Computational and Structural Biotechnology Journal 20 (2022) 1681–1690

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Automated coronary artery calcium scoring using nested U-Net and focal loss

Jia-Sheng Hong [a,b,1], Yun-Hsuan Tzeng [c,d,1], Wei-Hsian Yin [c,e], Kuan-Ting Wu [a,c], Huan-Yu Hsu [a,c], Chia-Feng Lu [b], Ho-Ren Liu [d,*], Yu-Te Wu [a,f,*]

[a] Institute of Biophotonics, National Yang Ming Chiao Tung University, Taipei 112, Taiwan
[b] Department of Biomedical Imaging and Radiological Sciences, National Yang Ming Chiao Tung University, Taipei 112, Taiwan
[c] School of Medicine, National Yang Ming Chiao Tung University, Taipei 112, Taiwan
[d] Division of Advanced Medical Imaging, Health Management Center, Cheng Hsin General Hospital, Taipei 112, Taiwan
[e] Heart Center, Cheng Hsin General Hospital, Taipei 112, Taiwan
[f] Brain Research Center, National Yang Ming Chiao Tung University, Taipei 112, Taiwan

ARTICLE INFO

Article history:
Received 17 August 2021
Received in revised form 24 March 2022
Accepted 24 March 2022
Available online 26 March 2022

Keywords:
Coronary artery calcium scoring
Multidetector computed tomography
U-Net++
Focal loss
Automated coronary artery calcium detection

ABSTRACT

Coronary artery calcium (CAC) is a great risk predictor of the atherosclerotic cardiovascular disease and CAC scores can be used to stratify the risk of heart disease. Current clinical analysis of CAC is performed using onsite semiautomated software. This semiautomated CAC analysis requires experienced radiologists and radiologic technologists and is both demanding and time-consuming. The purpose of this study is to develop a fully automated CAC detection model that can quantify CAC scores. A total of 1,811 cases of cardiac examinations involving contrast-free multidetector computed tomography were retrospectively collected. We divided the database into the Training Data Set, Validation Data Set, Testing Data Set 1, and Testing Data Set 2. The Training, Validation, and Testing Data Set 1 contained cases with clinically detected CAC; Testing Data Set 2 contained those without detected calcium. The intraclass correlation coefficients between the overall standard and model-predicted scores were 1.00 for both the Training Data Set and Testing Data Set 1. In Testing Data Set 2, the model was able to detect clinically undetected cases of mild calcium. The results suggested that the proposed model's automated detection of CAC was highly consistent with clinical semiautomated CAC analysis. The proposed model demonstrated potential for clinical applications that can improve the quality of CAC risk stratification.

1. Introduction

In 2019, cardiovascular diseases accounted for one-third of all deaths worldwide, and nearly 90% of these deaths were due to heart attack and stroke [1,2]. Atherosclerosis is one of the leading causes of cardiovascular disease, and the presence of coronary artery calcium (CAC) indicates its progressive risk. Therefore, screening is crucial for early detection and risk stratification of potential atherosclerotic cardiovascular disease. Computed tomography (CT) is the clinical standard modality for the non-invasive screening of CAC. Assessment of CAC is performed by electrocardiogram-gated CT without using contrast medium. Due to the characteristics of calcified plaques on CT images, a standardized quantitative calcium score, known as the Agatston score, was proposed and is now widely used in clinical practice to assess the risk of cardiovascular disease [3]. The calcification area on CT images has been proven to correlate with the pathological lesion size [4], and the calcium score is sensitive in detecting vascular stenosis [5]. Because the calcium score is closely associated with the overall levels of atherosclerotic plaques, it is used to assess the risk of coronary events [6–10]. Studies have demonstrated the importance of the calcium score as an indicator in cardiovascular disease risk stratification.

Currently, clinical analysis of calcium scores is performed in a semiautomated manner by radiologists and radiologic technologists, who use CAC analysis software to manually outline possible calcium locations on images. The software uses a threshold of 130

* Corresponding authors at: Institute of Biophotonics, National Yang Ming Chiao Tung University, No.155, Sec. 2, Linong St., Beitou Dist., Taipei City 112, Taiwan (Y.T. Wu). Health Management Center, Cheng Hsin General Hospital, No. 45, Zhenxing Street, Beitou District, Taipei City, 112, Taiwan (H.R. Liu).
E-mail addresses: ch6752@chgh.org.tw (H.-R. Liu), ytwu@nycu.edu.tw (Y.-T. Wu).
[1] Co-first Author: These authors contributed equally.

https://doi.org/10.1016/j.csbj.2022.03.025

Hounsfield units (HU) to segment possible calcium locations from areas that have been manually framed. However, this approach requires a complete review of all cardiac CT images, which is labor-intensive and time-consuming. Therefore, the purpose of this study is to establish an automated CAC detection model to reduce the burden of manual quantification and thus improve the quality of risk management in healthcare.

The application of artificial intelligence in the automated segmentation of medical images has matured markedly. Among them, U-Net is a well-known deep learning framework based on the convolution neural network [11]. Convolution neural network is a type of artificial neural network in deep learning, and it is powerful in tasks for image processing. It uses convolution and backpropagation to learn features and update model parameters automatically from the data. U-Net has been applied for the automated segmentation of medical images, demonstrating favorable performance. There have been studies using U-Net-based deep learning models for automatic detection of CAC. Gogin et al. used a three-dimensional (3D) architecture–based U-Net model to detect the CAC [12], and Zhang et al. used a multiview input architecture to train the U-Net-based model [13]. U-Net is based on encoder and decoder architecture composed of convolution neural networks. The encoder captures the features, and the image dimension reduces after passing the encoder. The direct skip connection that concatenates feature maps from the encoder to decoder in U-Net helps the decoder recover the target objects' fine-grained details to restore spatial information. However, the direct fast-forward concatenation connects semantically dissimilar feature maps, degrading segmentation performance. U-Net++ is an improved version of U-Net that can fully capture the extracted features at different scales using a dense skip connection [14]. The dense skip connection can enrich the feature maps from the encoder, making it semantically similar to the corresponding feature maps of the decoder. The performance of U-Net++ in various medical image segmentation tasks has been compared with that of U-Net, with results verifying that U-Net++ outperforms U-Net. In addition, although not using a U-Net-based model, Zhao et al. used the medical prior knowledge to increase the channels of the input images [15]. They binarized the CT images by a clinical threshold of 130 HU used to determine CAC. The binary image became an augmented input, allowing the deep learning model to detect CAC with improved accuracy.

