using binary cross-entropy loss and the Adam optimizer with a learning rate of 3×10^{-4}, incorporating exponential decay. Early stopping was triggered if validation AUC did not improve for 30 epochs.

At inference time, 110 stochastic forward passes were performed. This value was selected based on empirical validation experiments, where performance stabilized beyond 100 passes, balancing uncertainty estimation quality and computational efficiency. The average of predictions was used as the final output. Youden's J statistic was computed on the CUBS to determine the optimal uncertainty threshold. Samples with uncertainty below this threshold were considered high-confidence and were re-evaluated separately to assess model reliability under trusted conditions.

D. Deep Ensembles (DE)

The Deep Ensemble method combined three independently trained CNNS: VGG19, ResNet50, and DenseNet121. Each network was fine-tuned on the CUBS training set using transfer learning. Binary cross-entropy loss and the Adam optimizer (learning rate 3×10^{-4}) were used for training.

Each model generated a probability score for each sample. The final prediction was the mean of the three probabilities. Predictive uncertainty was estimated using entropy:

$$H(p) = -[p \log(p) + (1-p) \log(1-p)] \quad (1)$$

where p denotes the predicted probability for the positive class. Higher entropy values indicate lower model confidence. The same Youden-based thresholding was performed on the CUBS validation set to determine an entropy threshold for identifying high-confidence predictions. Predictions with entropy values below this threshold were considered more reliable and were re-assessed independently.

E. Evaluation framework

The evaluation process followed a structured framework that was consistently applied to both uncertainty estimation techniques. The evaluation framework, illustrated in Fig. 1, unfolded in sequential stages to assess both predictive accuracy and confidence robustness:

1) In-distribution scenario: The CUBS dataset was used as the main in-distribution source. It was divided into three distinct subsets: training (60%), validation (20%), and test (20%). The selected uncertainty estimation technique—either MCD or DE —was applied to train the corresponding model using the training and validation subsets. The model was evaluated on the CUBS test set, where both classification performance and uncertainty estimation were computed. In addition, precision, recall, and F1-score were computed on the CUBS dataset to provide complementary evaluation metrics.

2) Graceful degradation: The samples from the CUBS test set were filtered to identify high-confidence predictions based on uncertainty thresholds. The subset of low-uncertainty samples were isolated and the model's classification performance, measured by AUC, accuracy, precision, recall, was re-evaluated on this subset to examine whether predictive reliability improved under conditions of low uncertainty. This step aligns with the principle of graceful degradation, where the model identifies cases which might cause performance degradation and refrains from producing predictions on them.

3) Domain shift scenario: The trained model was applied to the ATTIKON dataset, which represents a domain-shifted set, not seen during training. This step assessed generalization capabilities and uncertainty behavior when the model was faced with new but contextually related data.

4) Out-of-distribution scenario: The model was evaluated on the BUSI breast ultrasound dataset, which constituted a fully out-of-distribution (OOD) scenario. This step was critical in determining whether the uncertainty estimation mechanisms can differentiate between in-distribution and unrelated data.

In each of these evaluation scenarios, the model's predictions were accompanied by a corresponding uncertainty estimate. The uncertainty was quantitatively interpreted to assess confidence levels, and the capacity of the model to distinguish between low-risk and high-risk patients under various confidence settings was critically examined.

Evaluation metrics included prediction AUC and accuracy. For the in-distribution CUBS dataset, precision, recall, and F1-score were additionally computed to provide complementary insights into classification performance. An "uncertainty AUC" was also computed by labeling correctly classified samples as 0 and incorrectly as 1. For domain shift and OOD analysis, the uncertainty separation score—a measure of how well the model distinguishes between in-distribution (CUBS) and shifted/OOD (ATTIKON/BUSI) samples was computed by labeling CUBS samples as 0 and ATTIKON/BUSI as 1. Higher scores indicated better shifted/OOD separation via uncertainty. This multi-layered approach enabled comprehensive assessment of model performance, robustness, and trustworthiness, which is particularly important in high-stakes medical settings.

