Confidence Estimation in Analyzing Intravascular Optical Coherence Tomography Images with Deep Neural Networks

Lennard Korte, Li Rong Wang, Xiuyi Fan
Nanyang Technological University
Singapore
{LENNARDA001,LIRONG002}@e.ntu.edu.sg, xyfan@ntu.edu.sg

***Abstract*—Atherosclerosis is a leading, yet preventable cause of death worldwide. To diagnose and assess plaque deposits within the arterial walls, experts employ intravascular optical coherence tomography (IVOCT) for the detection and characterization of lesions. Clinical routines acquire a high number of images, necessitating automatic plaque detection for fast and accurate decision support. Deep learning-based plaque detection methods demonstrated remarkable success, but come with frequent prediction inaccuracies. Therefore, a thorough understanding and assessment of predictive reliability is required for future clinical decision support systems. We address this issue by employing a novel uncertainty estimation method, extending deep learning (DL) models that directly learn plaque classification from an IVOCT dataset and form the basis of state-of-the-art architectures, through a confidence estimation branch (CEB). Our models allow for optimized treatment execution by providing physicians with confidence scores, along with disease predictions from DL models. Experiments demonstrate that the estimated confidence correlates to prediction accuracy with the proposed method. We show that generating uncertainty scores with a CEB is feasible, paving the way for more qualitative and effective decision-making in future medical predictive analysis.**

***Index Terms*—Confidence Estimation, Uncertainty Measure, Deep Learning, Medical Predictive Analysis, IVOCT**

I. Introduction

The interpretation of medical images for effective treatment often constitutes a critical challenge for experts [1]. Since atherosclerosis is a leading, yet preventable, cause of death worldwide , visualizing lesions and arterial walls with intravascular optical coherence tomography (IVOCT) is crucial for effective treatment [2]. Analyzing the high number of images captured during clinical routines involves significant time and monetary investments by experts. Thus, tremendous research efforts are dedicated to automatic plaque detection with deep learning (DL) for fast and accurate decision support in automated computer-aided diagnosis (CAD) systems. Automatic IVOCT data analysis methods, including lumen segmentation, stent detection, and plaque identification, have been successfully developed with a high predictive ability [3], [4]. These methods correlate IVOCT images with histology data and use various techniques based on DL with convolutional neural networks (CNNs) to improve accuracy in tasks like tissue layer segmentation and plaque detection [4], [5].

However, these methods are often limited in their ability to accurately capture model uncertainty, or do not capture it at all, which is crucial for deriving well-informed decisions based on AI predictions. Although Bayesian principles have been proposed in ML decades ago [6], only recently computationally feasible methods boosted Bayesian approaches in ML [7]. Consequently, these methods have gained considerable interest as opposed to traditional methods that tend to overfit, dimming their generalization capabilities and performance on unseen data. They are generally incapable of addressing uncertainties. Whereas some uncertainty estimating models have been developed, it has been observed that such models are generally overconfident [8]. To estimate uncertainties, Bayesian inference such as MC dropout via posterior distribution stands out as the main approach [9]. More recently, uncertainty estimations directly from models were explored, i.e., [10] presents an anomaly detection module to estimate prediction uncertainty.

This aspect gains even more importance in the medical sector, where decisions can profoundly affect end-users. In medical predictive analysis, uncertainty estimation enables decision-making based on confidence levels [9], enhancing treatment and risk mitigation. If the predictive model exhibits high confidence in a specific diagnosis, it may suggest allocating more resources or attention to that area to optimize patient treatment. Conversely, low confidence advises caution or alternative strategies to minimize health risks. Uncertainty estimation in plaque prediction using IVOCT plays a crucial role in more accurate diagnosis and treatment planning.

In our study, we advance the framework of deep neural networks, a state-of-the-art classification architecture for medical image classification [1], [5], without compromise to the network's prediction efficacy. We introduce an auto-encoder-based uncertainty estimation method focusing on scope compliance related aleatoric uncertainty by demonstrating its performance in predicting plaque on IVOCT images. Comparing with MC dropout, the proposed method does not require running the inference task multiple times, hence eliminating associated running cost.

This paper is organized as follows: Section II provides an overview of IVOCT image classification and confidence estimation. Section III details our model's architecture and functionality. Finally, in section IV we discuss test results by evaluating confidence estimation together with classification

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00073

performance and conclude in section V.

II. Background

A. IVOCT Image Classification

IVOCT is a key imaging tool in coronary angiography, offering higher axial and lateral resolution of microstructures in arterial walls, although with a smaller field of view compared to intravascular ultrasound (IVUS) [11], [12]. Automated analysis methods for IVOCT data include lumen segmentation [13], stent detection [3], [14], and plaque deposit detection, addressing the challenge of manual image review. Studies have shown that IVOCT images can be correlated with histology data, affirming the feasibility of inferring plaque characteristics from IVOCT data [12]. IVOCT is also utilized in measuring fibrous cap thickness in arterial walls, assessing rupture risks and potential myocardial infarction occurrences [11]. Calcified plaque regions, indicative of increased stenosis risk, have been quantified using IVOCT [15]. The optical backscattering and attenuation coefficient of tissue has been identified as a useful marker for characterizing coronary plaques [16], [17]. Fully-automated plaque characterization methods using texture and optical properties as features have been proposed for plaque detection, achieving classification accuracies between 72.1% and 89.5% for different tissue types [16], and segmentation [18]. K-means clustering and random forests have been used for segmenting calcified plaque regions, achieving sensitivities between 71% and 81% for various tissues. [19].