The voxel number of CAC on a CT image is substantially smaller than the non-CAC voxel number. This results in a severe positive and negative sample imbalance, which renders effective learning of the positive classes challenging for deep learning models. Focal loss is the modification from cross entropy loss [16], which allows the model to focus more on learning difficult samples by reducing the weight of easy-to-classify samples. Thus, the inefficient learning problem caused by the unbalanced positive and negative samples can be alleviated. In this study, we used the focal loss to train the model. We tested whether U-Net++ was better than U-Net in the segmentation of CAC. To the best of our knowledge, no study has been conducted to model the segmentation task of CAC using U-Net++. We retrospectively collected clinical routine cases of coronary calcium scans, including CT and manually segmented CAC images. The balanced sampling with focal loss was applied to optimize the model during training. The aim of this study was to develop an automated model for detecting and segmenting CAC in CT images. Eventually, the performance of the model was verified by comparing the CAC scores obtained by semiautomated clinical software with those obtained by the proposed model.

2. Materials and methods

2.1. Subjects and data collection

This study involved the retrospective collection of clinical routine images and was approved by the institutional review board of Cheng Hsin General Hospital ([807] 109A-46). A total of 1,811 CAC scans (from between November 2015 and January 2021) were retrieved from the hospital database. A total of 1,067 scans had a nonzero CAC score, and 744 scans had a CAC score of zero. An experienced radiologic technologist checked the data with nonzero CAC scores to ensure the accuracy of the data set. The 1,067 scans were divided into Training, Validation, and Testing Data Sets, with 754, 98, and 215 scans, respectively. Those 215 scans were termed Testing Data Set 1. Approximately 70% of the data was used to train the model, 10% was used to validate the optimal model, and the remaining 20% was used to test the model's performance in detection and segmentation of the unknown data. That series of 744 scans, which was denoted as Testing Data Set 2, with a CAC score of zero, was used to examine the sensitivity of the trained model in detecting CAC. Table 1 is the demographics of the data sets. Fig. 1 is a simple flowchart describing the mode of data separation.

2.2. Cardiac CT and semiautomated calcium analysis

All images were acquired using a dual-source 128-slice multidetector CT scanner (Somatom Definition FLASH, Siemens Healthineers, Erlangen, Germany). Clinical routine multidetector CT scans were performed with electrocardiogram gating and covered the entire heart region. Automated tube current modulation (CARE Dose4D, Siemens Healthineers) was conducted with a voltage of 120 kVp at the reference of 80 mAs. The rotation time was 0.28 s with the collimation of 128 $\times$ 0.6 mm and 2-mm slice thickness. The scan direction was craniocaudal and the pitch depended on the heart rate. For heart rates above 75 bpm and below 65 bpm, the phase was 40% and 70%, respectively, for electrocardiogram reconstruction. The semiautomated CAC scores obtained manually in clinical practice were recorded as per standard practice in hospital onsite software (syngo.CT CaScoring, Siemens Healthineers).

2.3. Quantitative calcium scores and CAC risk stratification

To calculate calcium scores, we first located the connected calcium voxels in three dimensions in the binarized calcium image. Voxels connected in six directions (anterior, posterior, left, right, superior, and inferior) were considered as a single calcified lesion. A total lesion area of <1 mm^2 was considered as noise and removed. After segmenting the calcified lesions, three quantitative

Table 1
Demographics of the CAC scans in the training and testing data sets.

	Total	Training	Validation	Testing 1	Testing 2
Number of Scans	1,811	754	98	215	744
Mean Age at Acquisition	58.1 (18–96)	62.5 (32–96)	61.6 (40–82)	61.5 (29–85)	52.1 (18–81)
Gender (Male/Female)	1,132/679 (63/37)	519/235 (69/31)	67/31 (68/32)	158/57 (73/27)	388/356 (52/48)

Values in brackets are the minimum and maximum values for age and percentages for gender.

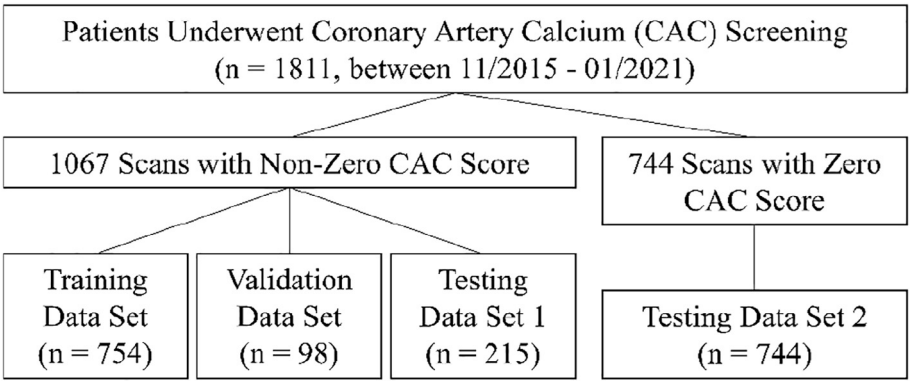


Fig. 1. Data separation flowchart.

scores were calculated, namely, the Agatston score, volume score, and mass score [3,17,18].

Fig. 2 shows an example of quantitative calcium score calculation. The equation of each score is as follows:

$$\text{Agatston Score} = W \times \text{Pixels} \times \text{Pixel Size} \quad (1)$$

$$\text{Volume Score} = \text{Pixels} \times \text{Pixel Size} \times \text{Slice Thickness, and} \quad (2)$$

$$\text{Mass Score} = C \times HU_{mean} \times \text{Pixels} \times \text{Pixel Size} \times \text{Slice Thickness.}$$

The Agatston score is calculated by multiplying the calcified area by a weighting factor *W*. *W* in Eq. (1) is derived from the maximum intensity within a calcified lesion; *W* is 1, 2, 3, and 4 when the HU value is 130–199, 200–299, 300–399, and ≥ 400, respectively. To calculate the volume score, the number of image pixels is multiplied by the pixel size and then by the slice thickness. Eq. (2) provides the total volume of the calcified lesion. The mass score is calculated by multiplying the calibration coefficient *C* by the mean HU value within the calcified lesion and then by the total volume. In Eq. 3, the calibration coefficient *C* is obtained clinically through a calibration phantom.

The scores for all calcified lesions were eventually summarized to produce an overall quantitative calcium score for the risk stratification of CAC. The overall Agatston score was divided into five categories according to the score values [19,20], as displayed in Table 2. The first category (C1) contained cases with a score of 0 (i.e., no calcium observed). As the score increased, the more severe the calcium area and the higher the degree of coronary atherosclerosis and risk of cardiovascular disease. The highest-risk category was that defined by an Agathon score of over 400 (C5).