IV. Results

A. Predictive Performance and Confidence Correlation

The obtained results for both MCD and DE demonstrated a clear relationship between predictive confidence and classification accuracy. As shown in the cumulative accuracy plots (Fig. 2 for MCD, Fig. 3 for DE), accuracy was significantly higher on samples with low estimated uncertainty. For the most confident predictions, cumulative accuracy began near 90% and gradually declined as samples with higher uncertainty were included.

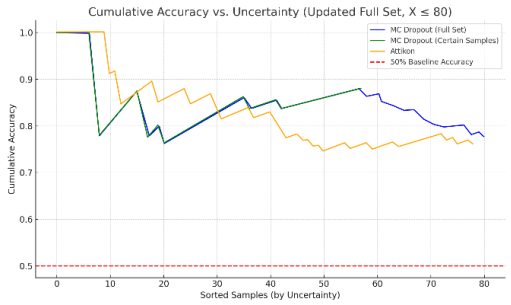

Fig. 2. Cumulative accuracy as a function of uncertainty for the MCD method across the in-distribution test set (CUBS), the high-confidence subset (Certain Samples), and the domain-shifted ATTIKON dataset. The model maintains high accuracy on confident predictions, while performance on the domain-shifted ATTIKON dataset reflects the model's ability to detect distributional shift and moderate its predictions accordingly.

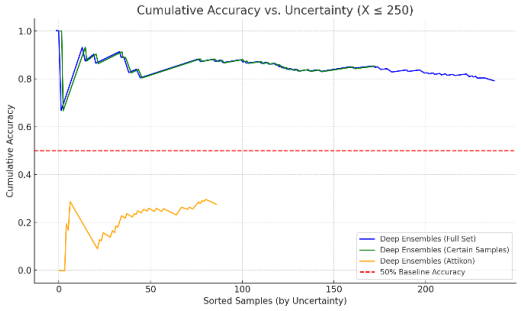

Fig. 3. Cumulative accuracy as a function of uncertainty for the DE method across the in-distribution test set (CUBS), the high-confidence subset (Certain Samples), and the domain-shifted ATTIKON dataset. The model demonstrates higher accuracy on low-uncertainty predictions (Certain Samples), with reduced performance under domain shift (ATTIKON).

Quantitatively, MCD achieved an AUC of 0.5951 for predictions on the full test set, while DE yielded a slightly higher score of 0.6485. On the same dataset, MCD reached a precision of 0.359, recall of 0.918, and F1-score of 0.515, indicating strong sensitivity, whereas DE achieved a precision of 0.503, recall of 0.863, and F1-score of 0.635, reflecting a more balanced trade-off between precision and recall. When restricting evaluation to the low-uncertainty samples (filtered using an uncertainty threshold), both models showed substantial performance improvements, with the AUC increasing to 0.6471 for MCD and 0.7554 for DE, thus highlighting the effectiveness of uncertainty-based filtering in improving prediction reliability.

Moreover, DE showed a notable advantage in uncertainty calibration, with an uncertainty-based AUC of 0.7511 compared to 0.5933 for MCD on the same dataset. This suggests that DE not only yielded better predictions for confident samples but also provided more discriminative uncertainty estimates.

B. Generalization on Shifted Data

To assess generalization, both models were evaluated on a dataset drawn from a different clinical source (ATTIKON), which differed in patient population and data distribution. As expected, performance decreased relative to the original test set. For MCD, the prediction AUC dropped to 0.6263 and the uncertainty AUC to 0.5470. DE showed a more intense decline, with a prediction AUC of 0.4673 and an uncertainty AUC of 0.4447. These results highlighted that both models struggled to maintain performance under distributional shift, though MCD appeared more robust in this context. However, both methods still provided meaningful uncertainty signals. The uncertainty separation score was 0.7934 for MCD and 0.7871 for DE, indicating comparable sensitivity to distributional changes.

C. Out-of-Distribution Detection

To evaluate OOD detection capability, both models were tested on ultrasound data from a breast imaging dataset (BUSI), lying completely outside the training distribution. The comparison of uncertainty distributions showed that OOD samples consistently received higher uncertainty scores.

