Received: 18 November 2021 | Revised: 1 June 2021 | Accepted: 20 June 2022

DOI: 10.1002/mp.15870

RESEARCH ARTICLE

MEDICAL PHYSICS

Active multitask learning with uncertainty-weighted loss for coronary calcium scoring

Bernhard Föllmer[1] | **Federico Biavati**[1] | **Christian Wald**[1] | **Sebastian Stober**[2] | **Jackie Ma**[3] | **Marc Dewey**[1,4] | **Wojciech Samek**[3]

[1] Department of Radiology, Charité-Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[2] Artificial Intelligence Lab, Otto-von-Guericke-Universität, Magdeburg, Germany

[3] Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, Berlin, Germany

[4] Berlin Institute of Health and DZHK (German Centre for Cardiovascular Research), Berlin, Germany

Correspondence
Bernhard Föllmer, Charité - Universitätsmedizin Berlin, Klinik für Radiologie, Campus Charité Mitte (CCM), Charitéplatz 1, 10117 Berlin.
Email: bernhard.foellmer@charite.de

Authors Marc Dewey and Wojciech Samek should be considered as joint senior authors.

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: GRK2260; SPP-Radiomics, Grant/Award Number: SPP2177; FP7 Program of the European Commission, Grant/Award Numbers: 603266-2, HEALTH-2012.2.4.-2; BIOQIC

Abstract

Purpose: The coronary artery calcification (CAC) score is an independent marker for the risk of cardiovascular events. Automatic methods for quantifying CAC could reduce workload and assist radiologists in clinical decision-making. However, large annotated datasets are needed for training to achieve very good model performance, which is an expensive process and requires expert knowledge. The number of training data required can be reduced in an active learning scenario, which requires only the most informative samples to be labeled. Multitask 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 multitask learning model for coronary calcium scoring in electrocardiogram-gated (ECG-gated), noncontrast-enhanced cardiac calcium scoring CT. The model was 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 samples needed for training. We compared our model with a single-task U-Net and a sequential-task model as well as other state-of-the-art methods. The model was evaluated on 1275 individual patients in three different datasets (DISCHARGE, CADMAN, orCaScore), and the relationship between model performance and various influencing factors (image noise, metal artifacts, motion artifacts, image quality) was analyzed.

Results: Joint learning of multiclass coronary artery region segmentation and binary coronary calcium segmentation improved calcium scoring performance. Since shared information can be learned from both tasks for complementary purposes, the model reached 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 abnormalities and metal artifacts.

Conclusions: Our multitask 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 multitask learning, uncertainty-weighted loss

---

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2022 The Authors. *Medical Physics* published by Wiley Periodicals LLC on behalf of American Association of Physicists in Medicine.

1 | INTRODUCTION

Cardiovascular disease (CVD) is the leading cause of death globally.[1] It is well established that coronary calcium is associated with coronary atherosclerosis, whereas its absence predicts a very low risk of adverse coronary events.[2] In clinical practice, the coronary calcium burden is assessed using semiautomatic software to select coronary artery calcifications (CACs) in computed tomography (CT) image slices from automatically labeled candidates, which is a tedious and time-consuming process, especially when performed in the setting of large studies.[3] Typically, ECG-gated, noncontrast-enhanced CT, known as calcium scoring CT (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 types of imaging data such as noncontrast-enhanced ECG-gated CSCT scans,[8] contrast-enhanced coronary CT angiography (CCTA),[9] or a combination of both.[3,10,11] Since segmentation of the cardiac tree is very challenging in CSCT, methods combining noncontrast-enhanced and enhanced CT usually map spatial information on the coronary arteries from enhanced onto unenhanced images.[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 (LAD) artery, left circumflex artery (LCX), and right coronary artery (RCA), and others directly perform a regression of the Agatston score without segmentation.[13]

Currently, most 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 for use in clinical practice. Unfortunately, this is an expensive and time-consuming process and requires expert knowledge. This situation calls for methods that can reduce labeling costs and improve the performance by integrating the radiologist into the training process.

Active learning techniques can improve the 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 selection strategies 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 on the 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. Separate extraction of spatial information about the coronary arteries in an auxiliary task can overcome this problem. Multitask learning (MTL) is a machine learning approach in which several related tasks are learned simultaneously, which improves model performance by sharing complementary information.[15] In coronary calcium scoring, spatial information on the coronary arteries is closely related to the calcium scoring task, and therefore, learning of coronary artery regions (CARs) is a good auxiliary task to support the primary calcium scoring task. However, the optimization of multiple loss functions for MTL 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 (DWA)[17], dynamic task prioritization [18], and uncertainty-weighted loss[19] have been developed, and results suggest that the best optimization method should be selected on a per-case basis.[20]

The training of multitask models in an active learning scenario can be challenging if the dataset is very small. In this work, we exploit an MTL model with uncertainty-weighted loss that outperforms a single-task U-Net and a sequential model. The MTL model achieves very good performance on small training sets and can therefore be used in active learning scenarios. Its performance is similar to that of other state-of-the-art methods, yielding similar results as our statically weighted MTL model with optimally chosen weighting parameters. The contributions of this paper can be summarized as follows:

- We propose a novel learning paradigm for coronary calcium scoring based on simultaneous learning of multiple related tasks to increase data efficiency and model performance by leveraging auxiliary information through shared informative features.
- We propose a multitask encoder–decoder model for simultaneous CAR segmentation (multiclass) and CAC segmentation (binary) to improve model performance compared to single-task models.
- We show that our model achieves 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 multitask model and show how uncertainty-weighted loss can facilitate active MTL.

TABLE 1 Number of candidate lesions (connected 3-D image voxels with intensities > 130 HU) in the three datasets and subsets used in our study. Candidate lesions include calcified coronary artery lesions (LAD, LCX, RCA) and other calcified structures such as bone and extra-coronary calcifications (OTHER_CAC). Since the orCaScore test set is not public, no information on the distribution of candidate lesions is available.

	No. of scans	LAD	LCX	RCA	OTHER_CAR	Candidate lesions 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

- 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.

2 | MATERIALS AND METHODS

This section presents our multitask model for the simultaneous segmentation of CARs and CACs. In Section 2.1, we describe the datasets we used and our annotation strategies. We introduce the multiclass CAR segmentation task (Section 2.2.1) and the binary lesion segmentation task (Section 2.2.2). We propose the multiloss optimization method using uncertainty-weighted loss (Section 2.2.3) and give a detailed description of our implemented network architecture and training procedure (Section 2.2.4). In Section 2.2.5, we introduce a single-task U-Net and sequential model, to which we compare the performance of our multitask model. In Section 2.3, 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.

2.1 | Datasets

Three different datasets were selected to test the performance of our MTL approach. A flowchart detailing the dataset selection process can be found in the Supporting Information.

