Active Multi-Task Learning with Uncertainty Weighted Loss for Coronary Calcium Scoring

Bernhard Föllmer[a)], **Federico Biavati, Christian Wald**
Department of Radiology, Charité-Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany
Sebastian Stober
Artificial Intelligence Lab, Otto-von-Guericke-Universität, Magdeburg, Germany
Jackie Ma
Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, Berlin, Germany
Marc Dewey[b)]
Department of Radiology, Charité-Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany
Berlin Institute of Health and DZHK (German Centre for Cardiovascular Research), partner site Berlin, Germany
Wojciech Samek[b)]
Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, Berlin, Germany

Version typeset February 10, 2022
bernhard.foellmer@charite.de

Abstract

Purpose: Coronary artery calcium (CAC) scoring is an independent marker for the risk of coronary heart disease events. Automatic methods for quantifying CAC could reduce work load and assist radiologists in clinical decision making. However, large annotated datasets must be acquired and labeled to achieve very good model performance, which is an expensive process and requires expert knowledge. By labeling only most informative samples, active learning can reduce the number of training data required. Multi-task learning techniques can improve model performance by joint learning of multiple related tasks and extraction of shared informative features.
Methods: We propose an uncertainty weighted multi-task model for coronary calcium scoring in ECG-gated, non-contrast enhanced cardiac calcium scoring CT. The model is trained to solve the two tasks of coronary artery region segmentation (weak labels) and coronary artery calcification segmentation (strong labels) simultaneously in an active learning scenario to improve model performance and reduce the number of required training samples. We compare our model with a single-task U-Net and a sequential-task model, as well as other state-of-the-art methods. The model is evaluated based on 1275 individual patients of three different datasets (DISCHARGE, CADMAN, orCaScore) and the relationship between performance and various influencing factors (image noise, metal artifacts, motion artifacts, image quality) is analyzed.
Results: The joint learning of multiclass coronary artery region segmentation and binary coronary calcium segmentation improves calcium scoring performance. Since shared complementary information can be learned from both tasks, the model reaches optimal performance with only 12% of the training data and one-third of the labeling

time in an active learning scenario. We identified image noise as one of the most important factors influencing model performance, along with anatomical anomalies and metal artifacts.

Conclusions: Our multi-task learning approach with uncertainty weighted loss improves calcium scoring performance by joint learning of shared features and reduces labeling costs when trained in an active learning scenario.

Keywords: Coronary Artery Calcium Scoring, Deep Learning, Neural Networks, Active Multi-Task Learning, Uncertainty Weighted Loss

Abbreviations

CAC Coronary artery calcification.
CAR Coronary artery region.
CCTA Contrast enhances CT.
CSCT Calcium Scoring CT.
CT Computed tomography.
CVD Cardiovascular disease.
HU Hounsfield unit.
ICC Interclass correlation coefficient.
LAD Left anterior descending artery.
LCX Left circumflex artery.
LM Left main artery.
MTL Multi-task learning.
PPV Positive predictive value.
RCA Right coronary artery.

I. Introduction

Cardiovascular disease (CVD) is the global leading cause of death.[1] Coronary calcium is commonly associated with coronary atherosclerosis and its absence is associated with a very low risk of adverse coronary events.[2] In clinical practice, semi-automatic software is used to manually select coronary artery calcifications (CAC) in computed tomography (CT) image slices from automatically labeled candidates, which is a tedious and time-consuming process in large-scale studies.[3] Typically, ECG-gated, non-contrast enhanced computed tomography, known as calcium scoring computed tomography (CSCT), is used to identify CAC.[4] The Agatston score[5] is the most common measure used to quantify CAC with the aim of defining appropriate cardiac risk categories. In recent years, deep learning models such as convolutional neural networks have been used to automatically quantify CAC based on 2D slices [3,4] or 2.5D/3D volumetric input data [6,7].

Methods have been developed for different examination types such as non-contrast enhanced ECG-gated calcium scoring CTs,[8] contrast-enhanced coronary CT angiography

(CCTA)[9] or a combination of both.[3,10,11] Since the segmentation of the cardiac tree is very challenging in CSCT, methods using non-contrast enhanced and contrast enhanced CT usually map spatial information about coronary arteries from the contrast enhanced CT to the non-contrast enhanced CT.[10,12] Most methods perform segmentation of the calcified lesion to estimate the Agatston score and classify detected calcification based on the corresponding left anterior descending artery (LAD), left circumflex artery (LCX), and right coronary artery (RCA), but some also perform regression directly.[13]

Most of the current state-of-the-art methods only learn from sparse calcifications. Therefore, very large and heterogeneous datasets need to be acquired and labeled to train models that are robust and achieve satisfactory performance to be used in a clinical setting. Unfortunately, this is an expensive and time consuming process and requires expert knowledge. It calls for methods that can reduce labeling costs and improve performance by integrating the radiologist into the training process.

Active learning techniques are able to reach higher performance while using a smaller number of annotated training samples by active sample selection and therefore reduce labeling costs. In active learning, the learner (deep neural network) iteratively selects only the most informative samples based on a selection criterion such as uncertainty sampling, query by committee, expected error reduction, or expected model for labeling.[14] The method integrates the radiologist into the training process and avoids labeling of uninformative samples.

Spatial information about coronary arteries and corresponding coronary calcifications is very important to distinguish between coronary and extra-coronary calcifications. Calcifications are usually very sparse, which makes it difficult to extract features with spatial information. Extraction of spatial information about the coronary arteries in an auxiliary task can aid in the localization of coronary calcifications. Multi-task learning (MTL) is a technique which learns multiple related tasks together, to improve model performance by sharing complementary information.[15] In coronary calcium scoring, the spatial information of coronary arteries is closely related to the calcium scoring task and therefore learning of coronary artery regions serves as a good auxiliary task to support the original calcium scoring task. However, the optimization of multiple loss functions for multi-task learning is a crucial factor and tuning loss weighting by hand is difficult and computationally expensive. Many task balancing approaches for dense predictions such as static weighting, GradNorm[16], dynamic weight average[17], dynamic task prioritization[18] or uncertainty weighted loss[19] have been developed

and have shown that the best optimization method should be selected on a per case basis.[20] The training of multi-task models in an active learning scenario can be challenging if the dataset is very small. In this work, we exploit a multi-task model (MTL-model) with uncertainty-weighted loss that outperforms a single-task U-Net and a sequential-model. The model achieves very good performance on small training sets and can therefore be used in active learning scenarios. The model performs as well as other state-of-the-art methods and achieves similar results compared to our statically-weighted MTL-model with optimally chosen weighting parameter. The contributions of this paper can be summarized as follows:

- We propose a novel learning paradigm for coronary calcium scoring by simultaneous learning of multiple related tasks to increase data efficiency and model performance by leveraging auxiliary information through shared informative features.
- We propose a multi-task encoder-decoder model for simultaneous coronary artery regions segmentation (multiclass) and coronary artery calcification segmentation (binary) to improve model performance compared to single-task models.
- We show that our model obtains optimal performance with substantially less training data (12%) and reduces annotation time to one-third in an active learning scenario compared to training on the full dataset.
- We demonstrate the importance of loss weighting for optimal model performance of our multi-task model and show how uncertainty weighted loss can facilitate active multi-task learning.
- We show that our model performs almost as well as the best state-of-the-art methods in terms of F1-score, intraclass correlation coefficient (ICC) and sensitivity of CAC volume, on a common benchmark dataset for coronary calcium scoring.

II. Materials and methods

In this Section we present our multi-task model which performs simultaneous segmentation of coronary artery regions and segmentation of coronary artery calcifications. In Section II.A. we describe the used datasets and corresponding annotation strategies. We introduce the multiclass coronary artery region segmentation task II.B.1. and the binary lesion segmentation task II.B.2.. We propose the multi-loss optimization method using uncertainty weighted loss II.B.3. and give a detailed description of our implemented network architecture and training

procedure II.B.4.. In Section II.B.5. we introduce a single-task U-Net and sequential model as comparison models. In Section II.C. we introduce our active learning approach in which we use only the most informative samples to decrease annotation costs and propose our hybrid sampling strategy.

II.A. Datasets

For the evaluation of our multi-task learning approach, we compare the performance on three different datasets. A detailed flowchart about the dataset selection process can be found in the supplementary material.

The DISCHARGE-trial is a prospective multi-center randomized controlled trial to examine for which patients with suspected coronary artery disease based on stable chest pain, cardiac CT or cardiac catheterization is best suited.[21,22] Our DISCHARGE dataset consists of calcium scoring CTs (CSCT) from 1262 patients (708 male, 554 female) of the trial. Image data were acquired from 26 clinical sites using 14 different scanner types. Annotations for coronary artery calcification were acquired for all scans. Weak annotations of coronary artery regions were only acquired for 215 randomly selected scans and randomly divided into 140 CT scans (6721 slices) for training (65%) and 75 CT scans (3636 slices) for validation (35%). All remaining 1047 CT scans (57452 slices) were used as test set. Only one CSCT from each patient was selected for the dataset. The reconstruction of the CT scans was performed using filtered back projection method (383) and iterative reconstruction methods (879). To keep the data as close as possible to real-life clinical data, diagnostic CTs including scans with metal artifacts (pace-maker, artificial valves, etc.), scans with severe motion artifacts, high level of noise or anatomical abnormalities were not excluded. Note, annotations of coronary artery regions (CAR) and coronary artery calcifications (CAC) are available for training and validation set. For the test set, only CAC annotations are available. The annotations were performed by two observers. Observer one was a trained physician who annotated coronary calcifications, observer two was a trained medical imaging scientist who annotated coronary artery regions. Available contrast-enhanced CT scans (CCTA) were not included because the overall goal of the method is to predict coronary heart disease risk without the need to inject a contrast agent.

The second dataset consists of CT scans from the publicly available orCaScore challenge

on (semi-)automatic coronary calcium scoring.[23] The framework provides 72 pairs of CSCT and corresponding contrast-enhanced CT angiography (CCTA) from the same patient acquired at four academic hospitals. The data has been divided into a 32-scan training set and a 40-scan test set. For the training set, a reference standard by two expert observers, a radiologist with > 12 years of experience in CAC scoring and a research physician, are provided. CT scans with anatomical abnormalities, intracoronary stents, and metal implants as well as CTs showing severe motion artifacts or extremely high levels of noise determined by visual inspection were excluded. Annotations of CAR were additionally acquired for the training set.

The third dataset consists of CSCT from the single-center randomized controlled Coronary Artery Disease Management (CAD-Man) study.[24] The dataset consists of 156 CT scans and annotations were only acquired for CAC. The dataset serves as an additional test set. The reconstruction of the CT scans was performed using filtered back projection method. Reference standards are provided by one expert observer. The annotations of coronary calcifications and artery regions were performed by a trained medical imaging scientist.

Differences between the three datasets regarding distribution of candidate lesions are shown in Table 1. A candidate lesion is defined as connected 3-D image voxels (6-connectivity) with intensities greater than 130 HU.

	# scans	LAD	LCX	RCA	OTHER_CAR	Candidates per scan
DISCHARGE Training	140	344	168	338	865k	6183
DISCHARGE Test	1047	2375	1042	1872	6254k	5978
DISCHARGE Validation	75	198	118	221	432k	5768
orCaScore Training	32	103	21	56	138k	3454
orCaScore Test	40	-	-	-	-	-
CADMAN Test	156	335	151	156	1400k	8980

Table 1: Number of candidate lesions (connected 3-D image voxels with intensities greater than 130 HU) distributed in the DISCHARGE training set, validation set and test set, orCaScore training set and CADMAN test set. Candidate lesions are distributed over calcified coronary artery lesions (LAD, LCX, RCA) and other structures such as bones or extra-coronary calcifications (OTHER_CAC). Since the orCaScore test set is not public, no information about distribution of candidate lesions is available.

II.A.1. Annotation procedure

The coronary artery tree is divided into three sub-trees corresponding to the left anterior descending artery (LAD), left circumflex artery (LCX) and right coronary artery (RCA) including main branch (LM) and its side-branches. The left main artery (LM) is included in the sub-tree of the LAD. The training of the multi-task model requires annotations for the two task of coronary artery calcification (CAC) segmentation and coronary artery region (CAR) segmentation.

II.A.2. Annotation of coronary artery calcifications (CAC)

The annotation of the CAC was performed by thresholding and highlighting all voxels with a density above 130 HU (candidate lesions). The observer annotates all highlighted voxels of calcified lesions and assigns a class according to the corresponding coronary artery ("LAD", "LCX", "RCA"). Calcified lesions corresponding to multiple arteries (e.g. calcified lesions in bifurcations) were divided by annotating the voxels according to the arteries. For model evaluation, lesions where defined as connected voxels (6-connectivity) with a minimum lesion volume of $1.5mm^3$ All remaining candidate lesions were annotated as "OTHER_CAC". The annotations of coronary artery calcifications were performed using an in-house developed semi-automatic segmentation module for 3D slicer [25].

II.A.3. Weak annotation of coronary artery regions (CAR)

The annotations of CARs are acquired using weak annotations (scribbles), since a precise segmentation of the vessel tree is impossible in non-contrast CTs due to the missing contrast between arteries and surrounding tissue. To overcome the problem of misleading labels, we did not label regions between arteries and surrounding tissue which are difficult to distinguish or a precise labeling of the boundary would be extremely time-consuming. To facilitate and speed up the annotation process, we used an in-house semi-automatic segmentation module developed for 3D Slicer [25]. At first, the annotator was using a scribble to annotate the three main arteries in each slice. In the second step, an additional scribble (closed contour) was used to surround the arteries and isolate the annotated artery scribble from the tissue. In the third step, connected component analysis was performed to divide the annotations

into different components. The largest component (background) was joint with the closed contour scribbles and labeled as OTHER_CAR. If no coronary artery was seen in the slice, the annotator only placed a single scribble for OTHER_CAR in the image. Examples of the performed annotations can be seen in Figure 2.