Deep learning, especially convolutional neural networks (CNNs), has significantly advanced IVOCT image processing for tissue layer and plaque segmentation [1], [20], [21]. An automated plaque segmentation method has been developed using CNN and an improved random walk algorithm, achieving a Jaccard coefficient of 0.864 [22]. Preliminary studies using CNNs have focused on patch-wise approaches for plaque segmentation [23]. Effective preprocessing, like lumen segmentation, is crucial but challenging due to artifacts [13]. While segmentation is vital in many medical applications [1], in IVOCT, the focus is shifting towards identifying plaque deposits in pullbacks, which is crucial for clinical decision support systems. The field has progressed from models like AlexNet [24] to more advanced architectures like ResNet [25], enhancing medical image analysis. The impact and optimal use of transfer learning from datasets like ImageNet in IVOCT are still under investigation [26]. Multi-path structures [5] may resolve the choice between traditional polar and cartesian representations [23], intuitively resembling the artery's anatomical structure. More recent advancements combine DL with hand crafted approaches [27] and trained a Random Forest using CNN features [28].

Most approaches for feature extraction in plaque detection within Intravascular Optical Coherence Tomography (IVOCT) images predominantly rely on DL models as their fundamental component [12], incorporating various architectures, including those with and without skip connections or residual elements [5], [24], [25]. Frequently utilized variants of CNNs are AlexNet, DenseNet, ResNet or VGGNet [29].

B. Uncertainty Estimation

Our uncertainty estimation architecture is inspired by auto-encoder designs, notably those detailed in [30]. A multi-layer perception (MLP) model $f_{\phi,\psi} : \mathbb{R}^i \mapsto \mathbb{R}^o$ featuring N layers of hidden nodes mapping from $\mathbb{R}^i$ to $\mathbb{R}^o$, is trained with input image set X and their labels Y 1. This function $f_{\phi,\psi}$ combines two mappings. In this model, ϕ and ψ represent the parameters for the first M layers and the final $N - M$ layers of nodes, respectively. If the Mth layer has k nodes, then the function f_ϕ maps from $\mathbb{R}^i$ to $\mathbb{R}^k$, and the function f_ψ maps from $\mathbb{R}^k$ to $\mathbb{R}^o$.

$$f_{\phi,\psi} = \arg\min_{\phi,\psi} \|Y - (f_\psi \circ f_\phi)(X)\|^2 \tag{1}$$

Furthermore, an additional MLP branch f_ω to f_ϕ with the same output cardinality i as the input set is appended. This forms the auto-encoder $g : \mathbb{R}^i \mapsto \mathbb{R}^o$ (2) for set X.

$$g = \arg\min_{g} \|X - (f_\omega \circ f_\phi)(X)\|^2 \tag{2}$$

The hypothesis posits that for any given instance x in the input space R_i and its associated label y in the output space R_o it holds that

$$|y - f_{\phi,\psi}(x)| \leftrightarrow |x - (g \circ f_\phi)(x)|. \tag{3}$$

We therefore define the uncertainty score σ for the prediction $f_{\phi,\psi}$ as

$$\sigma = |x - (g \circ f_\phi)(x)|. \tag{4}$$

The illustrated model architecture (Figure 1) offers an uncertainty measure as the reconstruction error σ, which is the absolute difference between the original and rebuilt inputs. In theory σ has a direct relationship with the error in prediction (3). This offers a way to gauge the model's effectiveness prior to seeing the actual results.

The following work is based on two assumptions. First, the prediction model exhibits higher performance on inputs that resemble those in the training set, while effectiveness diminishes on unfamiliar inputs. Secondly, the decoder, trained on the same dataset, accurately reconstructs familiar inputs, resulting in low reconstruction errors (small σ). Conversely, it struggles to reconstruct unfamiliar inputs, resulting in significant reconstruction errors (large σ).