Our first dataset, the DISCHARGE dataset, consists of calcium scoring CTs (CSCT) from 1262 patients (708 male, 554 female) enrolled in the DISCHARGE trial. The DISCHARGE trial is a prospective multicenter randomized controlled trial investigating for which patients with suspected coronary artery disease based on stable chest pain cardiac CT or cardiac catheterization are best suited as initial test.[21–23] In this trial, CT examinations were performed at 26 clinical sites using 14 different scanner types across Europe. Annotations for CAC were acquired for all scans. Weak annotations of CARs were only acquired for 215 randomly selected scans, which were 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 (57 452 slices) were used as test set. Only one CSCT from each patient was selected for the dataset. The CT scans were reconstructed using filtered back projection (383) and iterative reconstruction (879). To keep the data as close as possible to real-life clinical data, diagnostic CT scans with metal artifacts (pacemakers, artificial valves, etc.), severe motion artifacts, high noise levels, or anatomical abnormalities were not excluded. In the training and validation sets, both CARs and CACs are annotated. For the test set, only CAC annotations are available. Annotations were performed by two observers. Observer 1 was a trained physician certified by the Charité's cardiac CT training program[24], who annotated coronary calcifications. Observer 2 was a trained medical imaging scientist, who annotated CARs. Available contrast-enhanced CT scans (CCTA) were not included because the overall goal of the method is to predict the coronary heart disease risk from imaging studies without contrast agent administration.

The second dataset consists of CT scans from the publicly available orCaScore challenge on (semi-)automatic coronary calcium scoring.[25] The 72 pairs of CSCT and CCTA data were divided into a 32-scan training set and a 40-scan test set. For the training set, a reference standard established by two expert observers, a radiologist with >12 years of experience in CAC scoring and a research physician, is provided. CT scans with anatomical abnormalities, intracoronary stents, or metal implants as well as CTs showing severe motion artifacts or extremely high noise levels identified by visual inspection were excluded. For the training set, additional annotation of CARs was performed.

The third dataset consists of CSCT from the single-center randomized controlled Coronary Artery Disease

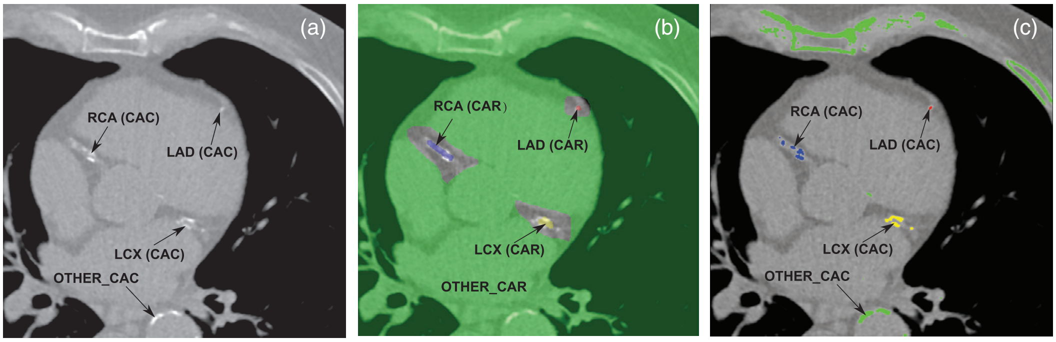

FIGURE 1 Multitask annotations of an image slice with coronary artery calcifications (CACs) in the left anterior descending artery (LAD), left circumflex artery (LCX), and right coronary artery (RCA) (A). Weak annotations of coronary artery regions (CARs) for the LAD - red, LCX - yellow, RCA - blue and OTHER_CAR - green (B). Strong annotations of coronary artery calcifications in the LAD - red, LCX - yellow, RCA - blue and other objects with attenuation > 130 HU (OTHER_CAC) - green (C).

Management (CAD-Man) study (further referred to as "CADMAN").[26] The dataset consists of 156 CT scans, in which only CACs were annotated. The dataset was used as an additional test set. CT scans were reconstructed using filtered backprojection. CAC identification by one expert observer was used as the reference standard. The annotations of coronary calcifications of the third dataset 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 was defined as connected 3-D image voxels (6-connectivity) with intensities greater than 130 Hounsfield units (HUs).

2.1.1 | Annotation procedure

For annotation, the coronary artery tree was divided into three subtrees (consisting of main branch and side branches): LAD artery, LCX, and RCA. The left main (LM) artery was included in the LAD subtree. CACs were annotated using an in-house semiautomatic segmentation module for 3D slicer[27]. Training of the multitask model required annotations for the two tasks of CAC segmentation and CAR segmentation. For illustration, an example of an annotated CT slice is shown in Figure 1.

2.1.2 | Annotation of coronary artery calcifications

Candidate lesions for annotation were identified using thresholding and highlighting all voxels with attenuation above 130 HU. For model evaluation, calcified lesions were defined as connected voxels (6-connectivity) with a minimum volume of 1.5 mm^3. The observer annotated all highlighted voxels of calcified lesions and assigned them to one of the three CARs ("LAD," "LCX," "RCA"). For calcified lesions affecting more than one artery (e.g., calcified lesions in bifurcations), individual voxels were classified. All candidate lesions not assigned to one of the three coronary artery segments were annotated as "OTHER_CAC."

2.1.3 | Weak annotation of coronary artery regions

Weak annotations (scribbles) were used for CARs because precise segmentation of arteries is impossible in unenhanced CT images due to poor contrast between arteries and surrounding tissue. To facilitate and speed up the annotation process, we used an in-house semiautomatic segmentation module developed for 3D Slicer[27]. To overcome the problem of misleading labels, we did not label regions between arteries and surrounding tissue that are difficult to distinguish or a precise labeling of the boundary would be extremely time-consuming. At first, the annotator used 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 joined 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 annotations are presented in Figure 1.

2.2 | Multitask segmentation network with uncertainty-weighted loss

We propose a multitask segmentation network based on an encoder–decoder structure with skip connections, inspired by the U-Net architecture.[28,29] The multitask

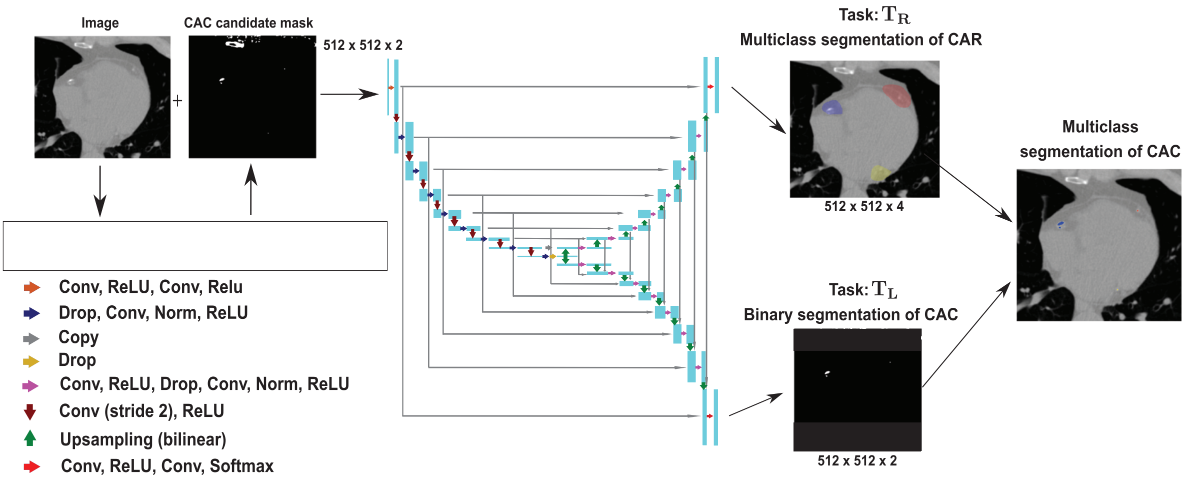

FIGURE 2 Multitask model for coronary artery calcification (CAC) scoring. The image and the CAC candidate lesion mask are concatenated to form the input tensor. The model consists of one encoder that shares feature maps with two decoders of the multiclass coronary artery region (CAR) segmentation task 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