II.B. Multi-task segmentation network with uncertainty weighted loss

We propose a multi-task segmentation network following an encoder-decoder structure with skip connections, inspired by the U-Net architecture.[26,27] The multi-task network architecture is illustrated in Figure 1 and performs multiclass coronary artery region (CAR) segmentation and coronary artery calcifications (CAC) segmentation at the same time. Features extracted by the encoder are shared with the two decoders for the tasks of multiclass coronary artery region segmentation (T_R) and binary segmentation of calcified lesions (T_L). Since the information of predicted coronary artery regions is a useful prior information for segmentation of calcified lesions, feature maps extracted by the decoder for coronary artery region segmentation are shared with the decoder for binary segmentation of calcified lesions. To utilize this prior information about candidate lesions, we concatenate the image slice (512 px x 512 px) with a candidate lesion mask to form the input tensor. The candidate lesion mask was created by thresholding the image using a constant threshold of 130 HU. During training, the losses of both task L_R (loss for task T_R) and L_L (loss for task T_L) are combined using an uncertainty weighting loss,[19] to jointly optimize the model parameters.

II.B.1. Coronary artery region segmentation task

The network aims to learn coronary artery regions from weakly labeled regions, as shown in Figure 2. Weak labels[28] are defined as segmentations, which are imprecise but less costly to obtain than pixel-level annotations. Since in non-contrast enhanced CT scans the spatial boundary between the vessels and the surrounding tissue cannot be determined precisely, pixels x of an image i are either annotated and belong to the annotated pixel set $\Omega_{R,i}$ with one of the CAR classes $\{LAD, LCX, RCA, OTHER\_CAR\}$, or are not annotated. The pixel-wise softmax function [26] and focal loss [29] are used to deal with large class imbalance

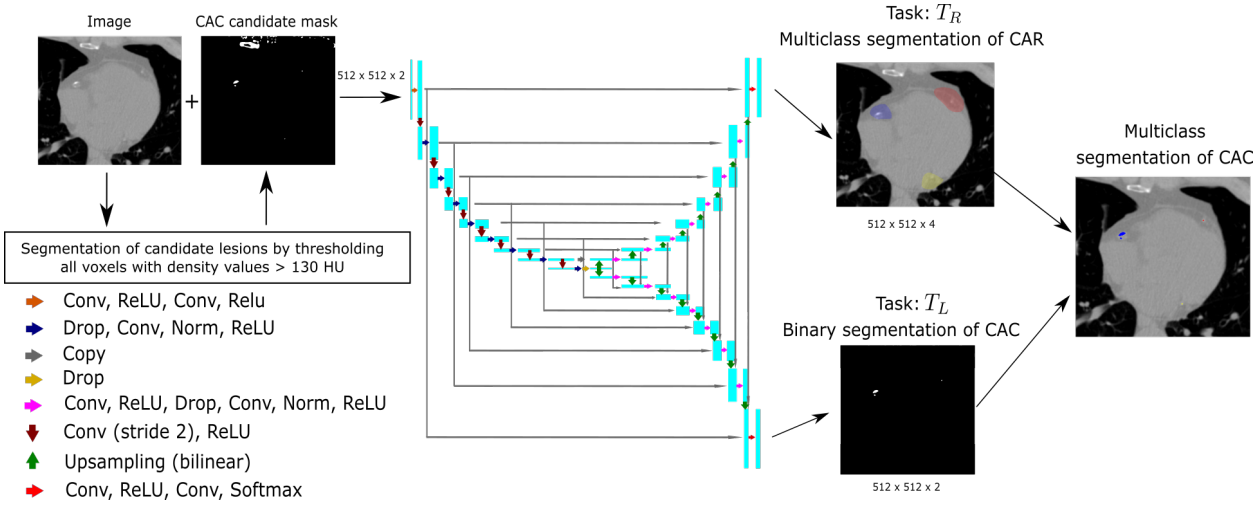


Figure 1: Multi-task model for coronary artery calcification (CAC) scoring. The image and the CAC candidate lesion mask are concatenated to form the input tensor. The model consists of one encoder that shares feature maps with two decoders of the multiclass coronary artery region (CAR) segmentation task T_R and binary CAC segmentation task T_L. Predictions are combined by multiplying binary CAC segmentation with multiclass CAR segmentation to perform multiclass calcification segmentation

between background pixels ($OTHER\_CAR$) and pixels of coronary artery regions.

$$\mathcal{L}_{R,i} = \sum_{x \in \Omega_{R,i}} \sum_{c_R=1}^{4} -w_{c_R} y_{c_R}(x)(1 - p_{c_R}(x))^{\gamma_R} \log(p_{c_R}(x)) \tag{1}$$

The γ_R parameter smoothly adjusts the rate at which easily segmented pixels are down-weighted and w_R balances the loss. $p_{c_R}(x)$ and $y_{c_R}(x)$ are the pixel-wise softmax output and the reference class of pixel $x \in \Omega_{R,i}$, respectively. The pixel set $\Omega_{R,i}$ contains all labeled pixels. Unlabeled pixels (gaps) $x \notin \Omega_{R,i}$, as shown in Figure 2, are ignored and not used for loss calculation. The parameter w_{c_R} is a weighting parameter that balances the importance of the classes and handles the data imbalance problem. The parameter c_R is the channel of the corresponding CAR class.

II.B.2. Binary lesion segmentation task

The lesion segmentation network performs a binary segmentation of candidate coronary artery lesions into the classes $\{CAC, OTHER\_CAC\}$. Feature maps extracted by the decoder for the coronary artery region segmentation are shared with the decoder for the binary

lesion segmentation. The binary focal loss $\mathcal{L}_{L,i}$ defined as (2) is calculated based on all voxels from candidate lesions grouped in the set $\Omega_{L,i}$.

$$\mathcal{L}_{L,i} = \sum_{x \in \Omega_{L,i}} \sum_{c_L=1}^{2} -w_{c_L} y_{c_L}(x)(1 - p_{c_L}(x))^{\gamma_L} \log(p_{c_L}(x)) \tag{2}$$

The parameters $p_{c_L}(x)$, $y_{c_L}(x)$ as well as $x \in \Omega_{L,i}$ and w_{c_L} are defined analogouesly to the region segmentation in Subsection II.B.1.. The output of the binary CAC segmentation decoder is multiplied (channel-wise) with the output of the CAR segmentation decoder to perform multiclass CAC segmentation.

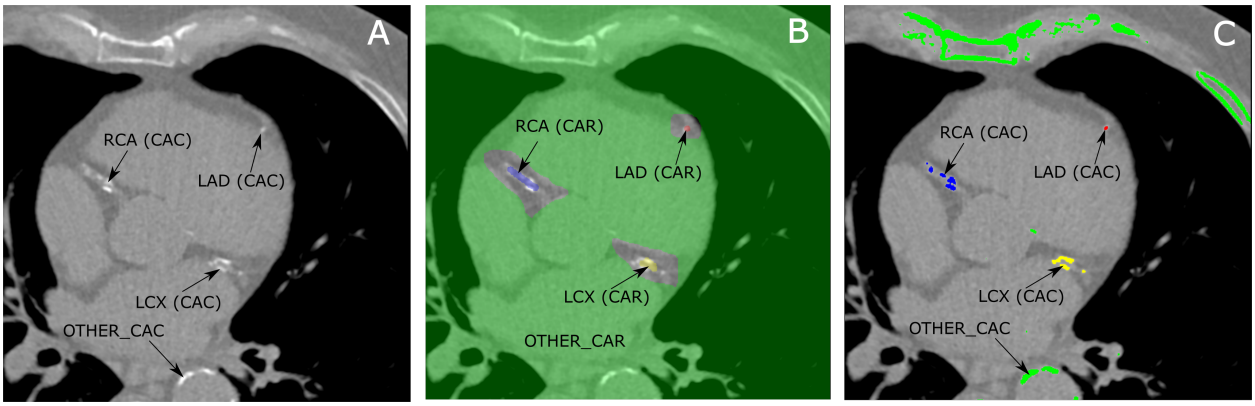


Figure 2: Multi-task annotations of an image slice with coronary artery calcifications (CAC) in the left anterior descending artery (LAD), left circumflex artery (LCX) and right coronary artery (RCA) (A). Weak annotations of coronary artery regions (CAR) for the LAD - red, LCX - yellow, RCA - blue and OTHER_CAR - green (B). Strong annotations of coronary artery calcifications in the LAD - red, LCX - yellow, RCA - blue and other objects with density higher 130 HU (OTHER_CAC) - green (C).

II.B.3. Uncertainty based weighted loss

The performance of a multi-task network depends strongly on the weighting of the losses. The most commonly used loss weighting strategy for multi-task learning is static weighting, which computes a weighted sum of the losses using balancing parameters α_i. The static weighted loss of our multi-task model is the weighted sum of the losses for multiclass segmentation of coronary artery regions $\mathcal{L}_R$ and the binary segmentation of coronary calcifications $\mathcal{L}_L$, as shown in Equation (3).

$$\mathcal{L}_{total}(\mathbf{W}) = \alpha \mathcal{L}_R(\mathbf{W}) + (1 - \alpha)\mathcal{L}_L(\mathbf{W}) \tag{3}$$

This method is simple but unfortunately computationally expensive to fine tune.[30] The determination of the optimal weighting parameter value α is even more challenging in active multi-task learning, since the model is initially trained with a very small number of annotated training data. Other methods are based on dynamic weight average (DWA) using task-specific feature-level attention[17] or use gradient normalization[16] to balance losses.
For our multi-task calcium scoring model, we use the uncertainty weighted loss method of Cipolla et al.[19]. The uncertainty based weighting uses homoscedastic uncertainty to weight loss functions of each task.[19] We combine the outputs of the last layers (softmax output) from the decoders based on the homoscedastic uncertainty. To model the uncertainty, we introduce the positive scalar σ_R for coronary artery region segmentation task and σ_L for the binary calcification segmentation task. The parameters can be interpreted as Boltzmann distributions (also called Gibbs distribution) where the input is scaled by σ_R^2 and σ_L^2 respectively. The total loss $\mathcal{L}_{total}$ in Equation (4) is an uncertainty weighted loss of $\mathcal{L}_R$ and $\mathcal{L}_L$ where $\mathbf{W}$ represents the parameters of the multi-task network. A detailed deviation can be found in the supplementary material.

$$\mathcal{L}_{total}(\mathbf{W}, \sigma_R, \sigma_L) = \frac{1}{\sigma_R^2}\mathcal{L}_R(\mathbf{W}) + \frac{1}{\sigma_L^2}\mathcal{L}_L(\mathbf{W}) + \log \sigma_R + \log \sigma_L \tag{4}$$

This loss is smoothly differentiable, and is well formed such that the task weights will not converge to zero. For practical reasons, we predict the log variance $\log \sigma^2$, which is more stable and avoids any division by zero.[19]

II.B.4. Multi-task network architecture and training procedure

To train the multi-task network, we oversample slices with calcifications to form balanced mini batches (20% samples with calcifications, 80% without calcifications). The encoder consists of eight downsampling blocks, each block consists of two convolutional layers, dropout layer, batch normalization [31] and ReLU activation function.[32] The two decoders consist of eight upsampling blocks, where each block consists of a bilinear upsampling layer and two convolution layers, dropout layer, batch normalization layer and ReLU activation function, respectively. The feature maps of the upsampling block of the coronary artery region (CAR) segmentation decoder are shared with each upsampling block of the coronary calcification

segmentation decoder, but not vice versa, to follow the causal relation between coronary artery regions and coronary artery calcifications. Skip connections between the encoder and the two decoders are implemented as concatenations and used to share feature maps from the respective downsampling block.[26] The model was trained with a batch size of 8, the Adam [33] optimizer, an initial learning rate of 5e-04, learning rate decay of 0.95 after every 5 epochs and L2 weight decay. During training, the dropout rate of the inner layer between the encoder and decoder was set to 0.5, and the dropout rate for all other layers was set to zero. During our experiments, we found that the convergence of the static weighted loss MTL-model depends strongly on the initial learning rate but due to high training times, we did not perform further detailed hyper parameter analysis. We use focal loss for both training tasks. The focal loss parameters of task T_R were set to $\gamma_R = 2.0$, $\alpha_{OTHER\_CAR} = 0.01$, $\alpha_{LAD} = 1.0$, $\alpha_{LCX} = 1.0$, $\alpha_{RCA} = 1.0$. The loss parameters of the coronary calcification segmentation task were set to $\gamma_L = 2.0$, $\alpha_{CAC} = 1.0$, $\alpha_{OTHER\_CAC} = 0.01$. To train the network, we augmented the image slices by small translations to prevent overfitting. To perform multiclass calcification segmentation, the predictions of the two tasks are combined by multiplying the binarized calcification segmentation with the coronary artery region segmentation. To avoid overtraining, we use early stopping based on the performance of the validation set. The training stopped after approximately 200k iterations (240 epochs), where one iteration corresponds to a batch of 8 slices. Training was performed on an NVIDIA Tesla V100, 32GB and PyTorch framework. More details and a pretrained model can be found at (https://github.com/Berni1557/MTAL-CACS).