2.4. Experimental setup and deep learning framework

All experiments were conducted on a computer with an Intel i7-9700 central processing unit and a NVIDIA GeForce RTX 3090 24 gigabyte graphics processing unit. The deep learning model was constructed based on PyTorch (version 1.7.1 + cu110), which uses the Python 3.8.5 programming language. Image processing was performed on the MATLAB R2020a programming platform (MathWorks; Natick, MA, USA). Statistical analysis was performed using SPSS Statistics 24 (IBM; Armonk, NY, USA).

Table 2
Summary of CAC risk stratification based on Agatston score.

Class	Agatston score	Plaque Burden
C1	0	None
C2	1–10	Minimal
C3	10–100	Mild
C4	100–400	Moderate
C5	>400	Extensive

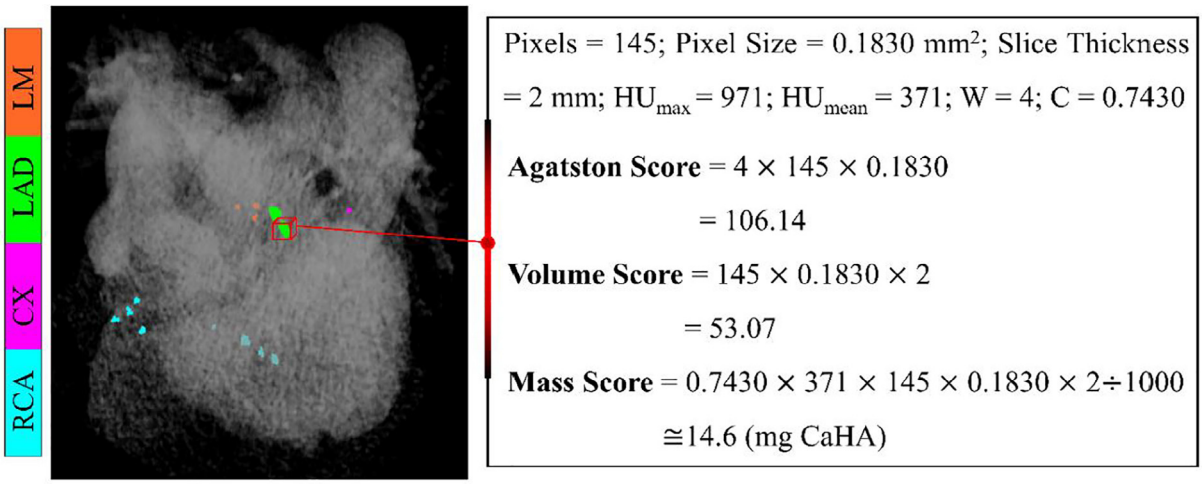


Fig. 2. Example of the calculation of quantitative calcium scores. In calculating the mass score, it was divided by 1000 to convert the volume from mm^3 to cm^3 because the unit of calibration coefficient is (mg CaHA)/(HU·cm^3).

The model training workflow is depicted in Fig. 3a. The Training Data Set was sampled using the Imbalanced Dataset Sampler for balancing positive and negative data [21]. The sampler made the positive and negative samples 1:1 ratio for each training batch by oversampling. Images without the calcified lesion were considered the negative, and those with the calcified lesion were considered the positive. After the data were passed through the sampler, the number of positive and negative samples that entered into the model was the same. Training of the model ceased when the difference between the focal loss of the model for Validation Data Set and the Training Data Set was greater than a certain value. We evaluated the Agatston scores of the Training Data Set and Validation Data Set during training. Among all trained weights, the one with the smallest difference in the Agatston score between the standard and the model-predicted scores (for Validation Data Set) was selected as the final model. Thereafter, Testing Data Set 1 and Testing Data Set 2 was inputted to the final model for the evaluation.

In addition to the original CT images as the model input, the binarized images with HU values greater than 130 as threshold served as the second channel for training input [15], as illustrated in Fig. 3b. As observed in the figure, the binarized image contained calcified lesions and bone tissue. The deep learning model, known as U-Net++, is illustrated in Fig. 3c, where $x^{i,j}$ is the convolution block of row i and column j. Each convolution block contained two modules, and each module consisted of a convolution layer followed by a batch normalization layer followed by the ReLU activation function. The number of channels in the convolution block was 32, 64, 128, 256, and 512 from top to bottom (i = 0,1,2,3,4). The loss function for optimization was the focal loss [16], formulated as follows:

$$\text{FocalLoss} = -\alpha_t(1-p_t)^{\gamma}\log p_t, \text{ and } p_t = \begin{cases} p & y=1 \\ 1-p & otherwise \end{cases}. \quad (4)$$

In Eq. (4), p is the probability of each category after softmax, rated between 0 and 1. y = 1 represents a pixel belonging to a certain category. α_t and γ are the scale and weighted control parameters, respectively. In our experiment, α_t = 0.5 and γ = 3 were applied. These parameters affect the overall loss function, making the loss tend toward the difficult-to-classify data.

The model was designed for semantic segmentation to segment and determine the location of the lesion at the same time. As depicted in Fig. 3d, the final model output was five channels corresponding to the non-CAC area, CAC on the left main coronary artery (LM), left anterior descending (LAD), left circumflex artery (CX), and right coronary artery (RCA). The output was then passed through the softmax activation function and became the value between 0 and 1. The argmax was used to determine the channel with the largest value for each voxel (i.e., the classification result). For instance, if a pixel has the maximum value in the second channel, it belongs to a CAC at LM. The output that belonged to the calcification then passed through the process for CAC score calculation as described in Section 2.3, in which the calcium smaller than 1 mm^2 was considered the noise and removed. The U-Net with the same structure but without the dense skip connection ($x^{0,1\text{-}3}$, $x^{1,1\text{-}2}$, and $x^{2,1}$) was trained for comparison. Five metrics were used to compare the performance between the U-Net and U-Net++, as shown in Fig. 4. During model training, the batch size was set to four images at a time, the optimizer selected was Adam, the initial learning rate was set to 0.0001, and the learning rate was reduced by 5% (using a scheduler) after each epoch. The models were finally trained for 11 and 26 epochs for the U-Net and U-Net++, respectively, over approximately two days.

2.5. Statistical analysis for the standard and model-predicted CAC scores

After the model outputted the predicted CAC locations, three quantitative calcium scores were calculated for each lesion-based and vessel-specific case, as described in Section 2.3. The results were compared and presented for the overall and vessel-specific CAC scores. The intraclass correlation coefficient (ICC) was used to examine the consistency between the standard and model-predicted CAC scores. For the risk stratification category, Cohen's kappa was applied for the comparison between the standard and model-predicted CAC scores.