The combined ROC curve, presented in Fig. 4, illustrates each model's ability to separate in-distribution (ID) from OOD data. MCD achieved a very high AUC of 0.9589, while DE scored an AUC of 0.7641. These results suggest that MCD was more effective in identifying unfamiliar inputs based on uncertainty, making it a strong candidate for safety-critical deployments.

D. Summary of Evaluation Metrics

TABLE II. presents a comprehensive summary of the obtained evaluation metrics across both uncertainty estimation methods and various evaluation scenarios. While DE excelled in calibrated uncertainty estimation and in-distribution prediction accuracy, MCD offered superior performance under distribution shift and OOD detection.

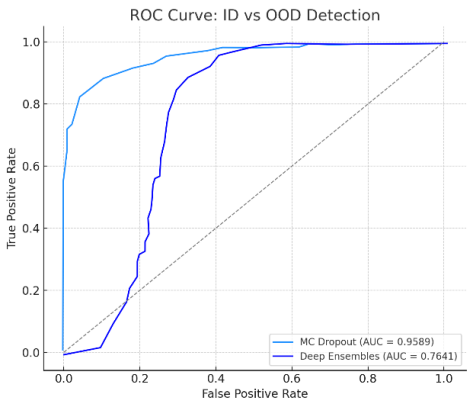

Fig. 4. ROC curve showing the ability of MCD and DE to distinguish in-distribution (CUBS) from out-of-distribution (BUSI) samples based on uncertainty. MCD achieves a significantly higher AUC (0.9589), suggesting enhanced reliability in detecting unfamiliar inputs.

V. Discussion

In this study, we explored the incorporation of uncertainty estimation techniques into deep learning models for carotid atherosclerosis risk stratification from ultrasound images. Given the high-stakes nature of medical diagnosis, quantifying model confidence becomes imperative for enhancing safety and interpretability.

We implemented two uncertainty-aware deep learning pipelines for carotid ultrasound risk stratification. Monte Carlo Dropout (MCD) employed a pre-trained InceptionV3 backbone, while Deep Ensembles (DE) combined multiple CNN architectures. Both approaches were evaluated across in-distribution, domain-shifted, and out-of-distribution datasets to comprehensively assess predictive performance and robustness.

Both the CUBS and ATTIKON datasets exhibit strong class imbalance. To address this, we applied inverse-frequency class weighting in the loss function, which compensates for minority under-representation without artificially altering the dataset distribution. Alternative strategies such as oversampling, undersampling, or balanced subsampling were considered less suitable. Resampling approaches in small datasets such as ATTIKON could cause overfit and amplify the contribution of noisy samples. In addition, we employed early stopping with validation AUC as the monitored criterion to prevent overfitting, ensuring that training stopped once generalization performance plateaued. This combined strategy stabilized optimization, improved sensitivity to the minority class, and supported more robust generalization. Finally, we highlight that our evaluation framework, externally validating to a shifted dataset with different class distributions, further underscores the importance of robustness under class imbalance, as it demonstrates the model's capacity to transfer across cohorts with heterogeneous prevalence rates.

Our evaluation revealed distinct advantages of each uncertainty estimation method across different dimensions of performance and robustness. The MCD approach demonstrated superior performance in distinguishing between in-distribution and OOD samples. Specifically, MCD achieved an AUC of 0.9589 in OOD detection (CUBS vs BUSI), indicating a strong capability to signal epistemic uncertainty when encountering unfamiliar inputs.

TABLE II. Summary of AUC scores for both MCD and DE across multiple evaluation scenarios. Results include in-distribution performance (CUBS), performance after filtering high-uncertainty samples, domain-shift generalization (ATTIKON), and OOD detection using breast ultrasound data (BUSI).