II.B.5. Single-task model, sequential-task model and multi-task model

We compare our multi-task model with a single-task and a sequential-task model in Figure 3 to show that simultaneous training of related tasks can extract informative shared features and improve model performance. The single-task model consists of a multiclass U-Net[26] with the same downsampling and upsampling block architecture as in the multi-task network. The last layer consists of four channels ($OTHER\_CAC$, LAD, LCX, RCA) for multiclass segmentation of coronary calcifications. The sequential model consists of two separated models. The first model is trained for multiclass coronary artery region (CAR) segmentation. After the training has finished, the predictions are used for the training of the coronary calcification (CAC) segmentation network. Therefore, the CAR predictions are concatenated

with the image and CAC candidate mask and serve as input for the binary segmentation network for coronary calcifications. The goal of the sequential model is to follow the causal relation between CAR and CAC.

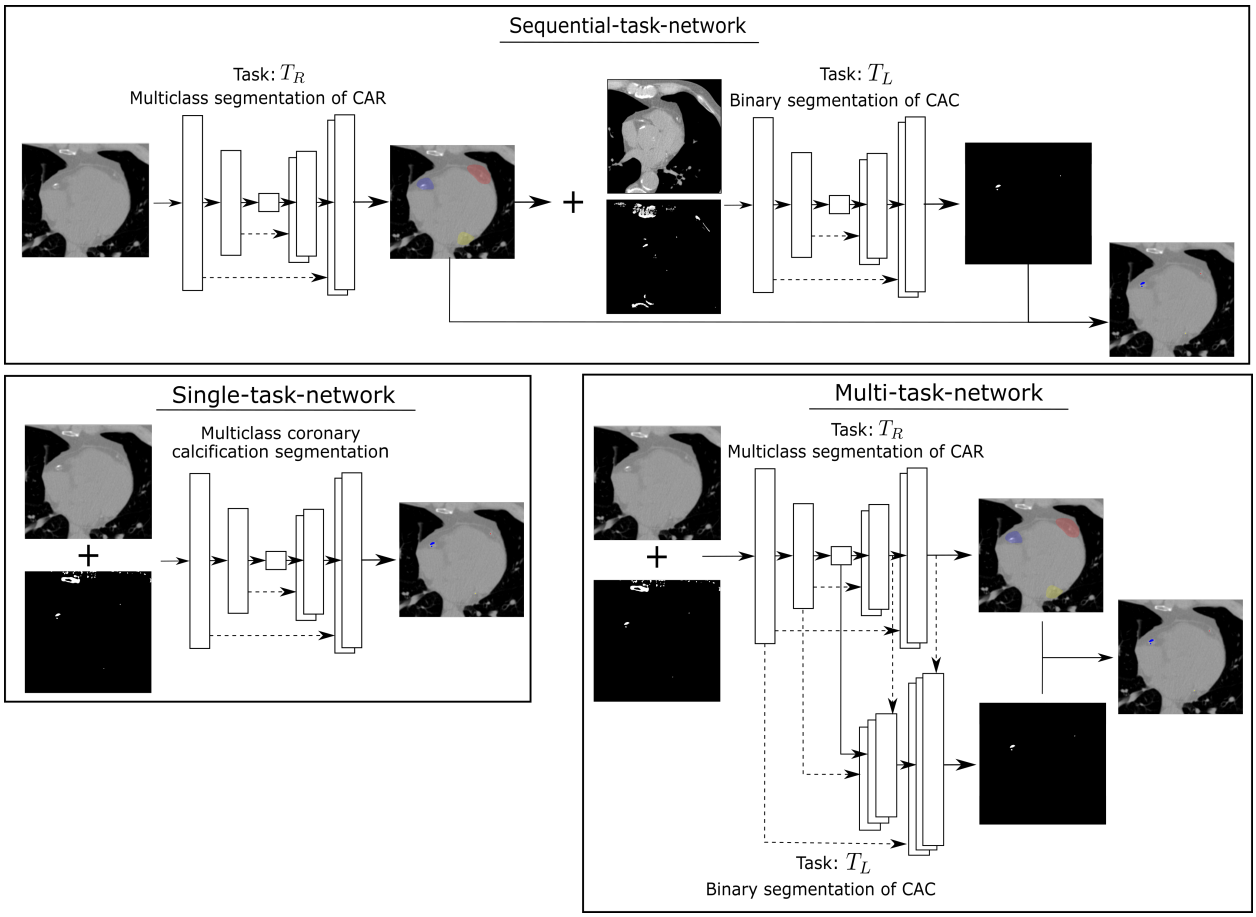


Figure 3: Single-task model, sequential-task model and multi-task model architecture comparison. The single-task model (lower, left) consists of a multiclass U-Net. The sequential model (upper) consists of a model for multiclass coronary artery region (CAR) segmentation whose predictions are used to train the coronary artery calcification (CAC) segmentation network. The multi-task model (lower, right) consists of one encoder and two decoders for the prediction of coronary artery region (CAR) segmentations and coronary calcification (CAC) segmentations which are combined for multiclass segmentation of CAC.

II.C. Active learning with uncertainty weighted multi-task model

Labeling of coronary calcifications in CT scans is a laborious and time-consuming task and requires significant expert knowledge.[34] Labeling for MTL methods tends to be more expensive since each task requires its own annotations. Active learning is able to reduce the costs

by iteratively labeling only most informative samples thus achieving optimal performance with a smaller number of samples.[35] In multi-task learning with static weighting parameter α, the best performing parameter has to be determined, which is a difficult and expensive process[19] and is often performed using grid search on the entire annotated dataset. In active learning, the estimation of a static weighting parameter α is even more challenging to tune, since the data distribution changes after each sampling round and thereby the optimal value of parameter α changes as well. Moreover, the estimation on small datasets can be very sensitive to the randomly drawn initial training samples.

We simulate active learning to investigate whether our uncertainty weighted loss model can overcome these problems. There are several approaches for active multi-task learning such as active learning via bandits,[36] active learning frameworks for adaptive filtering[37] or value of information based methods.[38]

For our approach, we developed a sampling strategy based on uncertainty sampling and random sampling, which we call the hybrid sampling strategy. First, we apply monte carlo dropout (MCD) [39] during inference for all samples which are not in the training set, predict segmentation maps and repeat this process $N_{MCD} = 10$ times. Dropout rate was set to 0.01 for all dropout layers. Based on the predictions, we estimate the MC sample variance[40] for each pixel, corresponding to candidate calcifications (pixel with density values greater than 130 HU) and calculate the average variance for each sample. We sort all samples in descending order and randomly sample from the top 20% with highest variance. Selected samples and respective annotations are added to the training set. We use this simple strategy, since it is not our goal to improve sampling strategies, but rather to investigate their general applicability. This sampling strategy has low computational complexity and increases the diversity of batch query samples.[35]. We compare our hybrid sampling method with the random sampling method, where we randomly select samples from the unlabeled dataset for labeling and training.

III. Results

In this Section we first introduse used performance metrics compare the performance of our proposed multi-task model with the singe-task U-Net and sequential-task model trained on the full DISCHARGE training set III.B.. In Subsection III.C., we compare our multi-task model with other state-of-the-art models. In Subsection III.D., we analyze our uncertainty weighted multi-task model in an active learning scenario and show that the number of required training samples and annotation time can be reduced compared to labeling the full training set. Finally, we analyze the influence of image noise, metal artifacts, motion artifacts and image quality on model performance in Subsection III.E..

III.A. Performance metrics

The performance of our models for coronary calcium scoring in Table 2 was evaluated on volume and lesion level with binary and multiclass segmentation metrics[41]. For the evaluation of the multiclass coronary artery region segmentation task T_R, we use the Micro F1-score on volume level. The Micro F1-score is the harmonic mean of Micro precision and Micro recall based on the the coronary arteries, excluding the $OTHER\_CAR$ class (5). For the Micro precision and Micro recall, the number of true positives (TP_{sum}) is the number of all correctly classified pixels of the coronary artery regions, excluding pixels of the class $OTHER\_CAR$. The number of false positives (FP_{sum}) is the number of pixels belonging to the class $OTHER\_CAR$ but being misclassified as one of the coronary arteries, plus all pixels of coronary arteries that are incorrectly assigned to another artery. The number of false negatives (FN_{sum}) is the number of pixels belonging to the coronary arteries but being misclassified as $OTHER\_CAR$, plus all pixels of coronary arteries that are incorrectly assigned to another artery. Therefore, misclassifications between arteries are counted as false negatives and false positives.

$$\text{Micro F1-score} = 2 * \frac{\text{Micro-precision} * \text{Micro-recall}}{\text{Micro-precision} + \text{Micro-recall}} \tag{5}$$

$$\text{Micro-recall} = \frac{TP_{sum}}{TP_{sum} + FN_{sum}} \tag{6}$$

$$\text{Micro-precision} = \frac{TP_{sum}}{TP_{sum} + FP_{sum}} \tag{7}$$

The evaluation of the binary coronary calcification task T_L was evaluated based on the positive predictive value (PPV), sensitivity and F1-score. The evaluation of the resulting multiclass calcification segmentation was evaluated based on the F1-score calculated irrespective of the artery-specific label, to be comparable with other methods. For the comparison with other methods, we use to intraclass correlation (ICC), sensitivity and F1-score in Table 4 and 3.

To evaluate our active learning method in Figure 5, we used the Micro-F1 score of the resulting multi-class calcification segmentation.

The risk categorization performance in Table 5 was evaluated based on the linearly weighted Cohen's kappa as a measure of agreement between the reference category and the categorization based on the MTL-model.

III.B. Comparison of single-task, sequential-task and multi-task model

We trained all three models described in Section II.B.5. on the full DISCHARGE training set and evaluate the performance based on the DISCHARGE test set. In Table 2 we compare results for coronary artery region (CAR) segmentation task T_R in terms of Micro F1-score and binary coronary calcification (CAC) segmentation task T_L in terms of F1-score, positive predictive value and sensitivity. The Micro F1-score is reported for the resulting multiclass calcification segmentation. For T_R we report Micro F1-scores only for the validation set, since annotations of the DISCHARGE test set were not available for CAR. For the static-MTL model we set weighting parameter to the optimal value $\alpha = 0.4$, determined based on the maximum Micro F1-score for calcification segmentation using grid-search method. Uncertainty weighted loss MTL-model and static weighted MTL-model with optimal weighting parameter value reached similar performance of F1-score=0.881 and F1-score=0.882, respectively. Both MTL-models (static weighted MTL and uncertainty weighted loss MTL) outperform the single-task model (F1-score=0.804) and sequential model (F1-score=0.769). The performance of the multi-task models is very good at the volume level, but lower at lesion level, due to a false positive predictions caused by misclassification of noise. As expected,

	CAR segmentation task T_R	Binary calcification segmentation task T_L			Multiclass calcification segmentation
	Micro F1 (Vol)	PPV (Vol/Num)	Sen. (Vol/Num)	F1 (Vol/Num)	F1 (Vol)
Single-task (U-Net)	-	0.900 (0.219)	0.726 (0.923)	0.804 (0.353)	0.775
Sequential-task model	**0.472**	0.916 (0.231)	0.663 (0.871)	0.769 (0.365)	0.740
Static weighted MTL-model ($\alpha = 0.4$)	0.459	**0.937** (0.412)	0.833 (0.883)	**0.882** (0.562)	**0.850**
Uncertainty weighted loss MTL-model	0.451	0.924 (0.413)	**0.842** (0.880)	0.881 (0.562)	0.849

Table 2: Performance comparison of the single-task model (U-Net), sequential-task model, static weighted MTL-model and uncertainty weighted loss MTL-model. Evaluation is based on Micro F1-score for coronary artery region segmentation task T_R (only available for DISCHARGE validation dataset) and F1-score, positive predictive value (PPV) and sensitivity (Sen.) for binary calcification segmentation task T_L and Micro F1-score for the combined multiclass calcification segmentation of the DISCHARGE test dataset on volume and lesion level.

the best performing model for coronary artery region segmentation task T_R is the sequential model (Micro F1-score=0.472), since the first of the two sequential networks performs only this task. Figure 4 shows an example for the predictions of CAR and CAC by the multi-task network. We compare the multi-task predictions for severe noise in Figure S3 (supplementary material) to show how the uncertainty weighted loss MTL-model outperformed the sequential model.