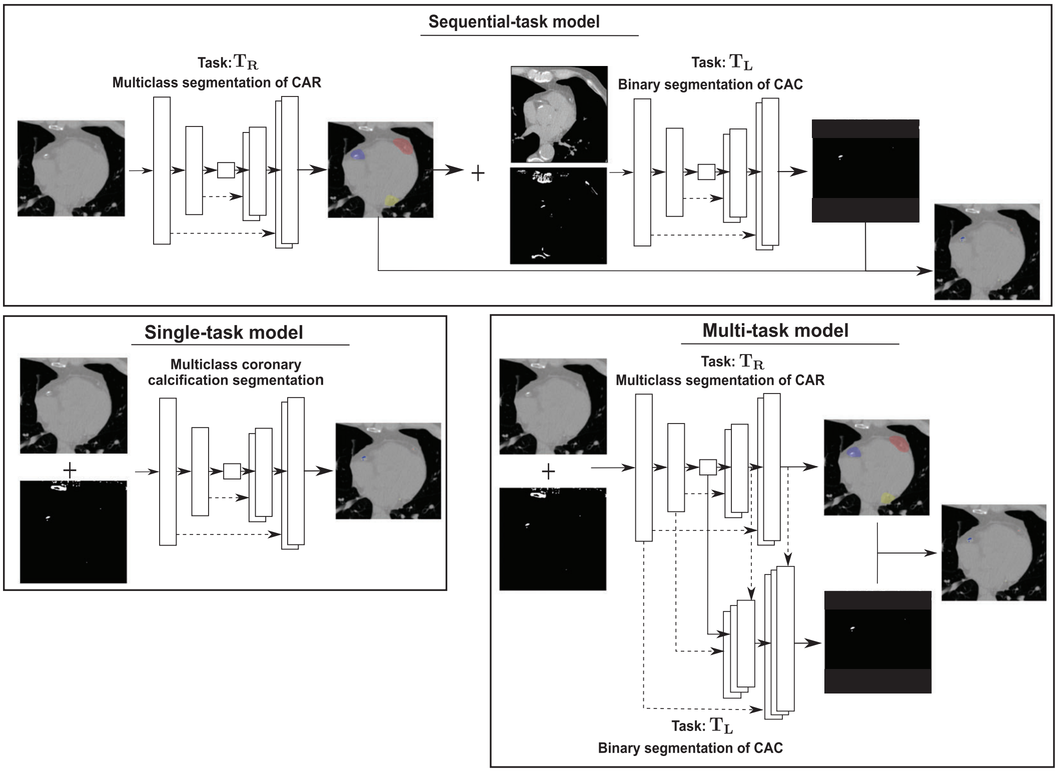
Sequential-task model
Task: $\mathrm{T_R}$
Multiclass segmentation of CAR
+
Task: $\mathrm{T_L}$
Binary segmentation of CAC
Single-task model
Multiclass coronary calcification segmentation
+
Multi-task model
Task: $\mathrm{T_R}$
Multiclass segmentation of CAR
+
Task: $\mathrm{T_L}$
Binary segmentation of CAC

FIGURE 3 Single-task model, sequential-task model, and multitask model architectures. The single-task model (bottom, left) consists of a multiclass U-Net. The sequential model (upper panel) 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 multitask model (bottom, right) consists of one encoder and two decoders for the prediction of CAR segmentations and CAC segmentations, which are combined for multiclass segmentation of CAC

network architecture is illustrated in Figure 2 and performs multiclass CAR segmentation and CAC segmentation at the same time. Features extracted by the encoder are shared with the two decoders for the tasks of multiclass CAR segmentation (T_R) and binary segmentation of calcified lesions (T_L). Since information on segmented CARs is useful prior information for segmentation of calcified lesions, feature maps extracted by the decoder for CAR segmentation are shared with the decoder for binary segmentation of calcified lesions. To utilize additional prior information about the location of candidate lesions, we concatenated the image slice (512 px × 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) were combined using an uncertainty-weighting loss[19] to jointly optimize the model parameters.

2.2.1 | Coronary artery region segmentation task

The network is trained to learn CARs from weakly labeled regions, as shown in Figure 1. Weak labels[30] are defined as segmentations that are imprecise but less costly to obtain than pixel-level annotations. Since, in unenhanced CT scans, spatial boundaries between vessels and surrounding tissue cannot be determined precisely, pixels x of an image i from the batch of size N are either annotated and belong to 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 function28 and focal loss[31] are used to deal with large class imbalances between background pixels (*OTHER_CAR*) and pixels of CARs.

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

The γ_R parameter smoothly adjusts the rate at which easily segmented pixels are downweighted, 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 1, are ignored and not used for loss calculation. Parameter w_{c_R} is a weighting parameter that balances the importance of the classes and handles the data imbalance problem. Parameter c_R is the channel of the corresponding CAR class.

2.2.2 | Binary lesion segmentation task

The lesion segmentation network performs a binary segmentation of candidate coronary artery lesions into the classes $\{CAC, OTHER\_CAR\}$. Feature maps extracted by the decoder for CAR segmentation are shared with the decoder for binary lesion segmentation. Binary focal loss, $\mathcal{L}_{L,i}$, defined by Equation (2), is calculated from all voxels of all candidate lesions grouped in set $\Omega_{L,i}$.

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

Parameters $p_{c_L}(x)$ and $y_{c_L}(x)$ as well as N, $x \in \Omega_{L,i}$ and w_{c_L} are defined analogously to region segmentation as outlined in Subsection 2.2.1. For multiclass CAC segmentation, the output of the binary CAC segmentation decoder is multiplied (channel-wise) by the output of the CAR segmentation decoder.

2.2.3 | Uncertainty-based weighted loss

How well a multitask network performs strongly depends on the weighting of losses. The most commonly used loss weighting strategy for MTL is static weighting, which computes a weighted sum of losses using balancing parameters α_i. The statically weighted loss of our multitask model is the weighted sum of the losses for multiclass segmentation of CARs $\mathcal{L}_R$ and for binary segmentation of coronary calcifications $\mathcal{L}_L$, as calculated according to 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.[32] Determination of the optimal weighting parameter value α is even more challenging in active MTL, because the model is initially trained with a very small number of annotated training data. Other methods use a DWA approach with task-specific feature-level attention[17] or gradient normalization[16] to balance losses.

For our multitask calcium scoring model, we use the uncertainty-weighted loss method proposed by Cipolla et al.[19]. Uncertainty-based weighting uses homoscedastic uncertainty to weight the loss functions of each task.[19] We used homoscedastic uncertainty to combine the outputs of the last layers (softmax output) from the decoders. To model the uncertainty, we introduced the positive scalar σ_R for the CAR segmentation task and σ_L for the binary calcification segmentation task. The parameters can be interpreted as Boltzmann distributions (also called Gibbs distribution) where input is scaled by σ_R^2 and σ_L^2, respectively. 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 multitask network. A detailed derivation can be found in the Supporting Information.