3. Results

3.1. Comparison for the model performance of U-Net and U-Net++

Table 3 presents the mean metrics for the U-Net and U-Net++. The CAC error in U-Net was between 5 and 7, and it was between

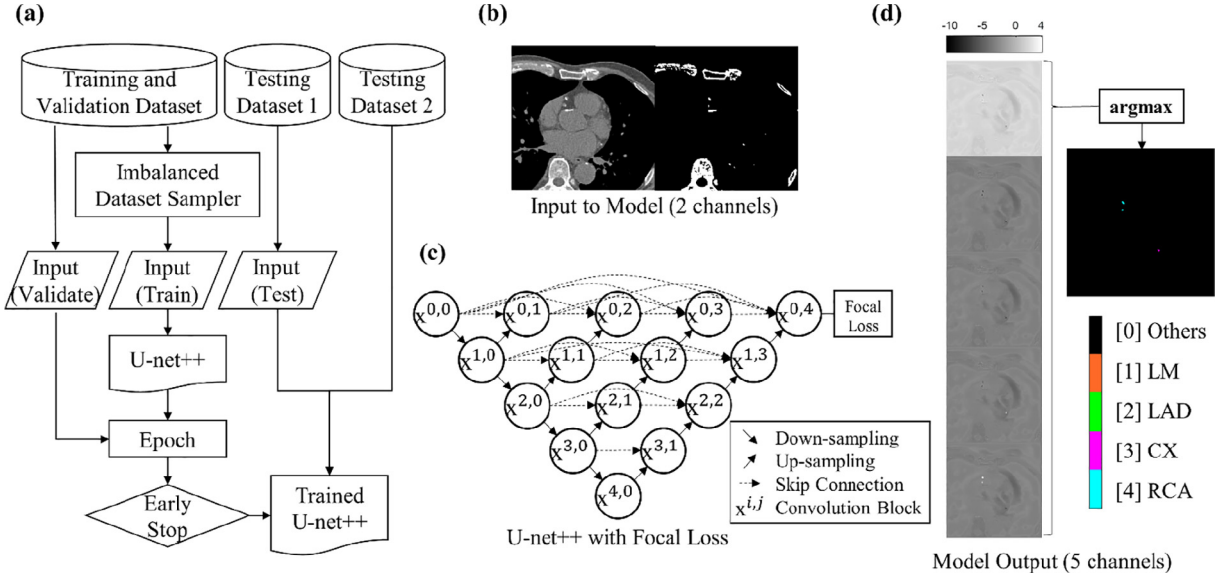
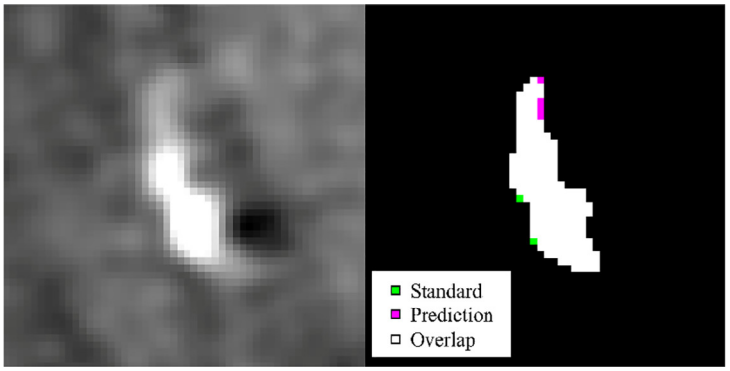


Fig. 3. Experimental flowchart and model architecture. (a) Model training workflow. (b) Schematic of the input images, including the original CT image and the thresholded binary image, to the model. (c) Schematic of U-Net++ architecture. (d) Schematic of output from the model.



True Positive (Overlap): 173 pixels
False Positive: 4 pixels
False Negative: 2 pixels

CAC Error
$= |(\text{Agatston score})_{\text{standard}} - (\text{Agatston score})_{\text{prediction}}|$
$= |86.54-87.52| = 0.98$
Precision $= TP/(TP + FP) = 0.98$
Recall $= TP/(TP + FN) = 0.99$
Dice $= 2 \times TP/[(TP + FP) + (TP + FN)] = 0.98$
IoU $= TP/(TP + FP + FN) = 0.97$

Fig. 4. Example for calculating the five metrics used to compare the U-Net with U-Net++.

Table 3
Mean metrics for the U-Net and U-Net++ in CAC detection.

	CAC Error	Precision	Recall	Dice	IoU
U-Net					
Training	6.27	0.53	0.53	0.53	0.53
Validation	7.42	0.54	0.54	0.54	0.54
Testing	5.48	0.54	0.54	0.54	0.54
U-Net++					
Training	0.12	0.91	0.92	0.91	0.90
Validation	1.57	0.87	0.86	0.85	0.83
Testing	0.48	0.88	0.87	0.86	0.84

0.1 and 1.5 in U-Net++. In the performance of the U-Net, the precision, recall, Dice, and IoU were between 0.5 and 0.6. In contrast, these metrics ranged from 0.8 to 1.0 for the performance of the U-Net++. The overall metrics indicated that the U-Net++ performed better than the U-Net.

3.2. Model performance for overall and vessel-specific calcium detection

Table 4 presents the case-based confusion matrix for calcium detection, where lesions smaller than 1 mm^2 were removed as noise, and few cases with CAC became cases without CAC after noise removal. Most of the CAC cases were correctly identified in both the Training Data Set and Testing Data Set 1. A sensitivity of 99% was achieved in both the Training Data Set and Testing Data Set 1, and specificity was 85% in the Training Data Set and 100% in Testing Data Set 1. In the Training Data Set and Testing Data Set 1, three and two cases with CAC, respectively, were determined in the standardized analysis but were not detected by the model. In the Training Data Set, five cases were identified as lacking CAC in standardized analysis but were detected as having so by the model.

Table 5 presents the confusion matrix for the model in vessel-specific calcified lesion detection. As long as the predicted voxels were included in the standard lesion (i.e., a lesion identified by standardized analysis), the lesion was considered as detected. Predicted voxels that were not included in any of the standard lesions were denoted as "not CAC." Standard lesions that were not included in any of the predicted voxels were denoted as "not detected." The results indicated that most of lesions were detected, with detection rates of 98.3%, 99.2%, 98.9%, and 95.7% in the Training Data Set and 80.0%, 96.6%, 88.7%, and 93.6% in the Testing Data Set 1 for LM, LAD, CX and RCA, respectively. The proportion of overall lesions not detected by the model was 1.9% and 5.1% in the Training Data Set and Testing Data Set 1, respectively. The additional lesions detected (not CAC) by the model was 3.9% and 5.7% in the Training Data Set and Testing Data Set 1, respectively.