III.C. Performance comparison with other methods

To compare our model (uncertainty weighted loss MTL-model) against other methods, we evaluate the performance based on the orCaScore test set described in Section II.A.. The orCaScore dataset does not provide any reference annotations for the test set and is therefore well suited for model comparison. For a fair comparison we trained our model twice: once on the DISCHARGE training set and once on the orCaScore training set. Results on the orCaScore test set compared with other methods are shown in Table 3. The results of

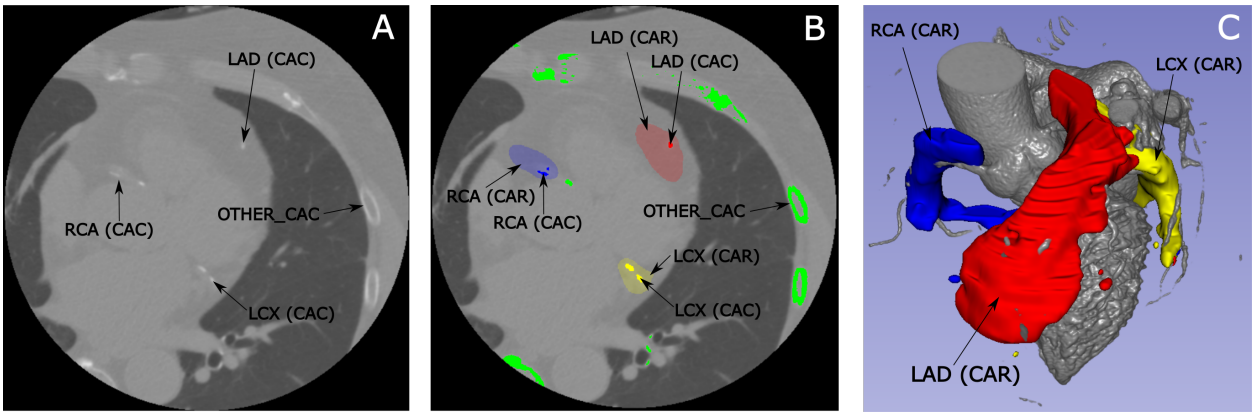


Figure 4: Visualization of overlap between predicted coronary calcifications (CAC) and coronary artery regions (CAR). Coronary calcifications in the left anterior descending artery (LAD), left circumflex artery (LCX) and right coronary artery (RCA) (A). Predicted coronary artery regions of the LAD - red, LCX - yellow and RCA - blue (B) and 3D surface model of the segmented CAR (C).

our model evaluated on the DISCHARGE test set and CADMAN test set is compared with methods evaluated on other non-public datasets in Table 4. Note that results are not directly comparable due to unknown data distributions.

We can see that our model trained on the DISCHARGE training set (F1-score=0.958) performs very good using only CSCTs. The best performing method of Gogin et al.[7] (F1-score=0.975) is using an ensemble of 3D CNNs to perform calcium scoring. Other methods are using CCTA to map segmentations from cardiac structures (heart, aorta, coronary arteries) in the CCTA to the CSCT [7] or use preprocessing by cylindrical cropping around an initial automatic segmentation of the ascending aorta.[3]. In the work of D. Eng et al. [45], two deep learning models were used to automate CAC scoring using gated unenhanced coronary CTs and non-gated unenhanced chest CTs, but reported performance metrics are not comparable with those in Table 4.

Note that our model performs well on the full DISCHARGE test set (F1-score=0.881), however due to the large variability in the dataset and inclusion of scans with motion and metal artifacts, the performance is lower than on the orCaScore test set (F1-score=0.958). Similar dataset dependent performance differences can be seen in Table 3 and 4 by Wolterink et al.[3] and Zhang et al.[8]. In Section III.E. we analyze different influencing factors to find reasons for these surprising findings.

The per patient risk categories were predicted based on the estimated agatston score of

Methods	Interaction	Dataset # scans (train, test)	ICC (Vol)	Sen. (Vol)	F1. (Vol)
Observer 1[3]	Manual	CSCT (-, 40)	0.998	0.985	0.9860
Observer 2[3]	Manual	CSCT (-, 40)	0.984	0.998	0.975
Shahzad et al.[42]	Automatic	CSCT+CCTA (209, 40)	0.971	0.621	0.893
Yang et al.[10]	Semi-Auto.	CSCT+CCTA (40, 40)	**0.992**	**0.940**	**0.968**
Kelm et al.[43]	Automatic	CSCT+CCTA (32, 40)	0.980	0.838	0.943
Kondo et al.[8]	Semi-Auto.	CSCT+CCTA (32, 40)	0.621	0.513	0.623
Durlak et al.[44]	Automatic	CSCT (32, 40)	0.989	0.835	0.951
Wolterink et al.[43]	Automatic	CSCT (373, 40)	0.986	0.845	0.947
Zhang et al.[8]	Automatic	CSCT (129, 40)	0.991	0.911	0.954
Gogin et al.[7]	Automatic	CSCT (783, 40)	**0.995**	**0.968**	**0.975**
Proposed network (DISCHARGE train)	Automatic	CSCT (215, 40)	0.994	0.955	0.958
Proposed network (orCaScore train)	Automatic	CSCT (32, 40)	0.984	0.961	0.928

Table 3: Performance comparison between our model and other state-of-the-art methods for automated coronary calcium scoring in cardiac CT on the orCaScore test set. Comparison is based on interclass correlation coefficient (ICC), Sensitivity (Sen.) and F1-score for CAC volume. The first block shows the performance of the two observers on the orCaScore test set. The second block shows results of all methods using non-contrast enhanced CT (CSCT) and contrast-enhanced coronary CT angiography (CCTA) on the orCaScore test set. The third block shows results of all methods using only CSCT on the orCaScore test set.

Methods	Dataset # scans (train, test)	ICC (Vol.)	Sen. (Vol.)	F1 (Vol.)
Kurkure et al.[46]	CSCT (100, 105)	-	0.921	-
Išgum et al.[47]	CSCT (228, 76)	-	0.738	-
Brunner et al.[48]	CSCT (30, 30)	-	0.863	-
Shahzad et al.[42]	CSCT (209, 157)	-	0.839	-
Zhang et al.[8]	CSCT (129 with 5-fold CV)	0.986	0.905	0.946
Wolterink et al.[43]	CSCT (373, 530)	0.96	0.79	0.85
Vos et al.[13]	CSCT (373, 530)	0.97	-	-
Zeleznik et al.[49]	CSCT (129, (441, 663, 4021))	0.89, 0.80, 0.792	-	-
Velzen et al.[50]	CSCT (373, 529)	0.970	-	-
Proposed network (DISCHARGE train)	CSCT (215, 1047) -DISCHARGE test	0.955	0.841	0.881
Proposed network (DISCHARGE train)	CSCT (215, 154) -CADMAN test	0.847	0.941	0.822

Table 4: Results of state-of-the-art methods for automated coronary calcium scoring in cardiac CT on non-public datasets. Results are compared in terms of interclass correlation coefficient (ICC), Sensitivity (Sen.) and F1-score. The method "Proposed network (DISCHARGE train)" refers to the proposed uncertainty weighted MTL-model trained on the DISCHARGE training set. The proposed network was evaluated on the DISCHARGE test set and CADMAN test set.

the CAC segmentations and compared with the risk categories based on the reference annotations. The confusion matrices of risk category predictions and corresponding linearly weighted Cohen's kappa (κ) for all three datasets are shown in Table 5. We use a linearly weighted kappa because risk categories are on an ordinal rating scale and the deviations are weighted differently depending on their size. It shows that κ is much higher for the orCaScore dataset (κ=0.97) compared to the DISCHARGE (κ=0.80) or CADMAN dataset (κ=0.80). Misclassifications of the risk category occurs mainly between category I and II because of false positive predictions.

a) DISCHARGE test set, $\kappa = 0.80$

	Automated risk category				
Risk	I	II	III	IV	Total
I	267	159	8	7	441
II	3	284	21	5	313
III	0	2	108	10	120
IV	0	0	2	171	173
Total	270	445	139	193	1047

b) orCaScore training set, $\kappa = 0.97$

	Automated risk category				
Risk	I	II	III	IV	Total
I	7	1	0	0	8
II	0	8	0	0	8
III	0	0	8	0	8
IV	0	0	0	8	8
Total	7	9	8	8	32

c) CADMAN test set, $\kappa = 0.80$

	Automated risk category				
Risk	I	II	III	IV	Total
I	39	16	4	0	59
II	0	49	3	2	54
III	0	1	18	2	21
IV	0	0	1	21	22
Total	39	66	26	25	156

Table 5: Confusion matrices show the agreement in CVD risk for the DISCHARGE test set (a), orCaScore training set (b) and CADMAN test set (c). Categorization is based on the total Agatston score with I: 0, II: [1,100), III: [100,300), IV: > 300.

III.D. Active learning evaluation with uncertainty weighted loss MTL-model

We evaluate our uncertainty weighted loss MTL-model in an active learning scenario by conducting two experiments. In the first experiment, we analyzed the model performance trained in an active learning scenario for different loss weighting strategies (uncertainty weighted loss, static weighted loss) and sampling strategies (random sampling, hybrid sampling) described in Subsection II.C.. We initially trained the model with only 100 randomly

selected samples (slices) and double the number of samples in each sampling round. Instead of retraining the model from scratch after each round, we continue training with the larger dataset and a reduced initial learning rate to 1e-04 compared to 5e-04 for the initial training. We useed early stopping based on the validation set to avoid overtraining in each sampling round. The Micro F1-score of the multi-class calcification segmentation is used to compare different models. As shown in Figure 5 with uncertainty weighted loss and hybrid sampling method, the model required only three sampling rounds and 800 annotated slices (12% of the training set) to achieve similar performance (Micro F1-score=0.846) as when trained on the full training set (Micro F1-score=0.849).

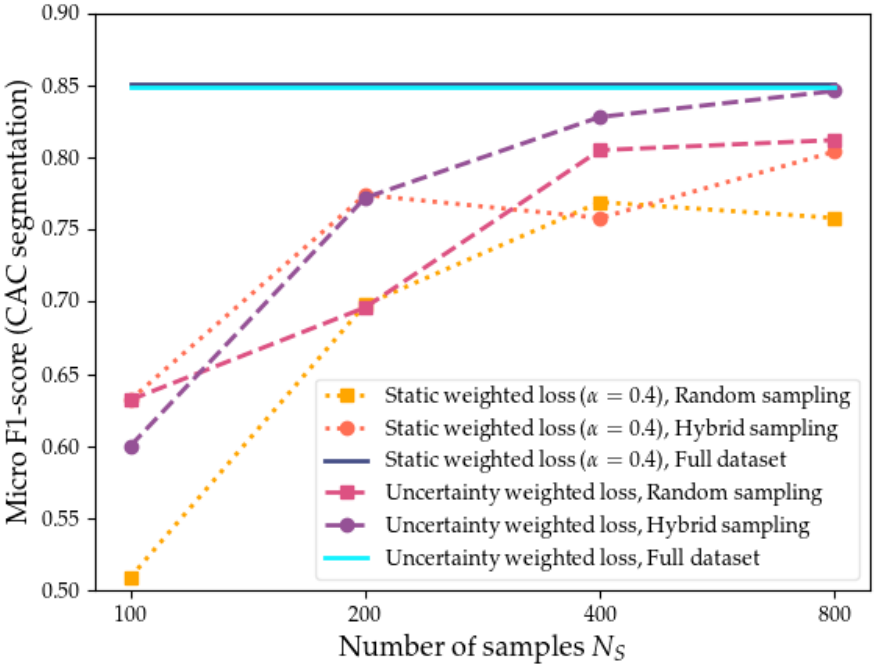


Figure 5: Performance comparison between different loss weighting methods (static weighted loss and uncertainty weighted loss) as well as different sampling methods (random sampling and hybrid sampling) in an active learning scenario.

To compare the model performance based on the used loss weighting strategy and sampling method after three sampling rounds, we compare the model performance proportion (compared to uncertainty weighted model on the full dataset) in an active learning scenario in Table 6. It can be seen that the uncertainty weighted loss outperforms static weighting for random and hybrid sampling by 6.4% and 4.9%, respectively. This can be explained by the

fact that the data distribution of the training set is changing in each sampling round and especially during the first sampling rounds. The uncertainty weighted loss method can compensate for this distribution shift, but static weighted loss cannot. It can also be seen that hybrid sampling outperforms random sampling for static weighted and uncertainty weighted loss by 5.4% and 4.0%, respectively. The hybrid sampling method selects only the most informative image slices and can therefore reduce number of required samples.

	Random sampling	Hybrid sampling
Static-weighted loss	89.28% (0.758/0.849)	94.70% (0.804/0.849)
Uncertainty-weighted loss	95.64% (0.812/0.849)	**99.65%** (0.846/0.849)

Table 6: Model performance proportion of the active learning model after three sampling rounds compared to the performance of the uncertainty weighted loss MTL-model trained on the full dataset.

The labeling of the additional coronary artery region annotations requires extra time, even if the annotation process is an efficient semi-automatic process described in Section II.A.3.. An approximation of the required annotation time for 1) coronary calcifications, 2) coronary calcifications and coronary artery regions and 3) only informative slices of coronary calcifications and coronary artery region was investigated empirically and is shown in Table 7. It shows that annotation of CAC and CAD with active learning reduces the annotation cost to approximately one-third, compared to labeling of calcifications on the full training set, even though labeling CAC and CAD is more time consuming.

	Annot. time per slice [s]	Number of labeled slices	Annot. time training set [s]	Improvement ratio
CAC	4.0	6721	26884	1.0
CAC + CAR	12.0	6721	80652	3.0
CAC + CAR + AL	12.0	800	9600	**0.36**

Table 7: Approximated annotation time for annotation of coronary artery calcifications (CAC) compared to annotation of coronary calcifications and coronary artery regions (CAC+CAR) and annotation of coronary calcifications and coronary artery regions using active learning (CAC+CAR+AL).

In a second experiment, we analyzed the impact of the number of training samples on the estimated optimal weighting parameter α in Equation (3) using a grid search method. Therefore, we trained the static weighted loss model multiple times with varying weighting parameter α from 0.1 to 0.9 and step size of 0.1 for a very small randomly selected dataset

(only 100 samples) and compared the results when trained on the full dataset. It turns out that the estimated optimal parameter of the full dataset $\alpha = 0.4$ does not match with the optimal weighting parameter of the small dataset $\alpha = 0.2$ because the small dataset does not represent the data distribution of the full dataset. If a non-optimal parameter value would have been selected after the first sampling round of the active learning method, optimal performance would not have been achieved. Alternatively, α could have been redetermined in each sampling round, but this would be computationally very expensive. A detailed analysis can be found in Figure S2 in the supplementary material.