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

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

2.2.4 | Multitask network architecture and training procedure

To train the multitask network, we oversampled slices with calcifications to form balanced mini batches (20% samples with calcifications, 80% without calcifications). The encoder consisted of eight downsampling blocks, each block consisted of two convolutional layers, dropout layer, batch normalization[33], and ReLU activation function.[34] The two decoders consisted of eight upsampling blocks, where each block consisted of a bilinear upsampling layer and two convolution layers, dropout layer, batch normalization layer, and ReLU activation function. The feature maps yielded by the upsampling block of the CAR segmentation decoder were shared with each upsampling block of the coronary calcification segmentation decoder, but not vice versa, to maintain the causal relationship between CARs and CACs. Skip connections between the encoder and the two decoders were implemented as concatenations and used to share feature maps from the respective downsampling block.[28] The model was trained with a batch size of 8, the Adam[35] 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 statically weighted loss MTL model strongly depended on the initial learning rate but, due to long training times, we did not perform further detailed hyperparameter analysis. We used focal loss for both training tasks. The focal loss parameters of task T_R were set to $\gamma_R = 2.0, \alpha_{OTHER\_CAR} = 0.01, \alpha_{LAD} = 1.0$, $\alpha_{LCX} = 1.0$, and $\alpha_{RCA} = 1.0$. The loss parameters of the coronary calcification segmentation task were set to $\gamma_L = 2.0, \alpha_{CAC} = 1.0$, and $\alpha_{OTHER\_CAC} = 0.01$. To prevent overfitting, we augmented the input tensor ($512 \times 512 \times 2$) by applying translation augmentation.

For multiclass calcification segmentation, we combined the predictions of the two tasks by multiplying the binarized calcification segmentation by the CAR segmentation. To avoid overfitting, we used early stopping based on the performance of the validation set. The training stopped after approximately 200k iterations (240 epochs), where one iteration corresponded to a batch of eight 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).

2.2.5 | Single-task model, sequential-task model, and multitask model

In Figure 3, we compare our multitask model with a single-task and a sequential-task model, showing that simultaneous training of related tasks can extract informative shared features and improve model performance.

The single-task model was a multiclass U-Net[28] with the same downsampling and upsampling block architecture as used in the multitask network. The last layer consisted of four channels (*OTHER_CAC*, *LAD*, *LCX*, *RCA*) for multiclass segmentation of coronary calcifications. The sequential model combined two separated models. The first model was trained for multiclass CAR segmentation. After completion of training, the predictions were used for the training of the coronary calcification (CAC) segmentation network. Therefore, the CAR predictions were concatenated with the CT image slice and CAC candidate mask and served 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.

2.3 | Active learning with uncertainty-weighted multitask model

Labeling of coronary calcifications in CT scans is a laborious and time-consuming task and requires significant expert knowledge.[36] Labeling for MTL methods tends to be more expensive because each task requires its own annotations. Active learning can reduce costs by iteratively labeling only the most informative samples, thus achieving optimal performance with a smaller number of samples.[37] In MTL with static weighting parameter α, the parameter with the best performance has to be identified, which is a difficult and expensive process [19] and is often performed by a grid search of the entire annotated dataset. In active learning, estimation of a static weighting parameter α is even more challenging to tune, because data distribution changes after each sampling round, and hence, the optimal value of parameter α changes as well. Moreover, estimation on small datasets can be very sensitive to the randomly drawn initial training samples.

We simulated active learning to investigate whether our uncertainty-weighted loss model can overcome these problems. There are several approaches for active MTL including active learning via bandits,[38] active learning frameworks for adaptive filtering[39], and methods based on value of information.[40]

For our approach, we developed a hybrid sampling strategy based on uncertainty sampling and random sampling. First, we applied Monte Carlo dropout (MCD)[41] during inference for all samples that were not in the training set, predicted segmentation maps, and repeated this process $N_{MCD} = 10$ times. The primary goal of the dropout layer during training was to prevent overfitting. In uncertainty estimation, we were interested in capturing uncertainty with respect to all layers, and therefore applied a small dropout rate of 0.01 to all layers. Based on the predictions, we estimated MC sample variance[42] for each pixel, corresponding to candidate lesions (pixel with attenuation greater than 130 HU) and calculated average variance for each sample. We sorted all samples in descending order and randomly sampled from the top 20% with highest variance. Selected samples and their annotations were added to the training set. We use this simple strategy, because it was not our goal to improve sampling strategies, but rather to investigate their general applicability. This sampling strategy has low computational complexity and increases diversity of batch query samples.[37]. We compared our hybrid sampling method with the random sampling method using randomly select samples from the unlabeled dataset for labeling and training.

3 | RESULTS

In this section, we first briefly outline the performance metrics we used to compare the performance of our proposed multitask model with the singe-task U-Net and the sequential-task model trained on the full DISCHARGE training set (Subsection 3.2). In Subsection 3.3, we present the results of the comparison of our multitask model with other state-of-the-art models. Subsection 3.4 reports the results we achieved with our uncertainty-weighted multitask model in an active learning scenario, showing that the number of training samples needed and annotation time can be reduced compared to labeling the full training set. Finally, Subsection 3.5 presents our results for effects of image noise, metal artifacts, motion artifacts, and image quality, on model performance.

3.1 | Performance metrics

Coronary calcium scoring with the three models studied here (Table 2) was evaluated on the volume and the lesion level using binary and multiclass segmentation metrics[43]. The multiclass CAR segmentation task, T_R, was evaluated by determining the Micro F1-score on the volume level. The Micro F1-score is defined as the harmonic mean of microprecision and microrecall (5). For microprecision and microrecall, the number of true positives (TP_{sum}) was the number of all pixels correctly assigned to one of the three CARs and did not include pixels representing other candidate lesions. The number of false positives (FP_{sum}) was defined as the number of pixels belonging to the class *OTHER_CAR* but being misassigned to one of the coronary arteries plus all coronary artery pixels incorrectly assigned to another artery. The number of false negatives (FN_{sum}) was defined as the number of pixels belonging to the coronary arteries but being misclassified as *OTHER_CAR* plus all coronary artery pixels incorrectly assigned to another artery. Hence, misclassifications between arteries were counted as both false negatives and false positives.

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

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

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

Performance of the binary coronary calcification task, T_L, was evaluated as positive predictive value (PPV), sensitivity, and F1-score. The resulting multiclass calcification segmentation was evaluated using the binary F1-score calculated irrespective of the artery-specific label, to be comparable with other methods. For comparison with other methods, we calculated the ICC, sensitivity, and F1-score. The results are presented in Tables 4 and 3.

The Micro-F1 score of the resulting multiclass calcification segmentation was used to evaluate our active learning method. The results are presented in Figure 5.

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 risk category and the predicted risk categorization based on the MTL-model.

3.2 | Comparison of single-task, sequential-task, and multitask models

We trained all three models described in Section 2.2.5 on the full DISCHARGE training set and evaluated their performance using the DISCHARGE test set. Table 2 presents the results for the comparison of the CAR segmentation task, T_R, in terms of the Micro F1-score and of the binary coronary calcification (CAC) segmentation task, T_L, in terms of the F1-score, PPV, 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, because annotations of the DISCHARGE test set were not available for CAR. For the statically weighted MTL model, we set the weighting parameter to the optimal value of $\alpha = 0.4$, determined as the maximum Micro F1-score for calcification segmentation using the grid-search method. The uncertainty-weighted loss MTL model and the statically weighted MTL model with the

TABLE 2 Performance comparison of the single-task model (U-Net), sequential-task model, statically weighted MTL model, and uncertainty-weighted loss MTL model. Results are provided as 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 combined multiclass calcification segmentation in the DISCHARGE test dataset on volume and lesion levels.