3.3. Model performance on quantitative calcium scores

Table 6 shows the statistical analysis between the standardized results and model predictions for the three quantitative calcium score variables. In the Training Data Set, all ICCs approached 1 with statistical significance, both for the overall and vessel-specific scores. In Testing Data Set 1, all ICCs were greater than 0.95 with statistical significance, except for LM, where the ICC was 0.88.

Fig. 5 illustrates the overall Bland–Altman plots for the standardized analysis and model-based prediction of each score. The data were more concentrated at the zero baseline in the Training Data Set than in Testing Data Set 1. Points were generally close to the interval of the 1.96-times standard deviation in both the Training Data Set and Testing Data Set 1.

Table 4
Confusion matrix of the model in the overall calcium detection.

Predicted	Standard			
	Training Data Set		Testing Data Set 1	
	No CAC	CAC	No CAC	CAC
No CAC	29	3	10	2
CAC	5	717	0	203
Sensitivity (%)	99.58%		99.02%	
Specificity (%)	85.30%		100%	

Table 5
Confusion matrix of the model in vessel-specific calcium detection.

Predicted	Standard									
	Training Data Set					Testing Data Set 1				
	LM	LAD	CX	RCA	Not CAC	LM	LAD	CX	RCA	Not CAC
LM	290	6	0	0	10	76	3	0	0	18
LAD	0	1993	2	0	52	4	602	2	0	23
CX	0	2	1059	3	55	3	2	260	3	21
RCA	0	0	0	1721	86	0	0	11	483	25
Not Detected	5	9	10	75	–	12	16	20	30	–

Table 6
Statistical analysis of the quantitative calcium scores identified by standardized analysis versus predicted by the model.

	Agatston score				Volume Score (mm^3)				Mass Score (mg CaHA)			
Location	Standard	Predicted	ICC	*P* value	Standard	Predicted	ICC	*P* value	Standard	Predicted	ICC	*P* value
Training Data Set												
LM	0 (11)	0 (12)	1.00	<0.001*	0 (9)	0 (9)	1.00	<0.001*	0.0 (1.3)	0.0 (1.4)	1.00	<0.001*
LAD	73 (230)	73 (232)	1.00	<0.001*	43 (115)	43 (116)	1.00	<0.001*	8.0 (27.0)	8.0 (26.9)	1.00	<0.001*
CX	0 (50)	1 (50)	1.00	<0.001*	0 (32)	3 (33)	1.00	<0.001*	0.0 (5.3)	0.3 (5.4)	1.00	<0.001*
RCA	8 (97)	7 (94)	0.99	<0.001*	8 (61)	8 (58)	0.99	<0.001*	1.1 (10.3)	1.1 (9.4)	0.99	<0.001*
Overall	136 (432)	133 (434)	1.00	<0.001*	80 (227)	80 (230)	1.00	<0.001*	15.0 (50.0)	14.9 (49.9)	1.00	<0.001*
Testing Data Set 1												
LM	0 (12)	0 (9)	0.88	<0.001*	0 (9)	0 (7)	0.88	<0.001*	0.0 (1.5)	0.0 (1.1)	0.88	<0.001*
LAD	64 (224)	65 (233)	0.98	<0.001*	36 (121)	37 (121)	0.98	<0.001*	6.4 (24.3)	7.0 (25.3)	0.98	<0.001*
CX	0 (34)	0 (35)	0.96	<0.001*	0 (23)	0 (22)	0.96	<0.001*	0.0 (3.8)	0.0 (3.8)	0.96	<0.001*
RCA	5 (89)	6 (81)	0.99	<0.001*	7 (58)	7 (51)	0.99	<0.001*	0.9 (9.2)	0.9 (8.3)	0.99	<0.001*
Overall	93 (416)	95 (399)	1.00	<0.001*	53 (234)	55 (213)	1.00	<0.001*	10.2 (50.4)	10.6 (46.9)	1.00	<0.001*

"*" represents that the *p* value is<0.05. The median of each score is presented, and the value in the bracket is the interquartile range.

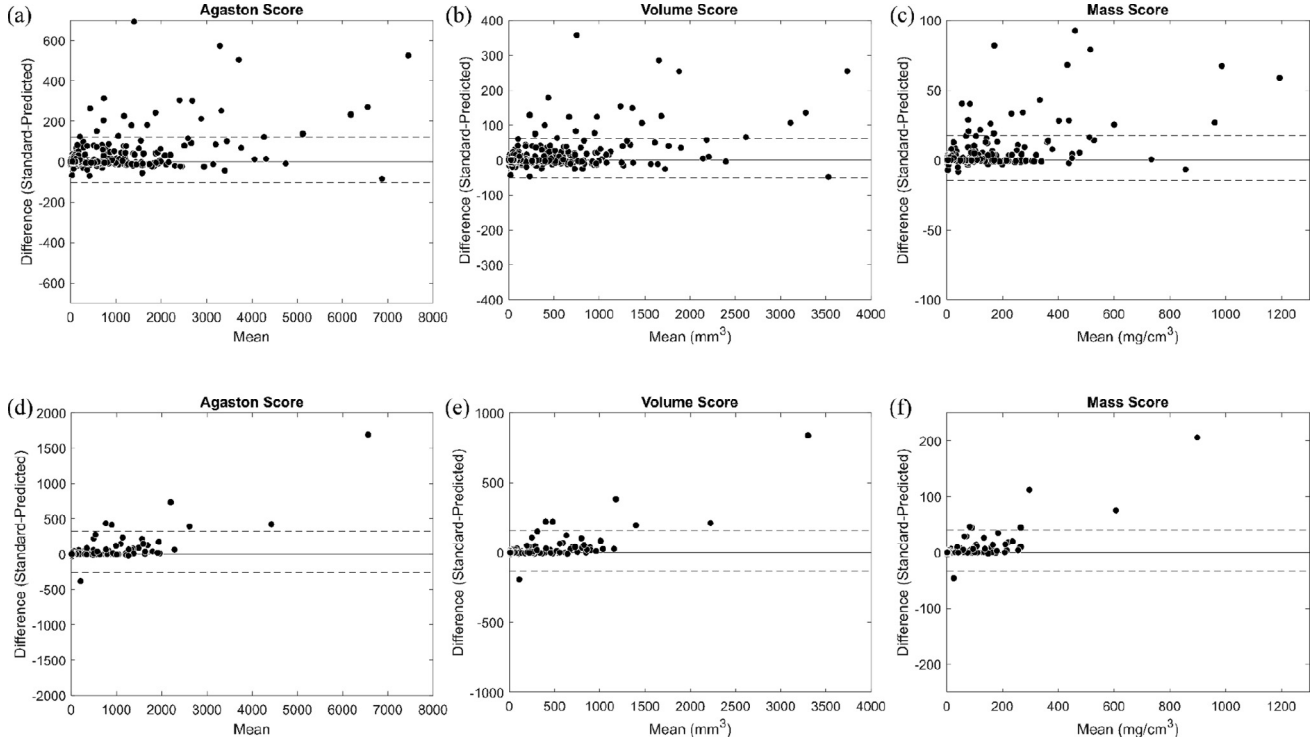


Fig. 5. Bland–Altman plots of the overall three quantitative calcification score variables in the Training Data Set [(a), (b), and (c)] and in Testing Data Set 1 [(d), (e), and (f)].