III.E. Influence of image noise, metal artifacts, motion artifacts and image quality on the model performance

The performance rises by 4.6% if CT scans with severe image noise, metal artifacts, motion artifacts and image quality are excluded. We can see in Table 3 and Table 4 that the test performance on the orCaScore test set (Micro F1-score = 0.961) is much higher compared to the test performance on the DISCHARGE dataset (Micro F1-score = 0.881). To explain the performance difference, we analyzed the influence of four factors 1) image noise, 2) metal artifacts, 3) motion artifacts, 4) image quality. We estimated the noise using a method similar to Christianson et al..[51] First we segmented the CT image into the heart-related tissue types (–200 to 140 HU), second, a noise image filter[52] was applied to the segmented region, third, a histogram was generated and the highest peak was selected as noise level.[51] The noise levels of all CT scans of the test dataset were normalized using z-score[53] and the most noisy 20% were labeled as noisy CT scans. Metal artifacts and motion artifacts were determined visually and scans were labeled according to their presence or absence. Image quality was visually assessed by a high level of disturbance or anatomical abnormalities and labeled as good or bad quality, accordingly. Note that none of the CT scans in the test set were deemed as nondiagnostic (unsatisfactory for diagnosis) by a radiologist. Examples for the four influencing factors are shown in Figure 6. It shows that the Micro F1-score ranges from 0.881 including all scans in the test set to 0.927 if all noisy scans, scans with metal or motion artefact and poor quality images are excluded. When only noisy images are excluded, performance increases by 3.1%. Surprisingly, when we exclude images with severe motion artifacts, performance drops only by 0.02%. This can by explained by the fact that in motion artifacts, calcifications appear very large, resulting in a high number of "falsely"

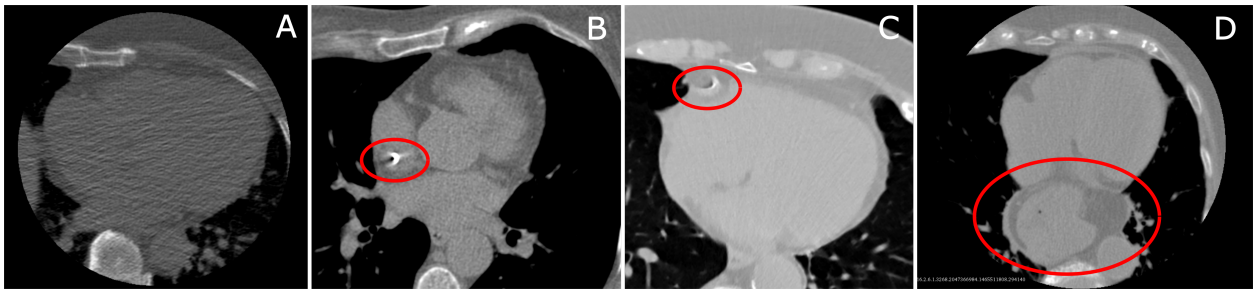


Figure 6: Image examples with severe image noise (A), metal artifacts (B), motion artifacts (C), low image quality (abnormality provoked by hiatal hernias) (D).

labeled true positives in the data set. Excluding samples with motion artifacts decreases the number of true positives and thus the corresponding Micro F1-score. A detailed analysis of the influencing factors and its influence on the model performance can be found in the supplementary material.

IV. Discussion

In this paper, we have proposed an MTL-model with uncertainty weighted loss for coronary calcium scoring in ECG-gated, non-contrast enhanced cardiac CTs. The model can be trained in an active learning scenario and requires only 12% of the training data and approximately one-third of the annotation time to achieve the same performance as when trained with the full dataset.

To the best of our knowledge, our model is the first that performs segmentation of coronary artery regions (CAR) based on weak annotations and segmentation of coronary calcifications (CAC) in an end-to-end framework. We compared our multi-task models with a single-task model (multiclass U-Net) and a sequential model. It can be seen in Table 2 that the multi-task models outperform other models in terms of positive predictive value, sensitivity, F1-score and Micro F1-score. The benefits of an MTL-model compared to a multiclass U-Net and a sequential model are the shared information between coronary artery region segmentation task T_R and calcification segmentation task T_L. In contrast to the multiclass U-Net, the MTL-model is able to learn important spatial information from weakly labeled samples and is able to transfer this knowledge for segmentation of coronary calcifications. Explanation techniques such as layer wise relevance propagation[54] could lead to a deeper understanding about different prediction strategies but they are beyond the scope of this

paper.

To investigate the uncertainty weighted loss, we compared the performance with a MTL-model trained using static weighted loss. When using optimal weighting parameter, the performance is similar, but it is important to note that the determination of the optimal weighting parameter value is a challenging and expensive process and even more difficult to estimate in an active learning scenario.

To reduce labeling costs, we investigated our uncertainty weighted multi-task network in an active learning scenario and could show that our model reaches optimal performance with substantial less training samples. The uncertainty weighted loss MTL-model is able to balance losses when the data distribution is changing after each sampling round. We compared different active learning scenarios and could show in Figure 5 that uncertainty weighted loss outperforms static weighted loss in random sampling and hybrid sampling. The biggest disadvantage of static weighting is the estimation of weighting parameter α, which is difficult to obtain and sensitive to the size of the training set shown in Figure S2 (supplementary material). In contrast to T. Gong et al.[55], we did not notice more instability issues when our uncertainty weighted loss model was trained on small datasets.

We compared our uncertainty weighted MTL-model with other methods in Table 3 on the orCaScore dataset and could show that our model performs very good in terms of F1-score, ICC and sensitivity using only CSCT. To compare the performance with respect to the dataset, we tested our model on three different datasets. To our surprise, the model trained on the DISCHARGE training set performed better on the orCaScore test set (Micro F1-score=0.958) than on the DISCHARGE test set (Micro F1-score=0.881). The test performance on the CADMAN dataset (Micro F1-score=0.822) was even lower than on the DISCHARGE test set due to a higher number of false positive predictions. This can be explained by a higher level of noise in the CADMAN dataset, since it contains only filtered back projections. The influence of noise can also be reflected in the higher number of lesion candidates per scan in Table 1 for the CADMAN and DISCHARGE dataset.

The predictions of cardiovascular disease (CVD) risk categories based on the segmentations in Table 5 show a very good agreement of $\kappa = 0.97$ for the orCaScore dataset but a lower agreement of $\kappa = 0.80$ for the DISCHARGE test set. Mislabeled noise leads to a high false positive rate between risk category I (total Agatston score is 0) and II (total Agatston score between 1 and 100) and similar findings have been made in R. Zeleznik et al..[49]

We also trained our model on the orCaScore training set with additional annotations for CAR and reached only slightly lower performance (Micro F1-score=0.928). To gain a better understanding of the different influencing factors (exclusion criteria) related to model performance, we compared the performance after exclusion of scans due to 1) image noise, 2) metal artifacts, 3) motion artifacts and 4) image quality). If all exclusion criteria were met, the Micro F1-score increased from 0.886 to 0.931. We have also shown that image noise is one of the most influencing factors on model performance beside metal artefacts and image quality. Scans with motion artifact had no effect on performance, which can be explained by visual expansion of the lesion area due to motion, mainly in the proximal RCA, leading to overestimation of the lesion volume in both the labeling phase by the radiologist and the prediction phase by the network.

IV.A. Limitations

We have seen that the convergence of the MTL-model trained with static weighted loss was more sensitive to changes of the learning rate compared to uncertainty weighted loss when trained on a small datasets. Nevertheless, training time requires several hours which makes tuning of the hyper-parameter challenging and limits the possibility to draw general conclusions.

Our method is processing 2D axial CT slices. The usage of 3D-information might be beneficial as shown in Zhang et al. [8] and recently published methods based on 3D-CNN ensembles [7] achieved very good results. Since our active learning approach is based on the labeling of only the most informative slices the 3D-annotations would be sparse. Learning dense 3D segmentations from sparse annotations can be challenging in a multi-task network [56] therefore we leave a 3D extension of our method for future work.

Furthermore, reference standards for the DISCHARGE and CADMAN datasets were provided by only one experienced observer for coronary calcifications and coronary artery regions. Independent annotations from a second observer and clarification of discrepancies by consensus would improve the quality of the dataset but since the annotation process requires expert knowledge and is tedious and time-consuming, we leave this improvement of the dataset for a further research project.

The analysis of influencing factors is limited to four factors (image noise, metal artifacts, mo-

tion artifacts, image quality), yet other factors such as reconstruction method, scanner type, slice thickness or slice spacing are known to influence model performance but are beyond the scope of this work and will be investigated in future work.

IV.B. Further research directions

With respect to our results, we have to critically reflect the question which loss and performance metrics are best suited for risk prediction of coronary heart disease events. Our model performs well on F1-scores, ICC and sensitivity of CAC volume but lacks precision on CVD risk agreement. In further analysis we will investigate how a direct prediction of the risk categories [13] can be integrated into our model to improve risk categorization. A major focus will be on improving the prediction of patients with zero calcium score. We also plan to extend our model from 2D input data to 3D to take advantage of 3D context information and overcome our limitations. We evaluated the uncertainty weighted MTL-model in an active learning scenario using our hybrid sampling method and believe that the model is also applicable with other sampling strategies but leave further analysis as future work. Additional future work may investigate how radiologist-in-the-loop frameworks might use explanations to guide a more efficient active learning based labeling process for coronary calcium scoring. A deeper understanding of the model behavior supported by explanations and quantification of model uncertainties would enable the radiologist to understand predictions and assist in medical decision making.

V. Conclusions

In this work we have proposed a multi-task model with uncertainty weighted loss for coronary calcium scoring. The model improves calcium scoring performance by extracting shared informative features from the two tasks of coronary artery region (CAR) segmentation and coronary artery calcifications (CAC) segmentation. The model performance was evaluated using a large multi-center dataset of the DISCHARGE trial (1047 CSCTs), a single-center dataset of the CAD-Man study (156 CSCTs) and the multi-center orCaScore test set (40 CSCTs). When trained in an active learning scenario, the model achieves optimal performance with only 12% of the training samples, reduces annotation time to one-third and

enables the integration of the radiologist into the training loop. The good performance and the reduction of required annotated image slices might enable the training of models applicable in a clinical setting.

Acknowledgement

Acknowledgment should be provided to the DISCHARGE Trial Group (Napp et al.)[22] for the collection and provision of the data from the DISCHARGE project (603266-2, HEALTH-2012.2.4.-2) funded by the FP7 Program of the European Commission (https://www.dischargetrial.eu):

Pál Maurovich-Horvat, M.D., Ph.D., M.P.H., Maria Bosserdt, Ph.D., Klaus F. Kofoed, M.D., D.M.Sc., Nina Rieckmann, Ph.D., Theodora Benedek, M.D., Ph.D., Patrick Donnelly, M.D., José Rodriguez-Palomares, M.D., Ph.D., Andrejs Erglis, M.D., Cyril Štěchovský, M.D., Gintarė Šakalytė, M.D., Nada Čemerlić Ađić, M.D., Matthias Gutberlet, M.D., Ph.D., Jonathan D. Dodd, M.D., Ignacio Diez, M.D., Gershan Davis, M.D., Elke Zimmermann, M.D., Cezary Kępka, M.D., Radosav Vidakovic, M.D., Ph.D., Marco Francone, M.D., Ph.D., Małgorzata Ilnicka-Suckiel, M.D., Ph.D., Fabian Plank, M.D., Ph.D., Juhani Knuuti, M.D., Rita Faria, M.D., Stephen Schröder, M.D., Colin Berry, M.D., Luca Saba, M.D., Balazs Ruzsics, M.D., Ph.D., Christine Kubiak, Ph.D., Iñaki Gutierrez-Ibarluzea, Ph.D., Kristian Schultz Hansen, Ph.D., Jacqueline Müller-Nordhorn, M.D., M.P.H., Bela Merkely, M.D., Ph.D., Andreas Dehlbæk Knudsen, M.D., Imre Benedek, M.D., Ph.D., Clare Orr, M.D., Filipa Xavier Valente, M.D., Ph.D., Ligita Zvaigzne, M.D., Vojtěch Suchánek, M.D., Antanas Jankauskas, M.D., Filip Ađić, M.D., Michael Woinke, M.D., Mark Hensey, M.B., B.Ch., B.A.O., Iñigo Lecumberri, M.D., Erica Thwaite, M.D., Michael Laule, M.D., Mariusz Kruk, M.D., Aleksandar N. Neskovic, M.D., Ph.D., Massimo Mancone, M.D., Donata Kuśmierz, M.D., Gudrun Feuchtner, M.D., Mikko Pietilä, M.D., Ph.D., Vasco Gama Ribeiro M.D., Tanja Drosch, M.D., Christian Delles, M.D., Gildo Matta, M.D., Michael Fisher, M.D., Ph.D., Bálint Szilveszter, M.D., Ph.D., Linnea Larsen, M.D., Ph.D., Mihaela Ratiu, M.D., Ph.D., Stephanie Kelly, Bruno Garcia del Blanco, M.D., Ph.D., Ainhoa Rubio, M.D., Zsófia D. Drobni, M.D., Birgit Jurlander, M.D., Ph.D., Ioana Rodean, M.D., Susan Regan, Hug Cuéllar Calabria, M.D., Ph.D., Melinda Boussoussou, M.D., Thomas Engstrøm, M.D., D.M.Sc., Roxana Hodas, M.D., Adriane E. Napp, Ph.D., Robert Haase, M.D., Georg M. Schuetz, M.D., Sarah Feger, M.D., Konrad Neumann, Ph.D., Henryk Dreger, M.D., Matthias Rief, M.D., Viktoria Wieske, M.D., Melanie Estrella, Ph.D., Peter Martus, Ph.D., Marc Dewey, M.D.