	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
Statically 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 3 Performance comparison between our uncertainty-weighted MTL model and other state-of-the-art methods for automated coronary calcium scoring in cardiac CT on the orCaScore test set. Performance results are provided as 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 unenhanced 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	Interaction	Dataset No. of 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. [52]	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. [53]	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. [54]	Automatic	CSCT (32, 40)	0.989	0.835	0.951
Wolterink et al. [53]	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

optimal weighting parameter value achieved similar performance with an F1-score = 0.881 and F1-score = 0.882, respectively. Both MTL models outperformed the single-task model (F1-score = 0.804) and sequential-task model (F1-score = 0.769). Performance of the two multitask models was very good at the volume level, but lower at the lesion level, which was attributable to false positive predictions due to misclassification of noise. As expected, the sequential model (Micro F1-score = 0.472) performed best for CAR segmentation task T_R, because the first of the two sequential networks performed only CAR segmentation. Figure 4 illustrates the predictions of CAR and CAC by the multitask network.

Comparison of the multitask predictions for severe noise is presented in Figure S3 (Supporting Information), showing how the uncertainty-weighted loss MTL model outperformed the sequential model.

3.3 | Performance comparison with other methods

To compare our model (uncertainty-weighted loss MTL model) against other methods, we investigated performances in the orCaScore test set described in Section 2.1. 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 obtained with the uncertainty-weighted loss MTL model and other methods on the orCaScore test set are compiled in Table 3. The results achieved with our model on the DISCHARGE and CADMAN test sets are provided, along with the results of methods evaluated on other nonpublic datasets, in Table 4. Note

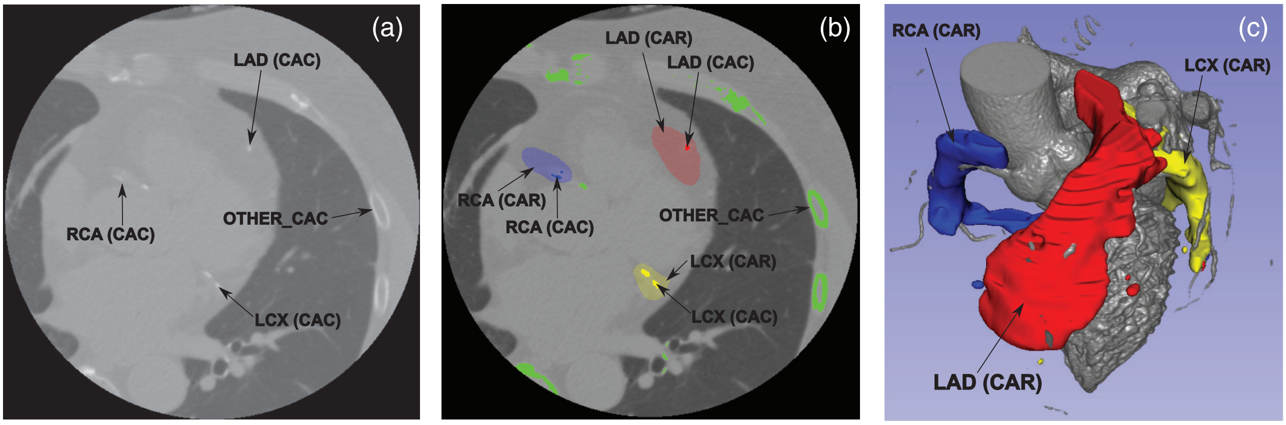

FIGURE 4 Visualization of overlap between predicted coronary calcifications (CAC) and coronary artery regions (CAR). Coronary calcifications in the left anterior descending artery (LAD), left circumflex artery (LCX), and right coronary artery (RCA) (A). Predicted coronary artery regions of the LAD -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

TABLE 4 Results of state-of-the-art methods for automated coronary calcium scoring in cardiac CT on nonpublic datasets. Results are compared in terms of interclass correlation coefficient (ICC), Sensitivity (Sen.), and F1-score. The method referred to as "Proposed network (DISCHARGE train)" is our uncertainty-weighted MTL model trained on the DISCHARGE training set. Performance of this network was investigated on the DISCHARGE and CADMAN test sets

Methods	Dataset # scans (train, test)	ICC (Vol.)	Sen. (Vol.)	F1 (Vol.)
Kurkure et al. [55]	CSCT (100, 105)	–	0.921	–
Isğum et al. [56]	CSCT (228, 76)	–	0.738	–
Brunner et al. [57]	CSCT (30, 30)	–	0.863	–
Shahzad et al. [52]	CSCT (209, 157)	–	0.839	–
Zhang et al. [8]	CSCT (129 with 5-fold CV)	0.986	0.905	0.946
Wolterink et al. [53]	CSCT (373, 530)	0.96	0.79	0.85
Vos et al. [13]	CSCT (373, 530)	0.97	–	–
Zeleznik et al. [50]	CSCT (129, (441, 663, 4021))	0.89, 0.80, 0.792	–	–
Velzen et al. [58]	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

that results are not directly comparable due to unknown data distributions.

The performance of our model trained on the DISCHARGE training set (F1-score = 0.958) was very good using only CSCTs. The best-performing method reported by Gogin et al.[7] (F1-score = 0.975) uses an ensemble of 3D CNNs for calcium scoring. Other methods use CCTA for segmentation of cardiac structures (heart, aorta, coronary arteries) and map the segmentations onto CSCT[7] or use preprocessing by cylindrical cropping around an initial automatic segmentation of the ascending aorta.[3]. Eng et al.[44] investigated two deep learning models to automate CAC scoring using gated unenhanced coronary CTs and nongated 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, performance was poorer than on the orCaScore test set (F1-score = 0.958). Similar dataset-dependent performance differences can be seen in Wolterink et al.[3] and Zhang et al.[8] listed in Tables 3 and 4. Possible factors influencing these results are discussed further in Section 3.5.

The per-patient risk categories were predicted from the estimated Agatston scores derived from the CAC segmentations and compared with the risk categories assigned on the basis of the reference annotations. The confusion matrices of risk category predictions and corresponding linearly weighted Cohen's kappa ($\varkappa$) for all three datasets are compiled in Table 5. We used a linearly weighted kappa because risk categories are on an ordinal rating scale and the deviations are weighted differently depending on their size. The results show that $\varkappa$ is much higher for the orCaScore dataset ($\varkappa = 0.97$)

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

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

	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, $\varkappa = 0.97$

	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, $\varkappa = 0.80$

	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

compared to the DISCHARGE ($\varkappa = 0.80$) or CADMAN dataset ($\varkappa = 0.80$). Misclassification of risk categories occurred mainly between categories I and II because of false positive predictions.

3.4 | Performance of the uncertainty-weighted loss MTL model in an active learning scenario

We conducted two experiments to investigate the performance of our uncertainty-weighted loss MTL model in an active learning scenario. In the first experiment, we analyzed model performance after training in an active learning scenario using two different loss weighting strategies (uncertainty-weighted loss, statically weighted loss) and sampling strategies (random sampling, hybrid sampling) described in Subsection 2.3. We initially trained the model with only 100 randomly selected samples (slices) and doubled the number of samples in each sampling round. Instead of retraining the model from scratch after each round, we continued training with the larger dataset and a reduced initial learning rate of 1e-04 compared to 5e-04 for the initial training. We used early stopping based on the validation set to avoid overtraining in each sampling round. Micro F1-scores for multiclass calcification segmentation were used to compare different models. As shown in Figure 5, with uncertainty-weighted loss and hybrid sampling, 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). All model variants were initially trained with the same, small dataset of only 100 slices, which leads to initial overfitting and explains the performance variability at the beginning of the active learning procedure.