III. Methodology

A. Model Architecture

We employ the ResNet18 [25] (Figure 2) deep neural network as the core of our prediction model as it represents classes of commonly used CNNs that proved their effectiveness in IVOCT image analysis even with small datasets [29]. ResNet18 processes images x of 224x224 pixels with three RGB channels, using an architecture divided into five distinct building blocks f_ϕ (orange). The first building block consists of a 7x7 convolution with stride two and 64 channels. The following 3x3 max-pooling layer likewise has a stride of two. The following four building blocks follow a similar structure. They consisting of two residual blocks with two

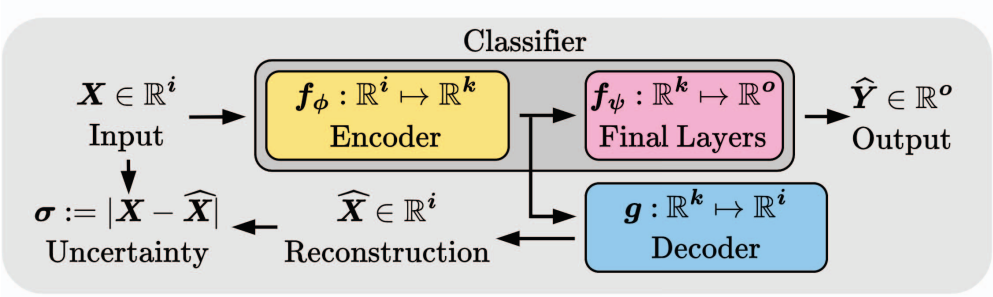

Fig. 1. MLP model architecture with decoder as uncertainty score branch.

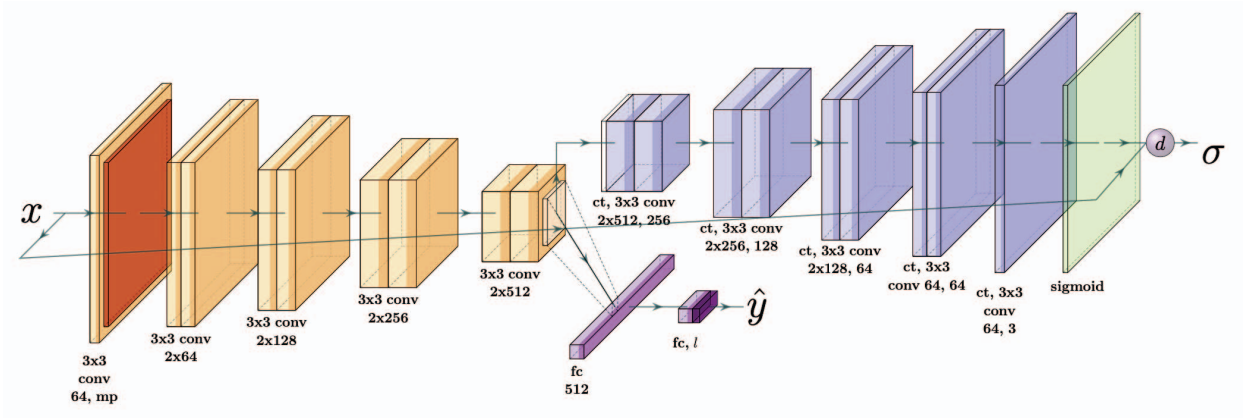

Fig. 2. Model architecture with ResNet18 classifier, consisting of encoder f_ϕ (orange), including max-pooling layers (mp, dark orange), and final layers f_ψ (purple) with input x and prediction $\hat{y}$ returned by loss function l (dark purple). Attached is the confidence estimation branch g (blue), including convolutional transpose layers (ct) and sigmoid layer (green), with loss function d returning reconstruction error σ.

3x3 convolutions each. Both residual blocks come with a skip connection, that feeds the input next residual block as well. Notable the first two blocks maintain an output depth of 64 channels. The first convolution of the later three layers comes with stride 2 that halves the feature resolution and double the output depth of the building block. Following each convolutional layer, batch normalization and a ReLU activation function are employed to stabilize and optimize the network's learning process. This graduated arrangement allows for a progressive refinement of features extracted from IVOCT images, resulting in a total number of $M = 17$ layers. They reduce the computational load and enhancing the extraction of dominant features. The classifier culminates in the final $N - M = 1$ layers f_ψ (purple), consisting of a fully connected (fc) layer, with 512 neurons, followed by a final fully-connected layer tailored to the number of classes ($|C|$) in the dataset, and loss function l for output $\hat{y}$ (dark purple).

Our novel contribution is the augmentation of the CNN with an additional CEB g (see Figure 2) that enables measuring the reconstruction error. Integrated with the last average-pooling layer of ResNet18, the CEB functions as a decoder, attempting to reconstruct the input image and thereby forming an auto-encoder in conjunction with the encoder. The architecture of the decoder mirrors that of encoder. Each block begins with a 2D convolutional transpose (ct) layer for upscaling (dark blue), followed by convolutional layers that directly reflect the design of the corresponding encoder block. Batch normalization and ReLU activation are applied before each convolution in the decoder to enhance performance. The final layer of each block reduces the depth of the feature maps by half, pivotal in reconstructing output images $\hat{X}$ with the original dimensions and number of channels. The loss function d, applied to the reconstruction output $\hat{x}$ of the final sigmoid layer (green) and the ground truth, returns the reconstruction error σ.