Affiliations:

MTA-SE Cardiovascular Imaging Research Group (P.M.-H., B.S., Z.D.D.) at Heart and Vascular Center (B.M., M.B.) and Department of Radiology, Medical Imaging Center (P.M.-H.), Semmelweis University, Budapest, Hungary; Department of Radiology (M.B., E.Z., A.E.N., R. Haase, G.M.S., S.F., M. Rief, V.W., M.E., M.D.) and Cardiology (M.L., H.D.) and Institute of Public Health (N.R., J.M.-N) and Institute of Biometry and Clinical Epidemiology (K.N.), Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin and DZHK (German Centre for Cardiovascular Research), partner site Berlin (H.D., M.D.) and Berlin Institute of Health (M.D.) and Berlin University Alliance (M.D.), Berlin, Germany; Department of Cardiology (K.F.K., A.D.K., T.E.) and Radiology (K.F.K., A.D.K.), Copenhagen University Hospital - Rigshospitalet & Department of Clinical Medicine, Faculty of Health and Medical Sciences and Department of Cardiology, Herlev-Gentofte Hospital (L.L.) and Department of Cardiology, Nordsjaellands Hospital, (B.J.), University of Copenhagen, Copenhagen, Denmark; Department of Internal Medicine, Clinic of Cardiology, (T.B., R. Hodas) and Department of Radiology and Medical Imaging (M. Ratiu), University of Medicine and Pharmacy, Science and Technology "G.E.Palade" and County Clinical Emergency Hospital Tirgu Mures (T.B.) and Center of Advanced Research in Multimodality Cardiac Imaging, Cardio Med Medical Center (I.B., I.R.), Tirgu Mures, Romania; Department of Cardiology, Southeastern Health and Social Care Trust, (P.D., C.O., S.K., S.R.) Belfast, United Kingdom; Department of Cardiology (J.R.-P, F.X.V., B.G.B,) and Radiology (H.C.C.), Hospital Universitari Vall d'Hebron, Institut de Recerca (VHIR), Universitat Autònoma de Barcelona, Barcelona, Spain; Department of Cardiology (A.E.) and Radiology (L. Z.), Paul Stradins Clinical University Hospital and University of Latvia (A.E.), Riga, Latvia; Department of Cardiology (C.Š.) and Department of Imaging Methods (V.S.), Motol University Hospital, Prague, Czech Republic; Department of Cardiology (G.Š.) and Department of Radiology (A.J.), Medical Academy, Lithuanian University of Health Sciences and Department of Cardiology, Hospital of Lithuanian University of Health Sciences (G.S.), Kaunas, Lithuania; Faculty of Medicine, University of Novi Sad, (N.Č.A., F.A.) and Department of Cardiology, Institute for Cardiovascular Diseases of Vojvodina (N.Č.A., F.A.), Novi Sad, Sremska Kamenica, Serbia; Department of Radiology (M.G.) and Department of Cardiology (M.W.), University of Leipzig Heart Centre, Leipzig, Germany; Department of Radiology (J.D.D.) and Department of Cardiology (M.H.), St. Vincent's University Hospital and School of Medicine (J.D.D.), University College Dublin,

Dublin, Ireland; Department of Cardiology (I.D., A.R.) and Department of Radiology (I.L.), Basurto Hospital, Bilbao Spain; Department of Cardiology (G.D.) and Department of Radiology (E.T.), Aintree University Hospital, Liverpool and School of Medicine, University of Central Lancashire (G.D.), Preston, United Kingdom; National Institute of Cardiology (C. Kępka, M.K.), Warsaw, Poland; Department of Cardiology, Internal Medicine Clinic, Clinical Hospital Center Zemun (R.V., A.N.N.) and Faculty of Medicine, University of Belgrade (R.V., A.N.N.), Belgrade, Serbia; Department of Radiological, Pathological and Oncological Sciences (M. Francone) and Department of Clinical Internal, Anesthesiologic and Cardiovascular Sciences (M.M.), Sapienza University of Rome, Rome, Italy; Department of Cardiology (M.I.S.) and Department of Radiology (D.K.), Wojewodzki Szpital Specjalistyczny We Wroclawiu, Wroclaw, Poland; Department of Internal Medicine III, Cardiology (F.P.) and Department of Radiology (G.F.), Innsbruck Medical University, Innsbruck, Austria; Turku PET Centre (J.K.) and Heart Center (M.P.), Turku University Hospital and University of Turku and Administrative Centre, Health Care District of Southwestern Finland (M.P.), Turku, Finland; Department of Cardiology, Centro Hospitalar de Vila Nova de Gaia/ Espinho (R.F., V.G.R.), Vila Nova de Gaia, Portugal; Department of Cardiology, ALB FILS KLINIKEN GmbH (S.S., T.D.), Göppingen, Germany; Institute of Cardiovascular & Medical Sciences, University of Glasgow (C.B., C.D.), Glasgow and Golden Jubilee National Hospital (C.B.), Clydebank, United Kingdom; Department of Radiology, University of Cagliari (L.S.) and Department of Radiology, Azienda Ospedaliera Brotzu (G.M.), Cagliari, CA, Italy; Department of Cardiology, Royal Liverpool University Hospital (B.R., M.F.) and Institute for Cardiovascular Medicine and Science, Liverpool Heart and Chest Hospital (B.R., M. Fisher), and Faculty of Health and Life Sciences, University of Liverpool (M. Fisher), Liverpool, United Kingdom; ECRIN-ERIC (European Clinical Research Infrastructure Network-European Research Infrastructure Consortium) (C. Kubiak), Paris, France; Basque Foundation for Health Innovation and Research, Barakaldo / Bizkaia and Basque Office for Health Technology Assessment (I.G.-I.), Vitoria - Gasteiz, Spain; University of Copenhagen, Department of Public Health, Section for Health Services Research (K.S.H.), Copenhagen, Denmark; Bavarian Cancer Registry, Bavarian Health and Food Safety Authority (J.M.-N.), Munich, Germany; Department of Clinical Epidemiology and Applied Biostatistics, Universitätsklinikum Tübingen, Tübingen (P.M.), Germany

Acknowledgment should also be provided for the CAD-MAN Trial Group (Dewey et al.)[24]

for the collection and provision of the data from the CAD-MAN project (ClinicalTrials.gov Identifier: NCT00844220) funded by a grant of the Heisenberg programme of the German Research Foundation to Marc Dewey.

The authors thank the organizers of the orCaScore Challenge for launching this international and open competition.

This work was funded by the German Research Foundation through the graduate program BIOQIC (GRK2260, project-ID: 289347353), the priority program SPP-Radiomics (SPP2177, project-ID: 402688427) and the DISCHARGE project (603266-2, HEALTH-2012.2.4.-2) funded by the FP7 Program of the European Commission.

Conflict of Interest

The author Marc Dewey declares relationships with the following companies: Prof. Dewey has received grant support from the FP7 Program of the European Commission for the randomized multicenter DISCHARGE trial (603266-2, HEALTH-2012.2.4.-2). He also received grant support from German Research Foundation (DFG) in the Heisenberg Program (DE 1361/14-1), graduate program on quantitative biomedical imaging (BIOQIC, GRK 2260/1), for fractal analysis of myocardial perfusion (DE 1361/18-1), the Priority Programme Radiomics for the investigation of coronary plaque and coronary flow (DE 1361/19-1 [428222922] and 20-1 [428223139] in SPP 2177/1). He also received funding from the Berlin University Alliance (GC_SC_PC 27) and from the Digital Health Accelerator of the Berlin Institute of Health. Prof. Dewey is European Society of Radiology (ESR) Research Chair (2019–2022) and the opinions expressed in this article are the author's own and do not represent the view of ESR. Per the guiding principles of ESR, the work as Research Chair is on a voluntary basis and only remuneration of travel expenses occurs. Prof. Dewey is also the editor of Cardiac CT, published by Springer Nature, and offers hands-on courses on CT imaging (www.ct-kurs.de). Institutional master research agreements exist with Siemens, General Electric, Philips, and Canon. The terms of these arrangements are managed by the legal department of Charité - Universitätsmedizin Berlin. Professor Dewey holds a joint patent with Florian Michallek on dynamic perfusion analysis using fractal analysis (PCT/EP2016/071551 and USPTO 2021 10,991,109).
Other authors declared no conflicts of interest.

[a] Author to whom correspondence should be addressed:
Bernhard Föllmer
Charité - Universitätsmedizin Berlin
Klinik für Radiologie
Campus Charité Mitte (CCM)
Charitéplatz 1, 10117 Berlin
E-Mail: bernhard.foellmer@charite.de

[b] These authors should be considered as joint senior authors.

References

[1] WHO, Cardiovascular diseases, fact sheet 317. World Health Organization. (2021).

[2] M. Dewey, Cardiac CT in Clinical Practice, in *Cardiac CT, Second Edition*, pages 33–42, Springer Berlin Heidelberg, 2014.

[3] J. M. Wolterink, T. Leiner, B. D. De Vos, J. L. Coatrieux, B. M. Kelm, S. Kondo, R. A. Salgado, R. Shahzad, H. Shu, M. Snoeren, R. A. Takx, L. J. Van Vliet, T. Van Walsum, T. P. Willems, G. Yang, Y. Zheng, M. A. Viergever, and I. Išgum, An evaluation of automatic coronary artery calcium scoring methods with cardiac CT using the orCaScore framework, Medical Physics **43**, 2361–2373 (2016).

[4] G. Santini, D. Della Latta, N. Martini, G. Valvano, A. Gori, A. Ripoli, C. L. Susini, L. Landini, and D. Chiappino, An automatic deep learning approach for coronary artery calcium segmentation, IFMBE Proceedings **65**, 374–377 (2017).

[5] A. S. Agatston, F. W. R. Janowitz, F. J. Hildner, N. R. Zusmer, M. Viamonte, and R. Detrano, Quantification Coronary Artery Calcium Using Ultrafast Cumputed Tomography, Journal of the American College of Cardiology **15**, 827–832 (1990).

[6] N. Lessmann, B. V. Ginneken, M. Zreik, P. A. D. Jong, B. D. D. Vos, M. A. Viergever, and I. Išgum, Automatic calcium scoring in low-dose chest CT using deep neural networks with dilated convolutions, **37**, 615–625 (2018).

[7] N. Gogin, M. Viti, L. Nicodème, M. Ohana, H. Talbot, U. Gencer, M. Mekukosokeng, T. Caramella, Y. Diascorn, J. Y. Airaud, M. S. Guillot, Z. Bensalah, C. Dam Hieu, B. Abdallah, I. Bousaid, N. Lassau, and E. Mousseaux, Automatic coronary artery calcium scoring from unenhanced-ECG-gated CT using deep learning, Diagnostic and Interventional Imaging **102**, 683–690 (2021).

[8] W. Zhang, J. Zhang, X. Du, Y. Zhang, and S. Li, An end-to-end joint learning framework of artery-specific coronary calcium scoring in non-contrast cardiac CT, Computing **101**, 667–678 (2019).

[9] J. M. Wolterink, T. Leiner, B. D. de Vos, R. W. van Hamersvelt, M. A. Viergever, and I. Išgum, Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks, Medical Image Analysis **34**, 123–136 (2016).

[10] G. Yang, Y. Chen, X. Ning, Q. Sun, H. Shu, and J.-L. Coatrieux, Automatic coronary calcium scoring using noncontrast and contrast CT images., Medical physics **43**, 2174 (2016).

[11] S. G. M. V. Velzen, N. Hampe, and B. D. D. Vos, AI for Calcium Scoring, pages 1–22.

[12] R. Shahzad, L. van Vliet, W. Niessen, and T. V. Walsum, Automatic Classification of Calcification in the Coronary Vessel Tree, Orcascore.Isi.Uu.Nl .

[13] B. D. de Vos, J. M. Wolterink, T. Leiner, P. A. de Jong, N. Lessmann, and I. Isgum, Direct Automatic Coronary Calcium Scoring in Cardiac and Chest CT., IEEE transactions on medical imaging **38**, 2127–2138 (2019).

[14] A. Smailagic, H. Y. Noh, P. Costa, D. Walawalkar, K. Khandelwal, M. Mirshekari, J. Fagert, A. Galdran, and S. Xu, MedAL : Accurate and Robust Deep Active Learning for Medical Image Analysis, (2018).

[15] S. Ruder, An Overview of Multi-Task Learning in Deep Neural Networks ∗ arXiv : 1706 . 05098v1 [cs . LG] 15 Jun 2017, (2017).

[16] Z. Chen, V. Badrinarayanan, C. Y. Lee, and A. Rabinovich, GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks, 35th International Conference on Machine Learning, ICML 2018 **2**, 1240–1251 (2018).

[17] S. Liu, E. Johns, and A. J. Davison, End-to-end multi-task learning with attention, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition **2019-June**, 1871–1880 (2019).