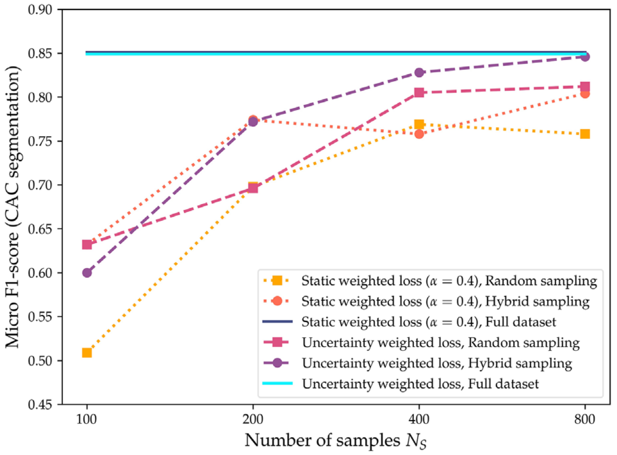

FIGURE 5 Performance comparison of the MTL model using two different loss weighting methods (statically weighted loss and uncertainty-weighted loss) and two different sampling methods (random sampling and hybrid sampling) in an active learning scenario

To assess MTL model performance for two different loss weighting and sampling strategies after three sampling rounds, we compared the proportion of model performance (compared to uncertainty-weighted model on the full dataset) in an active learning scenario. The results are presented in Table 6. The results show that the model with uncertainty-weighted loss outperformed use of static weighting for both random and hybrid sampling by 6.4% and 4.9%, respectively. This difference can be explained by the fact that data distribution in the training set changes in each sampling round and especially during the first sampling rounds. The uncertainty-weighted loss method can compensate for this distribution shift but statically weighted loss cannot. Moreover, hybrid sampling outperformed random sampling for both statically 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 the number of samples required.

Labeling of addition CAR annotations requires extra time, even if an efficient semiautomatic procedure as described in Section 2.1.3 is used. Annotation times required for (1) coronary calcifications, (2) coronary

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

	Random sampling	Hybrid sampling
Statically-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 7 Approximated annotation times for annotation of coronary artery calcifications (CACs) compared to annotation of coronary calcifications and coronary artery regions (CACs+CARs) and annotation of coronary calcifications and coronary artery regions using active learning (CACs+CARs+AL)

	Annot. time per slice [s]	Number of labeled slices	Annot. time training set [s]	Improvement ratio
CAC	4.0	6721	26 884	1.0
CAC + CAR	12.0	6721	80 652	3.0
CAC + CAR + AL	12.0	800	9600	**0.36**

calcifications and CARs, and (3) use of only informative slices for annotation of coronary calcifications and CARs were approximated empirically, and the results are shown in Table 7. The results show that annotation of CACs and CADs using active learning reduced annotation cost to approximately one-third compared to labeling of calcifications on the full training set, even though labeling both CACs and CADs is more time-consuming.

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 statically weighted loss model several times with values for weighting parameter α ranging from 0.1 to 0.9 and a step size of 0.1 using a very small randomly selected dataset (only 100 samples) and compared the results to the performance achieved with the model when trained on the full dataset. It turned out that the estimated optimal parameter for the full dataset $\alpha = 0.4$ was not the same as for the small dataset $\alpha = 0.2$ because the small dataset did not represent the data distribution of the full dataset. If a suboptimal parameter value had been selected after the first sampling round in the active learning scenario, optimal performance would not have been achieved. Alternatively, α could have been redetermined in each sampling round, but this would have been computationally very expensive. A detailed analysis is presented in Figure S2 in the Supporting Information.

3.5 | Influence of image noise, metal artifacts, motion artifacts, and image quality on model performance

Model performance was 4.6% higher when CT scans with severe image noise, metal artifacts, motion artifacts, and poor image quality were excluded. The results summarized in Tables 3 and 4 show much better performance on the orCaScore test set (Micro F1-score = 0.961) than on the DISCHARGE dataset (Micro F1-score = 0.881). To explain differences in performance, we analyzed possible effects of four factors: (1) image noise, (2) metal artifacts, (3) motion artifacts, and (4) image quality. We estimated image noise using a method similar to Christianson et al.[45] The method consisted of a series of steps: first, we segmented the CT image into the heart-related tissue types (–200 to 140 HU); second, a noise image filter[46] was applied to the segmented region; third, a histogram was generated and the highest peak was selected as the noise level.[45] The noise levels of all CT scans of the test dataset were normalized using z-score[47] and the most noisy 20% were labeled as noisy CT scans. Presence versus absence of metal artifacts and motion artifacts was assessed visually, and scans were labeled accordingly. Image quality was visually classified and labeled as good versus poor, when a high amount of disturbances or anatomical abnormalities was present. Note that none of the CT scans in the test set was classified as nondiagnostic (i.e., unsatisfactory for diagnosis) by a radiologist.

Examples of the four influencing factors investigated are shown in Figure 6.

Micro F1-scores ranged from 0.881 when all scans were included in the test set to 0.927 when scans with image noise, metal or motion artifacts, and poor image quality were excluded. When only noisy images were excluded, performance increased by 3.1%. Surprisingly, when we excluded images with severe motion artifacts, performance dropped only by 0.02%. This can be explained by the fact that, when motion artifacts are present, calcifications appear very large, resulting in a high number of “falsely” labeled true positives in the dataset. Excluding samples with motion artifacts decreased the number of true positives and thus the corresponding Micro F1-score. Detailed results of this subanalysis are presented in the Supporting Information.

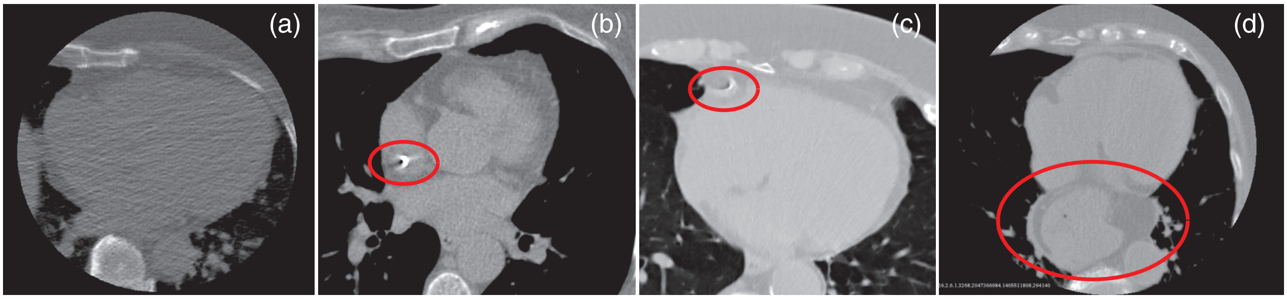

FIGURE 6 Examples of CT scans with severe image noise (A), metal artifacts (B), motion artifacts (C), and poor image quality (abnormality provoked by hiatal hernias) (D)

4 | DISCUSSION

In this paper, we have proposed an MTL model with uncertainty-weighted loss for coronary calcium scoring in ECG-gated, unenhanced cardiac CTs. The MTL 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 CARs based on weak annotations and segmentation of coronary calcifications (CACs) in an end-to-end framework. We compared different versions of our MTL model with a single-task model (multiclass U-Net) and a sequential model. Our results (Table 2) demonstrate that the MTL model outperforms other models in terms of PPV, sensitivity, F1-score, and Micro F1-score. Advantages of an MTL model over a multiclass U-Net and a sequential model are the shared information between CAR segmentation task T_R and calcification segmentation task T_L. Unlike the multiclass U-Net, the MTL model learns important spatial information from weakly labeled samples and can transfer this knowledge for segmentation of coronary calcifications. Explanation techniques such as layer-wise relevance propagation[48] could lead to a deeper understanding of different prediction strategies but they are beyond the scope of this paper.