B. Dataset

The data set we utilize for our experiments has a total of 3951 individual 16 bit greyscale IVOCT images. To prevent distorted results, none of the images contains a stent. All 347 A-scans from each 360° turn are assembled to a B-scan with a depth of 683 pixels. Three experienced physicians labeled each frame in partially overlapping subsets to mitigate intra-observer variability [14], indicating no plaque or one of six different plaque types. Distinguishing labels were classified repeatedly to ensure label quality by labeling with consensus.

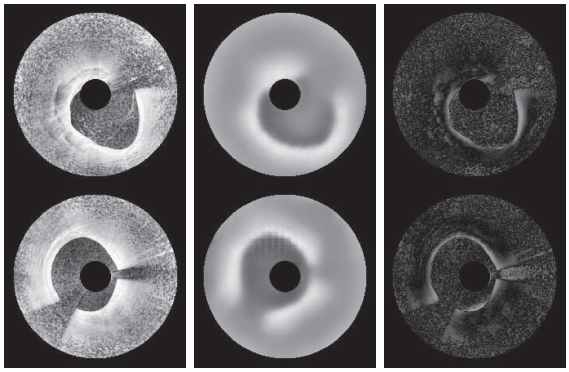

Fig. 3. Two IVOCT image samples (left), respective model outputs (center) and their absolute differences (right).

The IVOCT dataset consist of multiple intravascular serial scans (pullbacks) obtained by 49 different patients. Frames that originate from a single pullback are considered as a group due to their high similarities and strong relation. We adopt four fold cross-validation (CV), to find appropriate hyperparameters and avoid overfitting. The test set contains nine pullbacks. Each validation set contain 10 pullbacks of similar size in total. Because the pullback cardinalities are highly divergent and their label distributions within them are highly imbalanced, we randomly select a maximum of 50 samples of each class in every pullback.

C. Preprocessing and Data Augmentation

For simplicity, we restrict ourselves to a binary classification, distinguishing whether the image represents disease-affected (plaque) or healthy anatomy (no plaque). A one indicates plaque presence, while a zero indicates its absence. Thus, we obtain 2142 positive and 1808 negative image labels.

The polar image $I_p(d, \delta)$ is transformed into cartesian space with the transformations $x = d\cos(\delta)$ and $y = d\sin(\delta)$. We apply bilinear interpolation to each scan, converting it to a more human interpretable cartesian representation $I_c(x, y)$ that fits the models input resolution of 224x224 pixels 3. It is normalized and replicated to fit the three model input channels. We apply CLAHE filter [31] with tile size 5 on the padded image for contrast enhancement to help generalizing and prevent overfitting. Afterwards, the image is cut back to the original size. We apply an inverted circular mask with the diameter equal to the image size to remove the created artefacts. In addition, the catheter is removed with a circular mask x0.17 of the image diameter.

For data augmentation we apply random rotation with $a \in [0°, 360°]$ and random flipping along the x and y direction. The images are partially masked by a random square dimension of $x, y \in [0, 0.2]$, where d is the proportion of the images size. to 80% of the training images.

D. Training

As with [30], it is crucial that, in contrast to a typical auto-encoder model in which the encoder and decoder are trained simultaneously, in our approach the encoder f_ϕ is initially trained as component of the prediction model (see figure 1). Once training is completed, we proceed to train the decoder g independently (2). This method guarantees that the prediction accuracy of $f_{\phi,\psi}$ is not impaired.

In our experiment, the classifier as well as the decoder are trained in batches of size 8 and a weight decay of 0.01 to allow learning essential features, while minimizing overfitting. We use an initial learning rate of 5e-5. It rate undergoes an exponential decrease, being scaled down by 0.9 after each epoch. Each of the four models in its CV is trained for 10 epochs. Hyperparameters are found with a grid search observing BACC using the validation holdout. After disabling the gradients and backpropagation in the encoder, we train the decoder with the same training set with initial learning rate of 2e-5 for 20 epochs and likewise optimize it with the respective validation holdout for all CVs. As our dataset is small compared to, e.g., ImageNet [24], we make use of transfer learning which has been used successfully in the medical image domain [26].

$$l_n = -\sum_{c=1}^{C} w_c \log \frac{\exp\left(\hat{y}_{n,c}\right)}{\sum_{i=1}^{C} \exp\left(\hat{y}_{n,i}\right)} y_{n,c} \tag{5}$$

For training the classifier, we minimize the batch wise cross-entropy loss L as the mean over all losses l_n (5) in the batch dimension N as: $L = \frac{1}{N}\sum_{n=1}^{N} l_n$, commonly used for classification tasks. The target ground truth variable is indicated by y, while $\hat{y}$ signifies the model output. We determine the training set imbalance for the binary labels "plaque" and "no plaque" in each CV. Imbalance in each CV is handled by weighting the loss of each sample with weight w according to their class $c \in C$, where C is the set of all classes. This approach mitigates the risk of overfitting while concurrently minimizing information loss, a common issue encountered when balancing datasets through the exclusion of samples. The set of all sample losses is defined by mapping the cross-entropy loss to all pairs of y and $\hat{y}$ as: $l(y, \hat{y}) = \{l_1, \ldots, l_N\}$.