3.4. Performance of model prediction in calcium risk stratification

Table 7 presents the confusion matrix of the standard and model-predicted Agatston scores, divided into five categories for CAC risk stratification. The Cohen's kappa for the Training Data Set was 0.957 with statistical significance and a confidence interval of 0.940–0.974. The kappa for Testing Data Set 1 was 0.931 with statistical significance and a confidence interval of 0.891–0.971.

Table 7
Confusion matrix of the model performance in the overall calcium detection.

Predicted	Standard									
	Training Data Set					Testing Data Set 1				
	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5
C1	29	2	1	0	0	10	1	1	0	0
C2	3	70	2	0	0	0	20	1	0	0
C3	2	2	214	3	0	0	3	71	1	0
C4	0	0	4	214	2	0	0	2	46	2
C5	0	0	0	3	203	0	0	0	0	57

Table 8
Comparison with results of similar studies in terms of model performance on the CAC risk stratification.

	ICC	Kappa
de Vos et al. [26]	0.98	0.95
Gogin et al. [12]	0.97	0.894
Stanstedt et al. [27]	0.996	0.919
Wang et al. [28]	0.94	0.77
Zhang et al. [13]	0.988	–
Proposed Model	1.00	0.931

3.5. Model prediction performance in Testing Data Set 2

Testing Data Set 2 contained a total of 744 cases with a calcium score of zero. After parsing the images in Testing Data Set 2, the model detected that 78 of them have calcium. Examination by the experienced radiologic technologist indicated that 26 of the 78 were actual CAC cases. Two example cases for model prediction in Testing Data Set 2 are presented in Fig. 6. The model detected the suspected calcium at the CX and RCA locations, and the mask with HU values greater than 130 verified the detection by the model (Fig. 6b, e). In Fig. 6a, b, and c, an intravascular stent was mistakenly predicted as CAC; in Fig. 6d, e, and f, a confirmed CAC case was detected.

5. Discussion

The deep learning method proposed in this study was effective for detecting CAC. For the three quantitative CAC score variables, the model achieved an ICC of 1.00 with statistical significance in relation to the LM, LAD, CX, and RCA areas in the Training Data Set. In Testing Data Set 1, the ICC of all vessels was higher than 0.95, except for in LM, where an ICC of 0.88 was recorded, and all of the ICCs were statistically significant. The ICC for the overall score was 1.00 in relation to both the Training Data Set and Testing

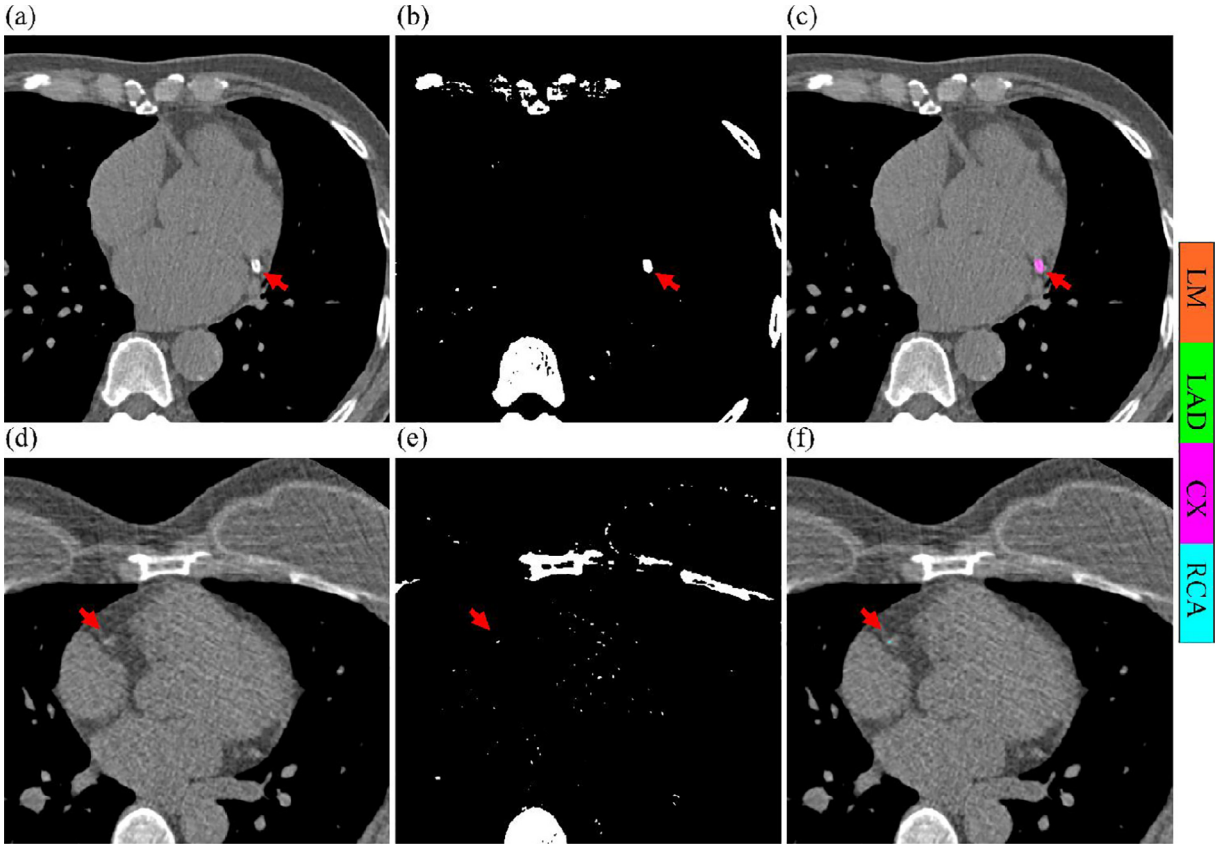


Fig. 6. Cases with calcium predicted by the model in Testing Dataset 2. The original CT images of the case (a, d), the binarized images with CT values greater than 130 HU (b, e) and the color superimposed images (c, f) of the location of the detected calcium. The red arrow is the location where the calcium was detected. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Data Set 1, indicating that the model was successful in detecting the overall CAC. In Testing Data Set 2, the model detected possible locations of CACs that were overlooked in clinical practice. For the risk stratification of CAC, the kappa values were above 0.93 with statistical significance for both the Training Data Set and Testing Data Set 1. The performance of the proposed model was highly consistent with that of the semiautomated standard CAC analysis used in clinical practice. In a clinical scenario, the physician can use the proposed model to predict and analyze the CAC and then confirm the results based on the model predictions. This has the following advantages. First, compared to the time-consuming semiautomated analysis by the physician, the model predicts a case in 50 ms. Second, the deep learning model is not affected by factors that can happen to humans, such as fatigue. Finally, the model does not have the problem of intraobserver and interobserver variability. The proposed model can be used as a clinical aid for automated CAC detection to reduce the clinical manpower burden and thus improve the quality of CAC diagnosis.