Uncertainty-weighted loss and statically weighted loss lead to similar performance of the MTL model when using optimal weighting parameter. However, determination of the optimal weighting parameter value is a challenging and expensive process and even more difficult in an active learning scenario.

To reduce labeling costs, we investigated our uncertainty-weighted MLT network in an active learning scenario and could show that our model reaches optimal performance with substantially fewer 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 (Supporting Information). Unlike Gong et al.[49], 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). 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, because it contains only filtered back projections. An effect of noise is also reflected in a higher number of lesion candidates per scan we found for the CADMAN and DISCHARGE datasets.

The predictions of CVD risk categories based on the segmentations in Table 5 show very good agreement of $\varkappa = 0.97$ for the orCaScore dataset but lower agreement of $\varkappa = 0.80$ for the DISCHARGE test set. Mislabeled noise leads to a high false positive rate between risk category I (total Agatston score of 0) and II (total Agatston score of 1–100), which is in agreement with the findings reported by Zeleznik et al.[50] We also trained our model on the orCaScore training set with additional annotations for CARs 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. Overall, our results show that image noise strongly affects model performance besides presence of metal artifacts and poor image quality. Conversely, motion artifacts have no effect on model 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.

4.1 | Limitations

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

Our method processes 2D axial CT slices. Use of 3D-information might be beneficial as shown by Zhang et al.[8] and recently published methods based on 3D-CNN ensembles[7] yielded promising results. Since our active learning approach is based on the labeling of only the most informative slices 3D-annotations would be sparse. Learning dense 3D segmentations from sparse annotations can be challenging in a multitask network[51]. Nevertheless, we plan to investigate how our method can be extended to 3D in future work.

Furthermore, reference standards for the DISCHARGE and CADMAN datasets were provided by only one experienced observer for both coronary calcifications and CARs. Independent annotations by a second observer and clarification of discrepancies by consensus would have improved the quality of the dataset. This will be done in a future research project.

We analyzed only four factors (image noise, metal artifacts, motion artifacts, image quality) that might affect model performance, yet other factors such as reconstruction method, CT scanner type, and slice thickness or slice spacing are known to influence model performance but were beyond the scope of this work.

4.2 | Future 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 in terms of F1-scores, ICC, and sensitivity of CAC volume but lacks precision regarding CVD risk agreement. In a future study, we will investigate how risk categorization can be improved by integrating direct prediction of risk categories[13] into our model. A major focus will be on improving the identification of patients with zero calcium scores. We also plan to extend our model from 2D input data to 3D to take advantage of 3D context information and overcome limitations of our current model. 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, which can be addressed in future work. Other 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.

5 | CONCLUSION

In this work, we have proposed an MTL 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 CAR segmentation and CAC segmentation. Model performance was evaluated using a large multicenter dataset of the DISCHARGE trial (1047 CSCTs), a single-center dataset of the CAD-Man study (156 CSCTs), and the multicenter 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 smaller number of required image slices needed might enable the training of models that can be used in a clinical setting.

ACKNOWLEDGMENT

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) and all authors of the DISCHARGE publication[23].

Acknowledgment should also be provided for the CAD-MAN Trial Group (Dewey et al.)[26] for the collection and provision of the data from the CAD-MAN project (ClinicalTrials.gov Identifier: NCT00844220) funded by a grant of the Heisenberg program of the German Research Foundation to Marc Dewey.

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

Acknowledgment should also be provided to Bettina Herwig for proofreading the manuscript.

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.

Open access funding enabled and organized by Projekt DEAL.

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. As 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.

REFERENCES

1. WHO. *Cardiovascular diseases*. Fact sheet 317, World Health Organization; 2021.
2. Latif MA, Budoff MJ, Greenland P. *Coronary Artery Calcium*. Springer; 2014:181-189.
3. Wolterink JM, Leiner T, De Vos BD, et al. An evaluation of automatic coronary artery calcium scoring methods with cardiac CT using the orCaScore framework. *Med Phys*. 2016;43:2361-2373.
4. Santini G, Della Latta D., Martini N, et al. An automatic deep learning approach for coronary artery calcium segmentation. *IFMBE Proc*. 2017;65:374-377.
5. Agatston AS, Janowitz FWR, Hildner FJ, Zusmer NR, Viamonte M, Detrano R. Quantification coronary artery calcium using ultrafast computed tomography. *J Am Coll Cardiol*. 1990;15:827-832.
6. Lessmann N, Ginneken BV, Zreik M, et al. Automatic calcium scoring in low-dose chest CT using deep neural networks with dilated convolutions. *IEEE Trans Med Imaging*. 2018;37:615-625.
7. Gogin N, Viti M, Nicodème L, et al. Automatic coronary artery calcium scoring from unenhanced-ECG-gated CT using deep learning. *Diagn Interv Imaging*. 2021;102:683-690.
8. Zhang W, Zhang J, Du X, Zhang Y, Li S. An end-to-end joint learning framework of artery-specific coronary calcium scoring in non-contrast cardiac CT. *Computing*. 2019;101:667-678.
9. Wolterink JM, Leiner T, Vos BD., Hamersvelt RW, Viergever MA, Išgum I. Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks. *Med Image Anal*. 2016;34:123-136.
10. Yang G, Chen Y, Ning X, Sun Q, Shu H, Coatrieux J-L. Automatic coronary calcium scoring using noncontrast and contrast CT images. *Med Phys*. 2016;43:2174-2186.
11. van Velzen SGM, Hampe N, de Vos BD, Išgum I*AI for Calcium Scoring*. https://doi.org/10.48550/ARXIV.2105.12558
12. Shahzad R, van Vliet L, Niessen WJ, Walsum T. Automatic classification of calcification in the coronary vessel tree. https://doi.org/10.13140/2.1.1787.8088
13. Vos BD, Wolterink JM, Leiner T, Jong PA, Lessmann N, Isgum I. Direct automatic coronary calcium scoring in cardiac and chest CT. *IEEE Trans Med Imaging*. 2019;38:2127-2138.
14. Smailagic A, Noh HY, Costa P, et al. MedAL: accurate and robust deep active learning for medical image analysis. arXiv, 2018.
15. Ruder S. An overview of multi-task learning in deep neural networks arXiv: 1706. 05098v1 [cs . LG]. 2017.
16. Chen Z, Badrinarayanan V, Lee CY, Rabinovich A. GradNorm: gradient normalization for adaptive loss balancing in deep multitask networks. In: *35th International Conference on Machine Learning, ICML 2018*. Vol. 2. 2018:1240-1251.
17. Liu S, Johns E, Davison AJ. End-to-end multi-task learning with attention. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2019:1871-1880.
18. Guo M, Haque A, Huang D-A, Yeung S, Fei-Fei L. Dynamic task prioritization for multitask learning. In: *15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*. 2018:282-299.
19. Cipolla R, Gal Y, Kendall A. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2018:7482-7491.
20. Vandenhende S, Georgoulis S, Van Gansbeke W, Proesmans M, Dai D, Van Gool L. Multi-task learning for dense prediction tasks: a survey. *IEEE Trans Pattern Anal Mach Intell*. 2022;44(7):3614-3633. https://doi.org/10.1109/TPAMI.2021.3054719
21. Dewey M. The Discharge trial. https://www.dischargetrial.eu/. 2021.
22. Napp AE, Haase R, Laule M, et al. Computed tomography versus invasive coronary angiography: design and methods of the pragmatic randomised multicentre DISCHARGE trial. *Eur Radiol*. 2017;27:2957-2968.
23. Maurovich-Horvat P, Bosserdt M, Kofoed KF, et al. CT or Invasive coronary angiography in stable chest pain. *N Engl J Med*. 2022;386:1591-1602.
24. Zimmermann E, Germershausen C, Greupner J, et al. Improvement of skills and knowledge by a hands-on cardiac CT course: before and after evaluation with a validated questionnaire and self-assessment. *RoFo: Fortschr Geb Rontgenstr Nuklearmed*. 2010;182:589-593.
25. Jelmer M, Wolterink BDdV, Tim L, Max AV, Ivana I. orCaScore https://orcascore.grand-challenge.org/. 2021.
26. Dewey M, Rief M, Martus P, et al. Evaluation of computed tomography in patients with atypical angina or chest pain clinically referred for invasive coronary angiography: randomised controlled trial. *BMJ (Clinical research ed.)*. 2016;355:i5441.