[18] M. Guo, A. Haque, D.-A. Huang, S. Yeung, and L. Fei-Fei, *Dynamic Task Prioritization for Multitask Learning: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*, pages 282–299, 2018.

[19] R. Cipolla, Y. Gal, and A. Kendall, Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition , 7482–7491 (2018).

[20] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool, Multi-Task Learning for Dense Prediction Tasks: A Survey, IEEE Transactions on Pattern Analysis and Machine Intelligence , 1–20 (2021).

[21] M. Dewey, The Discharge trial, https://www.dischargetrial.eu/, 2021.

[22] A. E. Napp, R. Haase, M. Laule, G. M. Schuetz, M. Rief, H. Dreger, G. Feuchtner, G. Friedrich, Š. Miloslav, and M. Dewey, Computed tomography versus invasive coronary angiography : design and methods of the pragmatic randomised multicentre DISCHARGE trial, pages 2957–2968 (2017).

[23] I. I. Jelmer M. Wolterink, Bob D. de Vos, Tim Leiner, Max A. Viergever, orCaScore, https://orcascore.grand-challenge.org/, 2021.

[24] M. Dewey, M. Rief, P. Martus, B. Kendziora, S. Feger, H. Dreger, S. Priem, F. Knebel, M. Böhm, P. Schlattmann, B. Hamm, E. Schönenberger, M. Laule, and E. Zimmermann, Evaluation of computed tomography in patients with atypical angina or chest pain clinically referred for invasive coronary angiography: Randomised controlled trial, BMJ (Online) **355** (2016).

[25] F. A. Jolesz, *Intraoperative Imaging and Image- Guided Therapy 123*.

[26] O. Ronneberger, P.Fischer, and T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241, Springer, 2015.

[27] R. Ke, A. Bugeau, N. Papadakis, P. Schuetz, and C. B. Schönlieb, Learning to Segment Microscopy Images with Lazy Labels, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12535 LNCS, pages 411–428, Springer Science and Business Media Deutschland GmbH, 2020.

[28] G. Papandreou, L.-c. Chen, K. Murphy, and A. L. Yuille, Weakly- and Semi-Supervised Learning of a Deep Convolutional Network for Semantic Image Segmentation.

[29] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, Focal Loss for Dense Object Detection, IEEE Transactions on Pattern Analysis and Machine Intelligence **42**, 318–327 (2020).

[30] S. Kongyoung, C. Macdonald, and I. Ounis, Multi-Task Learning using Dynamic Task Weighting for Conversational Question Answering, pages 17–26 (2020).

[31] S. Ioffe and C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, 32nd International Conference on Machine Learning, ICML 2015 **1**, 448–456 (2015).

[32] M. J. Brown, L. A. Hutchinson, M. J. Rainbow, K. J. Deluzio, and A. R. De Asha, Rectified Linear Units Improve Restricted Boltzmann Machines, Journal of Applied Biomechanics **33**, 384–387 (2017).

[33] D. P. Kingma and J. L. Ba, Adam: A method for stochastic optimization, 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings , 1–15 (2015).

[34] S. Budd, E. C. Robinson, and B. Kainz, A survey on active learning and human-in-the-loop deep learning for medical image analysis, Medical Image Analysis **71**, 102062 (2021).

[35] P. Ren, Y. Xiao, X. Chen, X. Wang, X. Chang, P.-Y. Huang, and Z. Li, A Survey of Deep Active Learning, Technical report, 2020.

[36] M. Fang and D. Tao, Active multi-task learning via bandits, SIAM International Conference on Data Mining 2015, SDM 2015 , 505–513 (2015).

[37] A. Harpale and Y. Yang, Active learning for multi-task adaptive filtering, ICML 2010 - Proceedings, 27th International Conference on Machine Learning , 431–438 (2010).

[38] Y. Zhang, Multi-task active learning with output constraints, Proceedings of the National Conference on Artificial Intelligence **1**, 667–672 (2010).

[39] Y. Gal and Z. Ghahramani, Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 1050–1059, JMLR.org, 2016.

[40] T. Nair, D. Precup, D. L. Arnold, and T. Arbel, Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation, in *Lecture Notes*

in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 11070 LNCS, pages 655–663, Springer Verlag, 2018.

[41] M. Grandini, E. Bagli, and G. Visani, Metrics for Multi-Class Classification: an Overview, pages 1–17 (2020).

[42] R. Shahzad, T. V. Walsum, M. Schaap, A. Rossi, S. Klein, A. C. Weustink, P. J. D. Feyter, L. J. V. Vliet, and W. J. Niessen, Vessel Specific Coronary Artery Calcium Scoring: An Automatic System, Academic Radiology **20**, 1–9 (2009).

[43] J. M. Wolterink, T. Leiner, R. A. P. Takx, M. A. Viergever, and I. Išgum, Automatic Coronary Calcium Scoring in Non-Contrast-Enhanced ECG-Triggered Cardiac CT With Ambiguity Detection, Tmi **34**, 1867–1878 (2015).

[44] F. Durlak, M. Wels, C. Schwemmer, M. Sühling, S. Steidl, and A. Maier, Growing a Random Forest with Fuzzy Spatial Features for Fully Automatic Artery-Specific Coronary Calcium Scoring, in *Machine Learning in Medical Imaging*, edited by Q. Wang, Y. Shi, H.-I. Suk, and K. Suzuki, pages 27–35, Cham, 2017, Springer International Publishing.

[45] D. Eng et al., Automated coronary calcium scoring using deep learning with multicenter external validation, npj Digital Medicine **4** (2021).

[46] U. Kurkure, D. R. Chittajallu, G. Brunner, Y. H. Le, and I. A. Kakadiaris, A supervised classification-based method for coronary calcium detection in non-contrast CT, International Journal of Cardiovascular Imaging **26**, 817–828 (2010).

[47] I. Išgum, A. Rutten, M. Prokop, and B. van Ginneken, Detection of coronary calcifications from computed tomography scans for automated risk assessment of coronary artery disease, Medical Physics **34**, 1450–1461 (2007).

[48] G. Brunner, D. R. Chittajallu, U. Kurkure, and I. A. Kakadiaris, Toward the automatic detection of coronary artery calcification in non-contrast computed tomography data., The international journal of cardiovascular imaging **26**, 829–838 (2010).

[49] R. Zeleznik et al., Deep convolutional neural networks to predict cardiovascular risk from computed tomography, Nature Communications (2021).

[50] S. G. M. V. Velzen, N. Lessmann, and B. K. Velthuis, Deep Learning for Automatic Calcium Scoring in CT : Validation Using Multiple Cardiac CT and Chest CT Protocols, Radiology (2020).

[51] O. Christianson, J. Winslow, D. P. Frush, and E. Samei, Automated Technique to Measure Noise in Clinical CT Examinations, American Journal of Roentgenology **205**, W93–W99 (2015).

[52] SimpleITK, SimpleITK, https://simpleitk.org, 2021.

[53] D. Sylvan, *Introduction to Mathematical Statistics*, volume 23, 2013.

[54] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation, PLOS ONE **10**, e0130140 (2015).

[55] T. Gong, T. Lee, C. Stephenson, V. Renduchintala, S. Padhy, A. Ndirango, G. Keskin, and O. H. Elibol, A Comparison of Loss Weighting Strategies for Multi task Learning in Deep Neural Networks, IEEE Access **7**, 141627–141632 (2019).

[56] J.-M. Bokhorst, H. Pinckaers, P. Van Zwam, I. Nagtegaal, J. Van Der Laak, F. Ciompi, and F. C. Nl, Learning from sparsely annotated data for semantic segmentation in histopathology images, Proceedings of Machine Learning Research **102**, 84–91 (2019).

Active Multi-Task Learning with Uncertainty Weighted Loss for Coronary Calcium Scoring

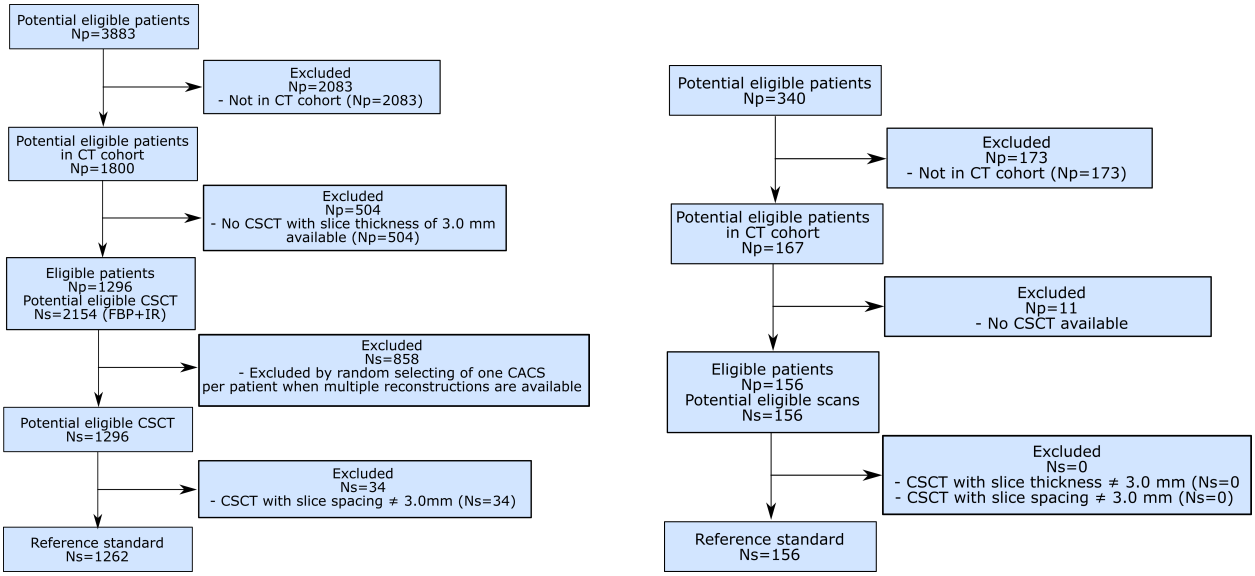
Supplemental Materials

I. Dataset selection process

Figure S1 shows the selection process for the DISCHARGE and CADMAN datasets.

For the DISCHARGE dataset, we considered all 3883 patients ($Np = 3883$) of the DISCHARGE trial as potential eligible patients. First, we excluded all patients who were not part of the study CT cohort. Second, at the date of 2020-09-01, all patient without any non-contrast enhanced cardiac CTs (CSCT) of 3.00 mm were excluded. Third, for each patients with multiple CSCT reconstructions, we randomly selected one of the reconstructions and excluded the rest. Fourth, we excluded all CSCTs with a slice spacing $\neq$ 3.0 mm.

For the CADMAN dataset, we considered all 340 patients of the CADMAN trial as potential eligible patients. First, we excluded all patients who are not part of the study CT cohort. Second, we excluded all patients without CSCT. Third, we excluded all CSCT scans with a slice thickness $\neq$ 3.0 mm or slice spacing $\neq$ 3.0 mm.



(a) Flowchart of the dataset selection process for the DISCHARGE dataset

(b) Flowchart of the dataset selection process for the CADMAN dataset

Figure S1: Flowchart of the dataset selection process for the DISCHARGE dataset and CADMAN dataset.

II. Dependency between model performance and loss weighting parameter

The estimation of the optimal weighting parameter is challanging when trained in a active learning scenario. In Figure S2 we see the dependency between model performance (Micro F1-score) and static weighting parameter value α, trained on a small, randomly selected training set (100 samples) and the full training set (6721 samples). The model was train with varying weighting parameter α from 0.1 to 0.9 and step size of 0.1. It shows, that the optimal parameter value $\alpha = 0.2$ of the small training does not match with the optimal parameter value $\alpha = 0.4$ of the full training set.

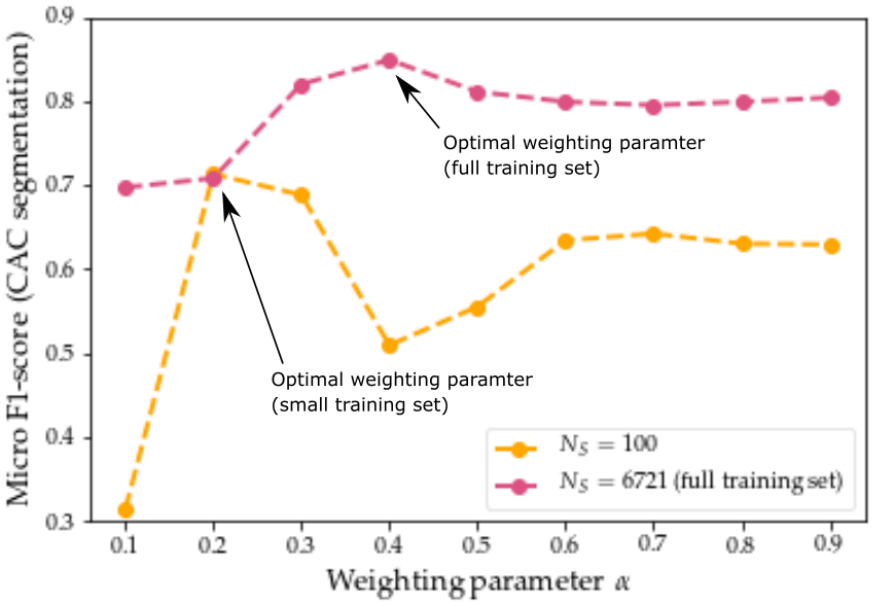


Figure S2: Dependency between model performance and loss weighting parameter value α for a very small training set with 100 samples and the full training with 6721 samples.