$$d_n(x, \hat{x}) = \frac{1}{|x|}\sum_{j=1}^{n}(x_j - \hat{x}_j)^2 \tag{6}$$

In the training of the decoder, mean squared error (MSE) loss is employed (6), which is characterized as the mean of the squared deviations for each pixel, where $|n|$ denotes the total number of pixels. Within the context of our model, the variable x denotes the actual values, corresponding to the training samples fed into the encoder. The variable $\hat{x}$ signifies the predicted values, which in our scenario are the output images generated by the decoder. Additionally, $|n|$ represents the total count of samples within the dataset. We likewise define the set of sample losses by mapping the MSE to all pairs of x and $\hat{x}$ as: $d(x, \hat{x}) = \{d_1, \ldots, d_N\}$. Just like with the classifier, batchwise loss d is the mean over all sample losses.

TABLE I
PERFORMANCE SCORES FOR CLASSIFIERS ON THEIR RESPECTIVE VALIDATION SET IN EACH CV.

CV	BACC	F1	MCC
1	0.71	0.78	0.42
2	0.73	0.74	0.45
3	0.70	0.68	0.41
4	0.66	0.61	0.30

TABLE II
MEAN LOSS FOR DECODERS ON THEIR RESPECTIVE VALIDATION SET IN EACH CV ACROSS THE CONFUSION MATRIX.

CV	TP 1	TN 2	FP 3	FN
1	0.065	0.068	0.068	0.067
2	0.063	0.066	0.065	0.068
3	0.064	0.066	0.066	0.070
4	0.065	0.068	0.067	0.070

IV. RESULTS

A. Performance Evaluation

Similar to [30], we compare three metrics between the base classifier and our extended model:

1) Model classification: plaque detection performance under ideal parameters,
2) Uncertainty estimation performance: MAE between ground truth and predicted image,
3) Uncertainty reliability: correlation between prediction error and uncertainty.

In the process of correlating classifier loss with decoder loss, the logarithmic and weights are eliminated when computing the classifier loss. Following the implementation of the softmax function (5), the outcome are the reciprocal probabilities $\hat{y}_n$ associated with the positive class. These probabilities, when multiplied by the negative target variable y_n, result in a simplified, linear, and non-weighted form of the cross-entropy loss l_n ranging from -1 to 0 (7).

$$l_n = -\text{softmax}(\hat{y}_n)^\intercal y_n \tag{7}$$

To ensure linearity in the decoder loss and limit its range between 0 and 1, we compute the mean absolute error (MAE) or F1 loss, instead of the MSE in the calculation of the decoder loss d_n, when correlating the two losses.

B. Model Classification

We report three main evaluation scores [32] that describe the correlation between the observed and predicted classifications and can be calculated from the confusion matrix only. First, the balanced accuracy (BACC) is retrieved by computing the arithmetic mean of the true positive rate (sensitivity) and true negative rate (specificity). Additionally, we evaluate the classification performance with the Mathews correlation coefficient (MCC or ϕ coefficient) [32]. Lastly, F1 score is the harmonic mean of the precision and recall. It symmetrically represents both precision and recall in one metric.

In the evaluation conducted on the test set, the mean performance scores of the four classifiers yield a BACC of 0.70, complemented by an F1 score of 0.71 and an MCC of 0.40. In addition, we report the performance scores for the metrics BACC, F1 and MCC of the four classifiers on their respective validation set I.

C. Uncertainty Estimation Performance

1) Mean Uncertainty Scores: Considering the test set we see a higher mean decoder loss of 0.68 for false classifications than for true classifications 0.65 II. This holds true for positive as well as negative classifications on every validation set. It can also be observed when applying the different models to their validation sets. When comparing the decoder performance on their respective validation set in each CV, we see an increased difference between mean decoder losses for true and false predictions, if images in the validation set entails a rather high domain shift. More specifically, the overall loss increases for the respective validation set, even more the mean loss for falsely classified samples in the validation set. This is also evident when we examine the classifier loss on the respective validation sets I. Accuracy decreases for sets, where the mean decoder loss for false classifications is higher.

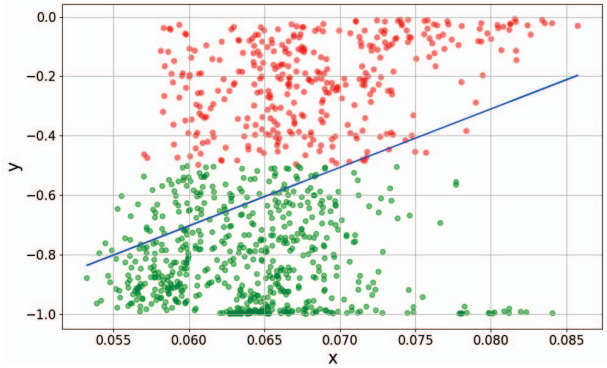

Fig. 4. Loss pair distribution of auto-encoder (x-axis) against classifier loss (y-axis) on test set. We mark positive predictions green and negative ones red.