In the three cases where the model did not detect CAC in the Training Data Set, CAC was actually detected in one of them but was removed in the noise removal step (because the segmentation area was smaller than 1 mm^2). In the case where the model detected CAC in the Training Data Set but the standard semiautomated analysis did not, the model did find the correct CAC location, but the calcium area was removed as noise because the area was<1 mm^2. The two cases not detected in Testing Data Set 1 were CAC lesions located at LM. In Testing Data Set 2, we assessed whether the proposed model could detect CAC cases that were missed in standard clinical practice, discovering that the proposed model did detect such cases. Furthermore, the model performed well in the Training Data Set regarding the vessel-specific level. In Testing Data Set 1, compared with its performance on classifying LAD and RCA, the model exhibited poorer prediction outcomes in the LM and CX regions because of the classification imbalance problem caused by the relatively small number of LM and CX cases.

Clinical semiautomated CAC analysis is not only time-consuming but also observer dependent. By contrast, automated CAC detection is fast and objective. Studies on automated CAC detection have used traditional image processing, machine learning–based methods [22–25], and, more recently, deep learning–based methods [12,13,15,26–28]. Traditional image processing methods essentially use mapping and rule-based approaches to define possible lesions using information such as calcium patterns and their characteristics and to then identify candidate lesions using classifiers. In recent years, deep learning methods have been rapidly developed, especially the models based on the convolution neural network, which is particularly effective in image object detection. The advantage of deep learning is that the computation speed is ultra-fast, and that the model can learn the image processing task efficiently provided the data set is sufficiently large. In our proposed model, the speed of CAC detection from a CT image was approximately 50 ms. The common guideline for obtaining CAC scores involves CAC risk classification based on a contrast-free cardiac CT scan [29]; thus, we compared our model's prediction results with those of similar studies in terms of the ICC and kappa coefficients between clinical semiautomated and fully automated CAC scores for the risk stratification (Table 8). In similar studies that used deep learning to model the automated detection of calcium, the final kappa values ranged from 0.77 to 0.95 in their Testing Data Sets, and the final ICCs of overall Agatston scores ranged from 0.94 to 0.988 in their Testing Data Sets. In the Testing Data Set 1, our model was able to achieve a kappa value of 0.931 for the CAC risk stratification and an ICC of 1.0 for the overall Agatston score. This only indicates that our model achieved comparable performance to other studies in detecting CAC because we did not use our data to build the models from these studies for comparison.

Wang et al. and Gogin et al. have used a 3D approach for model building, whereas Zhang et al. used three projections of two-dimensional (2D) images (coronal, sagittal, and axial) for model building [12,13,28]. In addition to CAC detection, Zhang et al. added a regression model to calculate quantitative calcium scores [13]. de Vos et al. used a two-stage model, with the first stage aligning the 3D cardiac structure and the second stage detecting CAC [26]. In our study, we used only the axial 2D images as the input. Besides, Zhao et al. added empirical definition features to the model input, such as a 130-HU threshold for defining the suspected CAC, in addition to the original CT images [15]. They compared different deep learning models and discovered that a model with the empirical feature image as input outperformed a model without it. We followed their approach in our model input by including binarized images (with 130-HU threshold) as the second channel of input. Additionally, because the proportion of the CAC voxels was several orders less than that of the non-CAC voxels, we introduced the imbalance data sampler to balance the proportion of 2D images with and without CAC [21], and then applied focal loss to further increase the ability of the model to learn difficult-to-classify voxels [16]. Finally, we used an improved U-Net architecture, namely U-Net++ [14]. In U-Net, images are downsampled four times before being upsampled. During each downsampling, image dimensions is reduced, which results in the loss of the fine feature information. Therefore, how many levels of downsampling should be selected for optimal model performance is unclear. Unlike traditional U-Net framework, U-Net++ implements upsampling after each downsampling, and this makes full use of the information obtained from each downsampling. Thus, although the U-Net++ model involves more parameters, it outperforms U-Net in many medical image segmentation tasks [14]. Our results demonstrated that the U-Net++ outperformed the U-Net architecture in the task for semantic segmentation of CAC.

CAC scoring is reliable, and its reproducibility has been extensively studied across different CT imaging modalities [18,30–33]. Numerous studies have found that the CAC score is meaningful and recommended for consideration in the risk stratification for coronary heart disease and atherosclerotic cardiovascular disease [6–9,20,34–37]. In addition, the presence or absence of CAC is a powerful indicator and can be used as a negative risk factor. Studies have shown that even in populations classified as at high risk (e.g., with high levels of hs-CRP and high-risk factors such as dyslipidemia and smoking), those with the zero CAC score have a lower risk of atherosclerotic cardiovascular disease [34]. According to the updated guideline [35], these patients could be reclassified to the low-risk group if CAC is not detected, and preventive interventions (e.g., statins) could be reasonably postponed. Conversely, if a CAC is detected, it should be noticed. The results of our experiment enabled physicians to review cases in which CAC was not diagnosed but was detected by the proposed model. Besides, in current clinical settings, risk stratification of CAC to cardiac disease is undertaken based on the overall CAC score, without considering the distribution of calcium in different vessels. Calcified lesions may be either concentrated in a single vessel or dispersed across many vessels for cases with the same CAC score. As such, identical overall CAC scores could result in completely different levels of coronary heart disease risk [8,10]. Vessel-specific risk stratification is superior to risk stratification by the overall CAC score in predicting coronary heart disease [36]. In addition, studies have shown that patients with LM calcium without symptoms have a higher risk of total cardiovascular mortality [37]. Therefore, analyzing vessel-specific CAC scores is crucial. The model proposed in this study is in excellent agreement with the standard semiautomated CAC analysis, both in detecting the overall CAC and vessel-specific CAC. Our proposed model can thus provide a comprehensive assessment of calcium for the totality of and across specific vessels.