	Metrics	MCD	DE
CUBS	Prediction AUC	59.51 %	64.85 %
	Uncertainty AUC	59.33 %	75.11 %
	Prediction AUC (after rejecting high uncertainty samples)	64.71 %	75.54 %
ATTIKON	Prediction AUC	62.63 %	46.73 %
	Uncertainty AUC	54.70 %	44.47 %
Domain shift / OOD detection	Uncertainty separation score (CUBS vs ATTIKON)	79.34 %	78.71 %
	Uncertainty separation score (CUBS vs BUSI)	95.89 %	76.41 %

Furthermore, the method maintained its discrimination performance when applied to the ATTIKON dataset—a shifted distribution compared to the training set—achieving a prediction AUC of 0.6263 and an uncertainty AUC of 0.5470, despite notable differences in risk profiles and acquisition settings.

This robustness to distributional shifts suggests that MCD is well-suited for clinical applications where the data distribution may vary, such as across institutions or populations. Its ability to produce calibrated uncertainty estimates further enables the reliable separation of high- and low-confidence predictions, supporting decisions under uncertainty.

In contrast, the DE method excelled in isolating high-confidence predictions, or certain samples. The model showed graceful degradation, retaining high prediction quality on samples with low uncertainty. On the CUBS dataset, filtering based on uncertainty yielded an AUC improvement from 0.6485 to 0.7554, demonstrating the method's utility in controlled settings where predictive reliability is prioritized. This characteristic aligns well with medical deployment scenarios where models are expected to abstain from predictions in borderline or ambiguous cases, thereby mitigating risk.

However, DE exhibited limitations in its generalization under domain shift. The uncertainty AUC scores on the ATTIKON and OOD datasets were comparably low, indicating a lack of sensitivity to distributional differences. This may be attributed to overfitting on the training set, leading to degraded OOD detection capabilities (AUC = 0.7641) compared to MCD. Moreover, the domain discrepancy between CUBS (predominantly low-risk) and ATTIKON (predominantly high-risk) may have further impacted the ensemble's ability to model the uncertainty landscape effectively.

Overall, neither method outperformed the other across all axes of evaluation. Rather, they were shown to offer complementary strengths:

- MCD is more appropriate for scenarios involving distributional shift, anomaly detection, or robustness under data variability.
- DE are advantageous when prediction confidence is paramount, such as in automated decision-support systems that emphasize reliability over coverage.

These insights pinpoint the potential of a hybrid or adaptive deployment strategy, where model selection and confidence thresholds could be dynamically adjusted based on clinical context.

Potential limitations of the study refer to the relatively small dataset sizes and class imbalances that may affect generalizability. The inverted class imbalance between the ATTIKON and CUBS datasets resemble a marginal, although realistic, scenario of distribution shift. Future work should therefore focus on applying these methods to larger and more balanced cohorts, where class ratios better reflect clinical prevalence and allow more representative evaluation. Finally, emerging hybrid approaches such as Masksembles, which combine the benefits of DE and MCD, represent a promising direction for further improving uncertainty quantification in clinical applications.

VI. Conclusion

This study investigated the use of the MCD and DE uncertainty estimation techniques for carotid atherosclerosis risk stratification using ultrasound imaging. MCD demonstrated exceptional generalization and out-of-distribution detection capabilities, while DE excelled in uncertainty calibration and high confidence prediction filtering. The results highlight the complementary nature of these approaches and their potential for enhancing the safety and reliability of AI-assisted medical diagnosis. Neither method universally outperformed the other; rather, each offered distinct advantages suited to different clinical scenarios and requirements. Future work should explore hybrid uncertainty estimation strategies, evaluate performance on larger and more balanced datasets, and investigate integration with human-in-the-loop systems for optimal clinical deployment. The development of uncertainty-aware medical AI represents a crucial step toward more transparent, reliable, and clinically acceptable automated diagnostic tools.

References

[1] "Cardiovascular diseases (CVDs)." Accessed: May 01, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] P. Song *et al.*, "Global and regional prevalence, burden, and risk factors for carotid atherosclerosis: a systematic review, meta-analysis, and modelling study," *Lancet Glob. Health*, vol. 8, no. 5, pp. e721–e729, May 2020, doi: 10.1016/S2214-109X(20)30117-0.