2) Single Uncertainty Scores: In the loss pair distribution graph 4, we consider single loss pairs. We observe an increasing mean classifier loss on the test set, when considering an increasing proportion of the smallest encoder loss sample. This proves the existence of a relation between decoder loss and classifier loss. It supports the observation from above that a higher mean decoder loss in false predictions corresponds to rather poor classifier performance. When the set does not contain samples from either known or an unknown data, the loss distribution graph does not directly reflect any relationship between the two losses.

Similar as above, we compare the performance of the validation sets on the respective model in the CV. In an optimal scenario, the RCC 5 represents a continuously escalating trend, illustrating the dynamic interplay between encoder and decoder losses. When the classifier performs relatively poor on a validation set, the risk-coverage curve (RCC) increases rather gradually. The rapid ascent at the beginning diminishes and the

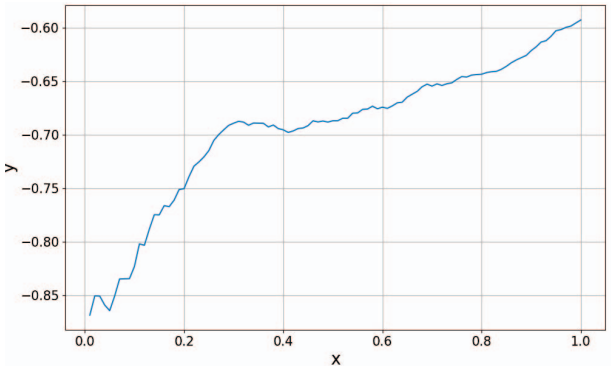

Fig. 5. The risk-coverage curve illustrates the relationship between classifier and decoder loss. The x-axis represents the the sample set size in percent, while the y-axis shows the mean of the computed values. Each data point is derived by initially selecting a set proportion of samples that exhibit the smallest MAE in the decoder evaluation. The mean classifier loss is calculated for these selected samples (7). This process is repeated iteratively 100 times, with each iteration involving an incremental increase in the proportion x.

curve converges more and more towards a straight line. The upper part of the loss pair distribution with a higher classifier loss moves to the right towards a higher decoder loss further away from the lower part of the distribution. This variability cannot be observed as strong when inspecting the performance of the models for different validations on the same test set. This is because the test set always stays the same and the training samples for the models mostly overlap.

D. Uncertainty Reliability

In assessing the classifier's familiarity with a newly introduced sample set, or inversely, the magnitude of the domain shift, two principal methods can be used as uncertainty measure. Both analyse the correlation between the decoder's responses to a known sample set, which bears close resemblance to the training data, and those to a new, unknown sample set. It is assumed that both sets are given nearly equal weight. The first method involves contrasting the mean disparity in outcomes for correct and incorrect classifications between both the known and unknown sample sets, as described above. The second method involves fitting a linear regression to the loss pair distribution. A steep gradient indicates a high degree of similarity between known and unknown dataset, with a higher prediction confidence and vice versa. The inverse provides a metric for quantifying the uncertainty or domain shift. However, simply applying linear regression does not automatically yield meaningful insights. The effectiveness of this method is contingent upon the presence of a "diagonal" correlation between classifier and auto-encoder, with a decoder learning effectively. This may be substantiated by taking Pearson's correlation coefficient (PCC) r [33] into account. For the third CV we report a gradient of 12.36 with a Pearson coefficient of 0.22. The model performs worse in the forth cross-validation, reflected by a stronger slope of 19.65 with a significantly larger Pearson coefficient 0.37.

Furthermore, we utilize the relationship between the two losses and define a threshold for decoder loss to be centered between the minimum and maximum decoder error. It serves to optimize the certainty when relying on the classifier prediction. By adopting this approach, we aim to enhance the overall expected prediction certainty, while selectively disregarding certain samples. This threshold may be varied by practitioners according to their needs. I.e. a high decoder threshold might be applied when analysing a cascade of images. On the other hand a lower threshold may be used to correct human analysis. When applying the model to the testset, we observe a BACC of 0.73, an F1 score of 0.74 and an MCC of 0.44.

E. MC Dropout

We employ MC Dropout during inference in DL models to estimate uncertainty. We do this by randomly deactivating neurons in the fully connected layer of the network. During multiple forward passes, we generate distributions of outputs instead of a single prediction. This way we introduce variability in the high-level features the network has learned. The variance across these multiple predictions serves as a measure of uncertainty. A high variance indicates low confidence in the prediction, and vice versa. This variance can be quantified for each image to provide an uncertainty score, reflecting the model's confidence in its prediction for that specific instance. Similar to the confidence estimation above, we derive the variance with MC Dropout across 50 predictions and define the threshold to be centered between its maximum and minimum. We observe a BACC of 0.73, an F1 score of 0.75 and an MCC of 0.44.