III. Comparison of model predictions between the sequential model and the uncertainty weighted loss MTL-model

To analyze different prediction strategies of the sequential model and the uncertainty weighted loss MTL-model, we compared the predition results of an image with severe image noise. The uncertainty weighted MTL-model performs both task simultaneously and extracts joint informative features which helps to avoid false positive predictions of calcifications in noisy image slices. Figure S3 shows that the sequential model has problems to distinguish noise from micro calcifications and predicts more false positive than the uncertainty weighted MTL-model.

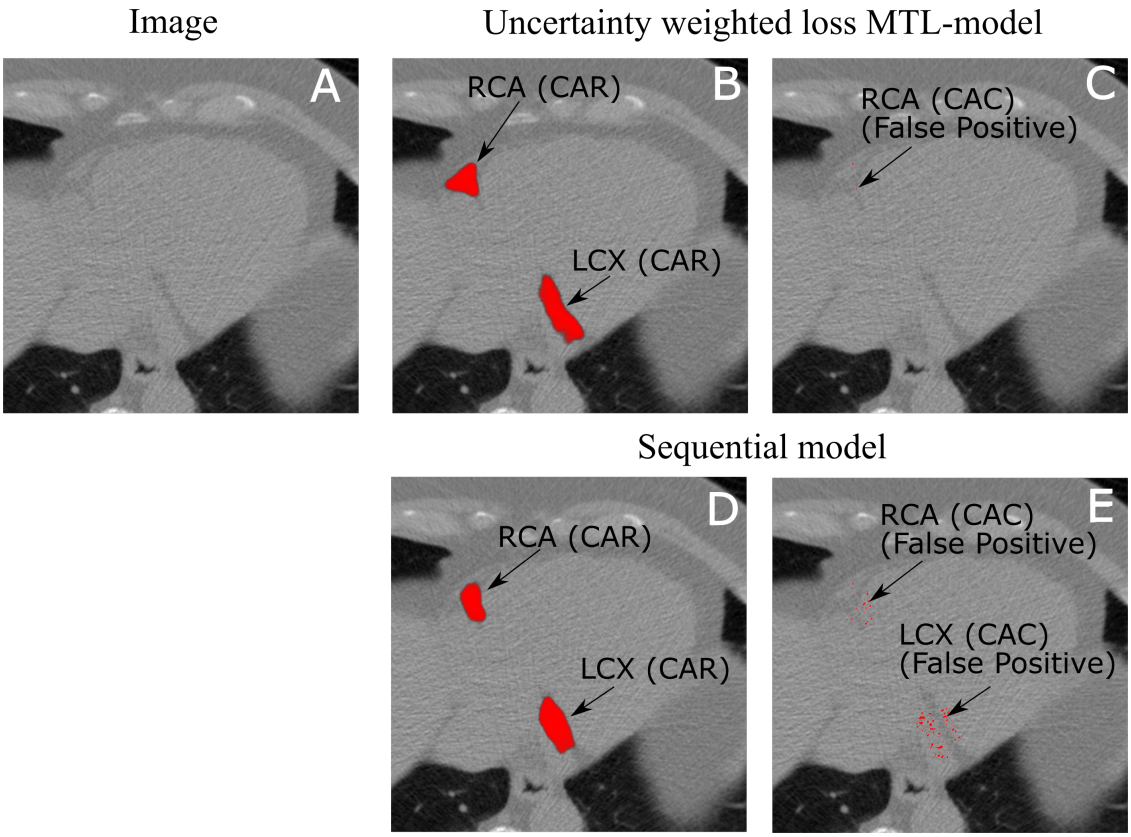


Figure S3: Comparison of model predictions for the uncertainty weighted MTL-model and sequential model. The first row shows the image slice (A) with a high level of noise, predicted coronary artery regions (CAR) (B) and coronary calcifications (CAC) (C) of the uncertainty weighted MTL-model. The second row shows the predicted coronary artery regions (CAR) (D) and coronary calcifications (CAC) (E) of the sequential model. In contrast to the sequential mode, joint learning of the two tasks and extraction of shared features, supports the ability to predicts fewer false positive calcifications in noisy images.

IV. Dependency between model performance and image noise, metal artifacts, motion artifacts and image quality

The model performance of the test set is strongly dependent on the selection criteria of the included CT scans. To analyze the influence of the four factors 1) image noise, 2) metal artifacts, 3) motion artifacts, 4) image quality, we performed a subset analysis of the DISCHARGE test set. Each row in Figure 1 corresponds to a subset of the DISCHARGE test set. The columns Noisy scan, Metal artifact, Motion artifact and Image quality are exclusion criteria according to which the scans are included (✓) or excluded (✗).

# Scans	Noisy scan	Metal artifacts	Motion artifacts	Image quality	F1-score	Sen.	PPV	κ
1047	✓	✓	✓	✓	0.881	0.842	0.924	0.800
924	✗	✓	✓	✓	0.914	0.887	0.942	0.834
1030	✓	✗	✓	✓	0.886	0.851	0.923	0.804
910	✗	✗	✓	✓	0.920	0.899	0.941	0.839
1025	✓	✓	✗	✓	0.878	0.833	0.929	0.795
907	✗	✓	✗	✓	0.914	0.881	0.950	0.828
1009	✓	✗	✗	✓	0.884	0.843	0.929	0.798
893	✗	✗	✗	✓	0.922	0.895	0.950	0.833
1028	✓	✓	✓	✗	0.886	0.851	0.924	0.809
910	✗	✓	✓	✗	0.919	0.897	0.942	0.839
1012	✓	✗	✓	✗	0.891	0.861	0.924	0.813
897	✗	✗	✓	✗	0.925	0.909	0.941	0.845
1006	✓	✓	✗	✗	0.884	0.843	0.930	0.804
893	✗	✓	✗	✗	0.920	0.892	0.951	0.833
991	✓	✗	✗	✗	0.890	0.854	0.929	0.807
880	✗	✗	✗	✗	0.927	0.905	0.950	0.839

Table 1: Dependency between model performance and exclusion criteria (image noise, metal artifacts, motion artifacts, image quality) on volume level for the DISCHARGE test dataset (✓- included, ✗- excluded).

V. Deviation of the uncertainty weighted total loss

We derive our total loss function based on the uncertainty weighted loss method of Cipolla et al.[19]. The outputs of the coronary calcification segmentation decoder and coronary artery region segmentation decoder are defined as $f^{\mathbf{W}}_{c_L}(\mathbf{x})$ and $f^{\mathbf{W}}_{c_R}(\mathbf{x})$, respectively. In the derivation of the total loss in Cipolla et al., a scaled version of the output $f^{\mathbf{W}}(\mathbf{x})$ is squashed through a softmax function.

$$p\left(\mathbf{y} \mid \mathbf{f}^{\mathbf{W}}(\mathbf{x})\right) = \text{Softmax}\left(\frac{1}{\sigma^2}\mathbf{f}^{\mathbf{W}}(\mathbf{x})\right) = \frac{\exp(\frac{1}{\sigma^2} f^{\mathbf{W}}_{c}(\mathbf{x}))}{\sum_{c'} \exp(\frac{1}{\sigma^2} f^{\mathbf{W}}_{c'}(\mathbf{x}))} \tag{E1}$$

The log likelihood for class c can be writte as

$$\log p\left(\mathbf{y} = c \mid \mathbf{f}^{\mathbf{W}}(\mathbf{x}), \sigma\right) = \frac{1}{\sigma^2} f^{\mathbf{W}}_{c}(\mathbf{x}) - \log \sum_{c'} \exp\left(\frac{1}{\sigma^2} f^{\mathbf{W}}_{c'}(\mathbf{x})\right) \tag{E2}$$

with $f^{\mathbf{W}}_{c}(\mathbf{x})$ the c'th element of the vector $f^{\mathbf{W}}(\mathbf{x})$.

We define the likelihood to factorise over the two outputs with the scaling parameters σ_R and σ_L, hence

$$\begin{aligned}
\mathcal{L}_{total}(\mathbf{W}, \sigma_R, \sigma_L) &= -\log p\left(\mathbf{y^R} = c_R, \mathbf{y^L} = c_L \mid \mathbf{f}^{\mathbf{W}}(\mathbf{x})\right) \\
&= \log(\text{Softmax}(y_R = c_R \mid \mathbf{f}^{\mathbf{W}}, \sigma_R) \cdot \text{Softmax}(y_L = c_L \mid \mathbf{f}^{\mathbf{W}}, \sigma_L)) \\
&= -\log \frac{\exp(\frac{1}{\sigma_R^2} f^{\mathbf{W}}_{c_R}(\mathbf{x}))}{\sum_{c'_R} \exp(\frac{1}{\sigma_R^2} f^{\mathbf{W}}_{c'_R}(\mathbf{x}))} - \log \frac{\exp(\frac{1}{\sigma_L^2} f^{\mathbf{W}}_{c_L}(\mathbf{x}))}{\sum_{c'_L} \exp(\frac{1}{\sigma_L^2} f^{\mathbf{W}}_{c'_L}(\mathbf{x}))}
\end{aligned} \tag{E3}$$

$$\frac{1}{\sigma_2} \sum_{c'} \exp\left(\frac{1}{\sigma_2^2} f^{\mathbf{W}}_{c'}(\mathbf{x})\right) \approx \left(\sum_{c'} \exp\left(f^{\mathbf{W}}_{c'}(\mathbf{x})\right)\right)^{\frac{1}{\sigma_2^2}} \tag{E4}$$

With the simplification[19] in equation (E4) we get:

$$\begin{aligned}
\mathcal{L}_{total}(\mathbf{W}, \sigma_R, \sigma_L) &= -\log \frac{\exp(\frac{1}{\sigma_R^2} f^{\mathbf{W}}_{c_R}(\mathbf{x}))}{\sigma_R(\sum_{c'_R} \exp(f^{\mathbf{W}}_{c'_R}(\mathbf{x})))^{\frac{1}{\sigma_R^2}}} - \log \frac{\exp(\frac{1}{\sigma_L^2} f^{\mathbf{W}}_{c_L}(\mathbf{x}))}{\sigma_L(\sum_{c'_L} \exp(f^{\mathbf{W}}_{c'_L}(\mathbf{x})))^{\frac{1}{\sigma_L^2}}} \\
&= -\frac{1}{\sigma_R^2} \log \frac{\exp(f^{\mathbf{W}}_{c_R}(\mathbf{x}))}{\sum_{c'_R} \exp(f^{\mathbf{W}}_{c'_R}(\mathbf{x}))} + \log(\sigma_R) \\
&\quad -\frac{1}{\sigma_L^2} \log \frac{\exp(f^{\mathbf{W}}_{c_L}(\mathbf{x}))}{\sum_{c'_L} \exp(f^{\mathbf{W}}_{c'_L}(\mathbf{x}))} + \log(\sigma_L)
\end{aligned} \tag{E5}$$

Analogous, we propose our total loss based on the focal losses defined in (1) and (2) claimed on the basis of our experiments.

$$\begin{aligned}\mathcal{L}_{total}(\mathbf{W},\sigma_R,\sigma_L) =& -\frac{1}{\sigma_R^2}(\log\frac{\exp(f_{c_R}^{\mathbf{W}}(\mathbf{x}))}{\sum_{c'_R}\exp(f_{c'_R}^{\mathbf{W}}(\mathbf{x}))}w_{c_R}(1-\frac{\exp(f_{c_R}^{\mathbf{W}}(\mathbf{x}))}{\sum_{c'_R}\exp(f_{c'_R}^{\mathbf{W}}(\mathbf{x}))})^{\gamma_R}) + \log(\sigma_R)\\ &-\frac{1}{\sigma_L^2}(\log\frac{\exp(f_{c_L}^{\mathbf{W}}(\mathbf{x}))}{\sum_{c'_L}\exp(f_{c'_L}^{\mathbf{W}}(\mathbf{x}))}w_{c_L}(1-\frac{\exp(f_{c_L}^{\mathbf{W}}(\mathbf{x}))}{\sum_{c'_L}\exp(f_{c'_L}^{\mathbf{W}}(\mathbf{x}))})^{\gamma_L}) + \log(\sigma_L)\\ =&\frac{1}{\sigma_R^2}\mathcal{L}_R(\mathbf{W}) + \frac{1}{\sigma_L^2}\mathcal{L}_L(\mathbf{W}) + \log\sigma_R + \log\sigma_L\end{aligned} \tag{E6}$$

with the two losses:

$$\mathcal{L}_R(\mathbf{W}) = -\log(\text{Softmax}(y_R = c_R \mid \mathbf{f}_{c_R}^{\mathbf{W}}))w_{c_R}(1-\text{Softmax}(y_R = c_R \mid \mathbf{f}_{c_R}^{\mathbf{W}}))^{\gamma_R} \tag{E7}$$

$$\mathcal{L}_L(\mathbf{W}) = -\log(\text{Softmax}(y_L = c_L \mid \mathbf{f}_{c_L}^{\mathbf{W}}))w_{c_L}(1-\text{Softmax}(y_L = c_L \mid \mathbf{f}_{c_L}^{\mathbf{W}}))^{\gamma_L} \tag{E8}$$

The final weighted loss depends on the model parameters $\mathbf{W}$ and the two task specific scalars σ_R and σ_L.