[3] D. Panagiotakos *et al.*, "The burden of cardiovascular disease and related risk factors in Greece: the ATTICA epidemiological study (2002-2022)," *Hell. J. Cardiol. HJC Hell. Kardiologike Epitheorese*, pp. S1109-9666(24)00113–1, Dec. 2024, doi: 10.1016/j.hjc.2024.05.009.

[4] G. Litjens *et al.*, "A Survey on Deep Learning in Medical Image Analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.

[5] A. Esteva *et al.*, "Prostate cancer therapy personalization via multi-modal deep learning on randomized phase III clinical trials," *NPJ Digit. Med.*, vol. 5, no. 1, p. 71, Jun. 2022, doi: 10.1038/s41746-022-00613-w.

[6] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," Oct. 04, 2016, *arXiv*: arXiv:1506.02142. doi: 10.48550/arXiv.1506.02142.

[7] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," Nov. 03, 2017, *arXiv*: arXiv:1612.01474. Accessed: Nov. 02, 2024. [Online]. Available: http://arxiv.org/abs/1612.01474

[8] 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, Aug. 2021, doi: 10.1016/j.ultrasmedbio.2021.03.022.

[9] S. Golemati, E. Patelaki, A. Gastounioti, I. Andreadis, C. D. Liapis, and K. S. Nikita, "Motion synchronisation patterns of the carotid atheromatous plaque from B-mode ultrasound," *Sci. Rep.*, vol. 10, no. 1, p. 11221, Jul. 2020, doi: 10.1038/s41598-020-65340-2.

[10] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data Brief*, vol. 28, p. 104863, Feb. 2020, doi: 10.1016/j.dib.2019.104863.

[11] H. Zhang and F. Zhao, "Deep Learning-Based Carotid Plaque Ultrasound Image Detection and Classification Study," *Rev. Cardiovasc. Med.*, vol. 25, no. 12, p. 454, Dec. 2024, doi: 10.31083/j.rcm2512454.

[12] S. Candemir *et al.*, "Automated coronary artery atherosclerosis detection and weakly supervised localization on coronary CT angiography with a deep 3-dimensional convolutional neural network," *Comput. Med. Imaging Graph.*, vol. 83, p. 101721, Jul. 2020, doi: 10.1016/j.compmedimag.2020.101721.

[13] T. Ganitidis, M. Athanasiou, K. Dalakleidi, N. Melanitis, S. Golemati, and K. S. Nikita, "Stratification of carotid atheromatous plaque using interpretable deep learning methods on B-mode ultrasound images," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Aug. 2021, pp. 3902–3905. doi: 10.1109/EMBC46164.2021.9630402.

[14] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, and S. Wahl, "Leveraging uncertainty information from deep neural networks for disease detection," *Sci. Rep.*, vol. 7, no. 1, p. 17816, Dec. 2017, doi: 10.1038/s41598-017-17876-z.

[15] L. Goetz, "Uncertainty estimation for out-of-distribution detection in computational histopathology," Oct. 18, 2022, *arXiv*: arXiv:2210.09909. doi: 10.48550/arXiv.2210.09909.

[16] P. Mojiri Forooshani *et al.*, "Deep Bayesian networks for uncertainty estimation and adversarial resistance of white matter hyperintensity segmentation," *Hum. Brain Mapp.*, vol. 43, no. 7, pp. 2089–2108, Dec. 2022, doi: 10.1002/hbm.25784.

[17] A. Malinin and M. Gales, "Predictive Uncertainty Estimation via Prior Networks," Nov. 29, 2018, *arXiv*: arXiv:1802.10501. doi: 10.48550/arXiv.1802.10501.

[18] Y. Ovadia *et al.*, "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift," Dec. 17, 2019, *arXiv*: arXiv:1906.02530. doi: 10.48550/arXiv.1906.02530.

[19] S. Ye, Y. Xu, D. Chen, S. Han, and J. Liao, "Learning a Single Network for Robust Medical Image Segmentation With Noisy Labels," *IEEE Trans. Med. Imaging*, vol. 43, no. 9, pp. 3188–3199, Sep. 2024, doi: 10.1109/TMI.2024.3389776.