V. Conclusion

Interpreting medical images correctly is crucial for effective treatment. Experts utilize IVOCT with integrated automated plaque detection methods to analyze critical intravascular structures. However, current deep learning based methods in IVOCT lack precise uncertainty estimation, vital for accurate decision-making in medical diagnostics. Our study addresses this by presenting an analysis of confidence estimation in the context of IVOCT plaque detection using deep neural networks. We employ ResNet18 for classification tasks on cartesian images and employ data augmentation techniques tailored to this specific image representation. Additionally, we use transfer learning and tune the model on four folds to obtain a representative classifier performance. Moreover, we integrate an innovative uncertainty estimation method by extending the classifier with a Confidence Estimation Branch (CEB), measuring the models reconstruction error. This involves adapting and analyzing both loss functions to understand their interrelation and using the confidence score to assess uncertainty in individual samples and sample sets. Finally, we compare our performance with the Bayesian approach, MC Dropout. Our results demonstrate that incorporating a CEB into DL models improves their interpretability and robustness in analyzing IVOCT images and other medical imaging contexts. It paves the way more informed decision-making, particularly crucial

where accurate image predictions directly impact patient treatment and health. Based on this, additional analysis with on CEB training behaviour and filters to reduce noise may be performed. Furthermore, extensions on different models and medical image datasets may be considered.

ACKNOWLEDGMENT

This research is supported by LKCMedicine Start up grant funding from Ministry of Education, Singapore.

REFERENCES

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[2] R. Zhang, Y. Fan, W. Qi, A. Wang, X. Tang, and T. Gao, "Current research and future prospects of ivoct imaging-based detection of the vascular lumen and vulnerable plaque," *Journal of Biophotonics*, vol. 15, no. 5, p. e202100376, 2022.

[3] Z. Wang, M. W. Jenkins, G. C. Linderman, H. G. Bezerra, Y. Fujino, M. A. Costa, D. L. Wilson, and A. M. Rollins, "3-d stent detection in intravascular oct using a bayesian network and graph search," *IEEE transactions on medical imaging*, vol. 34, no. 7, pp. 1549–1561, 2015.

[4] A. Abdolmanafi, L. Duong, N. Dahdah, and F. Cheriet, "Deep feature learning for automatic tissue classification of coronary artery using optical coherence tomography," *Biomedical optics express*, vol. 8, no. 2, pp. 1203–1220, 2017.

[5] N. Gessert, M. Lutz, M. Heyder, S. Latus, D. M. Leistner, Y. S. Abdelwahed, and A. Schlaefer, "Automatic plaque detection in ivoct pullbacks using convolutional neural networks," *IEEE transactions on medical imaging*, vol. 38, no. 2, pp. 426–434, 2018.

[6] D. J. MacKay, "Probable networks and plausible predictions-a review of practical bayesian methods for supervised neural networks," *Network: computation in neural systems*, vol. 6, no. 3, p. 469, 1995.

[7] M. Magris and A. Iosifidis, "Bayesian learning for neural networks: an algorithmic survey," *Artificial Intelligence Review*, pp. 1–51, 2023.

[8] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[9] B. Ghoshal, A. Tucker, B. Sanghera, and W. Lup Wong, "Estimating uncertainty in deep learning for reporting confidence to clinicians in medical image segmentation and diseases detection," *Computational Intelligence*, vol. 37, no. 2, pp. 701–734, 2021.

[10] J. Zhang, Y. Xie, G. Pang, Z. Liao, J. Verjans, W. Li, Z. Sun, J. He, Y. Li, C. Shen *et al.*, "Viral pneumonia screening on chest x-rays using confidence-aware anomaly detection," *IEEE transactions on medical imaging*, vol. 40, no. 3, pp. 879–890, 2020.

[11] G. J. Tearney, E. Regar, T. Akasaka, T. Adriaenssens, P. Barlis, H. G. Bezerra, B. Bouma, N. Bruining, J.-m. Cho, S. Chowdhary *et al.*, "Consensus standards for acquisition, measurement, and reporting of intravascular optical coherence tomography studies: a report from the international working group for intravascular optical coherence tomography standardization and validation," *Journal of the American College of Cardiology*, vol. 59, no. 12, pp. 1058–1072, 2012.

[12] A. Gudigar, S. Nayak, J. Samanth, U. Raghavendra, A. AJ, P. D. Barua, M. N. Hasan, E. J. Ciaccio, R.-S. Tan, and U. Rajendra Acharya, "Recent trends in artificial intelligence-assisted coronary atherosclerotic plaque characterization," *International Journal of Environmental Research and Public Health*, vol. 18, no. 19, p. 10003, 2021.

[13] A. G. Roy, S. Conjeti, S. G. Carlier, P. K. Dutta, A. Kastrati, A. F. Laine, N. Navab, A. Katouzian, and D. Sheet, "Lumen segmentation in intravascular optical coherence tomography using backscattering tracked and initialized random walks," *IEEE journal of biomedical and health informatics*, vol. 20, no. 2, pp. 606–614, 2015.