27. Jolesz FA. *Intraoperative Imaging and Image- Guided Therapy*. Springer; 2014.
28. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI) 9351 of LNCS*. Springer; 2015:234-241.
29. Ke R, Bugeau A, Papadakis N, Schuetz P, Schönlieb CB. 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)*. LNCS, vol. 12535. Springer Science and Business Media Deutschland GmbH; 2020:411-428.
30. Papandreou G, Chen L-c, Murphy K, Yuille AL. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: *IEEE International Conference on Computer Vision*. 2015.
31. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. *IEEE Trans Pattern Anal Mach Intell*. 2020;42:318-327.
32. Kongyoung S, Macdonald C, Ounis I. Multi-Task Learning using Dynamic Task Weighting for Conversational Question Answering. In: *Proceedings of the 5th International Workshop on Search-Oriented Conversational AI (SCAI)*, 2020:17-26. https://doi.org/10.18653/v1/2020.scai-1.3
33. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: *32nd International Conference on Machine Learning, ICML 2015*. 2015;1:448-456.
34. Brown MJ, Hutchinson LA, Rainbow MJ, Deluzio KJ, De Asha AR. Rectified linear units improve restricted Boltzmann machines. *J Appl Biomech*. 2017;33:384-387.
35. Kingma DP, Ba JL. Adam: a method for stochastic optimization. In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. 2015: 1-15.
36. Budd S, Robinson EC, Kainz B. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Med Image Anal*. 2021;71:102062.
37. Ren P, Xiao Y, Chen X, et al. *A Survey of Deep Active Learning*. Technical Report. 2020.
38. Fang M, Tao D. Active multi-task learning via bandits. In: *SIAM International Conference on Data Mining 2015, SDM 2015*. 2015:505-513.
39. Harpale A, Yang Y. Active learning for multi-task adaptive filtering. In: *ICML 2010 - Proceedings, 27th International Conference on Machine Learning*. 2010:431-438.
40. Zhang Y. Multi-task active learning with output constraints. In: *Proceedings of the National Conference on Artificial Intelligence*. 2010;1:667-672.
41. Gal Y, Ghahramani Z. 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; 2016:1050-1059JMLR.org.
42. Nair T, Precup D, Arnold DL, Arbel T. 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)*. LNCS. Springer-Verlag; 2018;11070:655-663.
43. Grandini M, Bagli E, Visani G. Metrics for multi-class classification: an overview. arXiv. 2020:1-17.
44. Eng D, Chute C, Khandwala N, et al. Automated coronary calcium scoring using deep learning with multicenter external validation. *npj Digital Med*. 2021;4.
45. Christianson O, Winslow J, Frush DP, Samei E. Automated technique to measure noise in clinical CT examinations. *Am J Roentgenol*. 2015;205:W93-W99.
46. SimpleITK. SimpleITK. https://simpleitk.org. 2021.
47. Sylvan D. Introduction to Mathematical Statistics. *J Biopharm Stat* 2013;23:716-717. https://doi.org/10.1080/10543406.2013.756334
48. Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One*. 2015;10:e0130140.
49. Gong T, Lee T, Stephenson C, et al. A comparison of loss weighting strategies for multi task learning in deep neural networks. *IEEE Access*. 2019;7:141627-141632.
50. Zeleznik R, Foldyna B, Eslami P, et al. Deep convolutional neural networks to predict cardiovascular risk from computed tomography *Nat Commun*. 2021;12:715.
51. Bokhorst J-M, Pinckaers H, Van Zwam P, et al. Learning from sparsely annotated data for semantic segmentation in histopathology images. *Proceedings of Machine Learning Research*. 2019;102:84-91.
52. Shahzad R, Walsum TV, Schaap M, et al. Vessel specific coronary artery calcium scoring: an automatic system. *Academic Radiol*. 2009;20:1-9.
53. Wolterink JM, Leiner T, Takx RAP, Viergever MA, Išgum I. Automatic coronary calcium scoring in non-contrast-enhanced ECG-triggered cardiac CT with ambiguity detection. *Tmi*. 2015;34:1867-1878.
54. Durlak F, Wels M, Schwemmer C, Sühling M, Steidl S, Maier A. Growing a random forest with fuzzy spatial features for fully automatic artery-specific coronary calcium scoring. In: Wang Q, Shi Y, Suk H-Il, Suzuki K, eds. *Machine Learning in Medical Imaging*. Springer International Publishing; 2017:27-35.
55. Kurkure U, Chittajallu DR, Brunner G, Le YH, Kakadiaris IA. A supervised classification-based method for coronary calcium detection in non-contrast CT. *Int J Cardiovasc Imaging*. 2010;26:817-828.
56. Išgum I, Rutten A, Prokop M, Ginneken B. Detection of coronary calcifications from computed tomography scans for automated risk assessment of coronary artery disease. *Med Phys*. 2007;34:1450-1461.
57. Brunner G, Chittajallu DR, Kurkure U, Kakadiaris IA. Toward the automatic detection of coronary artery calcification in non-contrast computed tomography data. *Int J Cardiovasc Imaging*. 2010;26:829-838.
58. Velzen SGMV, Lessmann N, Velthuis BK. Deep learning for automatic calcium scoring in CT: validation using multiple cardiac CT and chest CT protocols. *Radiology*. 2020;295:66-79.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Föllmer B, Biavati F, Wald C, et al. Active multitask learning with uncertainty-weighted loss for coronary calcium scoring. *Med Phys*. 2022;49:7262–7277. https://doi.org/10.1002/mp.15870