[14] H. Lu, M. Gargesha, Z. Wang, D. Chamie, G. F. Attizzani, T. Kanaya, S. Ray, M. A. Costa, A. M. Rollins, H. G. Bezerra *et al.*, "Automatic stent detection in intravascular oct images using bagged decision trees," *Biomedical optics express*, vol. 3, no. 11, pp. 2809–2824, 2012.

[15] G. S. Mintz, "Intravascular imaging of coronary calcification and its clinical implications," *JACC: Cardiovascular Imaging*, vol. 8, no. 4, pp. 461–471, 2015.

[16] G. J. Ughi, T. Adriaenssens, P. Sinnaeve, W. Desmet, and J. D'hooge, "Automated tissue characterization of in vivo atherosclerotic plaques by intravascular optical coherence tomography images," *Biomedical optics express*, vol. 4, no. 7, pp. 1014–1030, 2013.

[17] C. Xu, J. M. Schmitt, S. G. Carlier, and R. Virmani, "Characterization of atherosclerosis plaques by measuring both backscattering and attenuation coefficients in optical coherence tomography," *Journal of biomedical optics*, vol. 13, no. 3, pp. 034 003–034 003, 2008.

[18] S. Celi and S. Berti, "In-vivo segmentation and quantification of coronary lesions by optical coherence tomography images for a lesion type definition and stenosis grading," *Medical image analysis*, vol. 18, no. 7, pp. 1157–1168, 2014.

[19] L. S. Athanasiou, C. V. Bourantas, G. Rigas, A. I. Sakellarios, T. P. Exarchos, P. K. Siogkas, A. Ricciardi, K. K. Naka, M. I. Papafaklis, L. K. Michalis *et al.*, "Methodology for fully automated segmentation and plaque characterization in intracoronary optical coherence tomography images," *Journal of biomedical optics*, vol. 19, no. 2, pp. 026 009–026 009, 2014.

[20] L. S. Athanasiou, M. L. Olender, J. M. de la Torre Hernandez, E. Ben-Assa, and E. R. Edelman, "A deep learning approach to classify atherosclerosis using intracoronary optical coherence tomography," in *Medical Imaging 2019: Computer-Aided Diagnosis*, vol. 10950. SPIE, 2019, pp. 163–170.

[21] H. Tang, Z. Zhang, Y. He, J. Shen, J. Zheng, W. Gao, U. Sadat, M. Wang, Y. Wang, X. Ji *et al.*, "Automatic classification and segmentation of atherosclerotic plaques in the intravascular optical coherence tomography (ivoct)," *Biomedical Signal Processing and Control*, vol. 85, p. 104888, 2023.

[22] H. Zhang, G. Wang, Y. Li, F. Lin, Y. Han, and H. Wang, "Automatic plaque segmentation in coronary optical coherence tomography images," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 33, no. 14, p. 1954035, 2019.

[23] S. He, J. Zheng, A. Maehara, G. Mintz, D. Tang, M. Anastasio, and H. Li, "Convolutional neural network based automatic plaque characterization for intracoronary optical coherence tomography images," in *Medical Imaging 2018: Image Processing*, vol. 10574. SPIE, 2018, pp. 800–806.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[26] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[27] J. Lee, D. Prabhu, C. Kolluru, Y. Gharaibeh, V. N. Zimin, L. A. Dallan, H. G. Bezerra, and D. L. Wilson, "Fully automated plaque characterization in intravascular oct images using hybrid convolutional and lumen morphology features," *Scientific reports*, vol. 10, no. 1, p. 2596, 2020.

[28] A. Abdolmanafi, F. Cheriet, L. Duong, R. Ibrahim, and N. Dahdah, "An automatic diagnostic system of coronary artery lesions in kawasaki disease using intravascular optical coherence tomography imaging," *Journal of Biophotonics*, vol. 13, no. 1, p. e201900112, 2020.

[29] Y. A. Bayhaqi, A. Hamidi, F. Canbaz, A. A. Navarini, P. C. Cattin, and A. Zam, "Deep learning models comparison for tissue classification using optical coherence tomography images: toward smart laser osteotomy," *OSA Continuum*, vol. 4, no. 9, pp. 2510–2526, 2021.

[30] L. R. Wang, T. C. Henderson, and X. Fan, "An uncertainty estimation model for algorithmic trading agent," *International Conference on Intelligent Autonomous Systems*, no. 18, pp. 4–7, 2023.

[31] H. S. Ahmed and M. J. Nordin, "Improving diagnostic viewing of medical images using enhancement algorithms," *Journal of Computer Science*, vol. 7, no. 12, p. 1831, 2011.

[32] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC genomics*, vol. 21, no. 1, pp. 1–13, 2020.

[33] I. Cohen, Y. Huang, J. Chen, J. Benesty, J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," *Noise reduction in speech processing*, pp. 1–4, 2009.