This study demonstrated some limitations. First, in our retrospectively collected data set, clinical CAC analyses were not performed by the one individual. Therefore, we could not ensure that the criteria for assessing CAC were absolutely identical across all cases. In other words, the model learned from the subjective perceptions of different individuals to become more objective. Second, we did not exclude the cases that had intravascular stents in our experiment. The stents were not included in the clinical CAC evaluation. The HU values of stents are similar to those of calcium and are located in the coronary artery. In our experiment, the model sporadically misidentified stents as calcium. Third, because an image can contain different vessel types, we only considered the image sampling balance with and without CAC and did not balance the sampling for different vessel types. This is why our model performed less effectively in the LM and CX regions than in LAD and RCA. Finally, our model was constructed based on images scanned using the same machine and was not evaluated for images obtained from other scanning machines. Images scanned using different machines can vary. Therefore, if the proposed model is to be applied to the scanned images from other machines, either retraining the model or using transfer learning [38] to speed up the model training and enhance its performance is recommended.

6. Conclusion

In this study, a fully automated CAC detection model was developed using U-Net++ in conjunction with focal loss to effectively detect the location of calcium in different coronary arteries. Our model could detect mild CAC cases that are not identified in clinical practice. The proposed fully automated CAC detection model is highly consistent with standard clinical semiautomated CAC analysis, both in terms of overall and vessel-specific quantitative CAC score determination. The proposed model can effectively assist CAC analysis in clinical settings, thereby facilitating overall CAC risk assessment.

CRediT authorship contribution statement

Jia-Sheng Hong: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Yun-Hsuan Tzeng:** Conceptualization, Resources, Data curation, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Wei-Hsian Yin:** Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Kuan-Ting Wu:** Validation, Formal analysis. **Huan-Yu Hsu:** Validation. **Chia-Feng Lu:** Resources, Writing – review & editing, Supervision, Project administration. **Ho-Ren Liu:** Conceptualization, Resources, Data curation, Writing – review & editing, Visulization, Supervision, Project administration, Funding acquisition. **Yu-Te Wu:** Methodology, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors gratefully acknowledge partial financial support by the National Yang Ming Chiao Tung University, Cheng Hsin General Hospital, Veteran General Hospitals University System of Taiwan, the Brain Research Center, and National Yang Ming Chiao Tung University from the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan under grant no. CY11016, CY11102, and VGHUST110-G7-2-1. This manuscript was edited by Wallace Academic Editing.

References

[1] Sahin B, Ilgun G. Risk factors of deaths related to cardiovascular diseases in World Health Organization (WHO) member countries. Health Soc Care Community 2020. https://doi.org/10.1111/hsc.13156.
[2] Cardiovascular diseases (CVDs) n.d. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed February 9, 2022).
[3] Agatston AS, Janowitz WR, Hildner FJ, Zusmer NR, Viamonte M, Detrano R. Quantification of coronary artery calcium using ultrafast computed tomography. J Am Coll Cardiol 1990;15:827–32.
[4] Rumberger JA, Simons DB, Fitzpatrick LA, Sheedy PF, Schwartz RS. Coronary artery calcium area by electron-beam computed tomography and coronary atherosclerotic plaque area. A Histopathol Correlative Study Circulation 1995;92:2157–62. https://doi.org/10.1161/01.cir.92.8.2157.
[5] Budoff MJ, Diamond GA, Raggi P, Arad Y, Guerci AD, Callister TQ, et al. Continuous probabilistic prediction of angiographically significant coronary artery disease using electron beam tomography. Circulation 2002;105:1791–6. https://doi.org/10.1161/01.cir.0000014483.43921.8c.
[6] LaMonte MJ, FitzGerald SJ, Church TS, Barlow CE, Radford NB, Levine BD, et al. Coronary artery calcium score and coronary heart disease events in a large cohort of asymptomatic men and women. Am J Epidemiol 2005;162:421–9. https://doi.org/10.1093/aje/kwi228.
[7] Budoff MJ, Shaw LJ, Liu ST, Weinstein SR, Mosler TP, Tseng PH, et al. Long-term prognosis associated with coronary calcification: observations from a registry of 25,253 patients. J Am Coll Cardiol 2007;49:1860–70. https://doi.org/10.1016/j.jacc.2006.10.079.
[8] Blaha MJ, Mortensen MB, Kianoush S, Tota-Maharaj R, Cainzos-Achirica M. Coronary artery calcium scoring: is it time for a change in methodology? JACC Cardiovasc Imaging 2017;10:923–37. https://doi.org/10.1016/j.jcmg.2017.05.007.
[9] Nakao YM, Miyamoto Y, Higashi M, Noguchi T, Ohishi M, Kubota I, et al. Sex differences in impact of coronary artery calcification to predict coronary artery disease. Heart 2018;104:1118–24. https://doi.org/10.1136/heartjnl-2017-312151.
[10] Dzaye O, Dudum R, Mirbolouk M, Orimoloye OA, Osei AD, Dardari ZA, et al. Validation of the Coronary artery calcium data and reporting system (CAC-DRS): dual importance of CAC score and CAC distribution from the coronary artery calcium (CAC) consortium. J Cardiovasc Comput Tomogr 2020;14:12–7. https://doi.org/10.1016/j.jcct.2019.03.011.
[11] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation, Springer; 2015, p. 234–41.
[12] Gogin N, Viti M, Nicodème L, Ohana M, Talbot H, Gencer U, et al. Automatic coronary artery calcium scoring from unenhanced-ECG-gated CT using deep learning. Diagn Interv Imaging 2021.
[13] Zhang N, Yang G, Zhang W, Wang W, Zhou Z, Zhang H, et al. Fully automatic framework for comprehensive coronary artery calcium scores analysis on non-contrast cardiac-gated CT scan: total and vessel-specific quantifications. Eur J Radiol 2021;134:. https://doi.org/10.1016/j.ejrad.2020.109420109420.
[14] Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. Unet++: A nested u-net architecture for medical image segmentation. Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support, Springer; 2018, p. 3–11.
[15] Zhao M, Che X, Liu H, Liu Q. Medical prior knowledge guided automatic detection of coronary arteries calcified plaque with cardiac CT. Electronics 2020;9:2122.
[16] Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 2020;42:318–27. https://doi.org/10.1109/TPAMI.2018.2858826.
[17] Callister TQ, Cooil B, Raya SP, Lippolis NJ, Russo DJ, Raggi P. Coronary artery disease: improved reproducibility of calcium scoring with an electron-beam CT volumetric method. Radiology 1998;208:807–14. https://doi.org/10.1148/radiology.208.3.9722864.
[18] Hong C, Bae KT, Pilgram TK. Coronary artery calcium: accuracy and reproducibility of measurements with multi-detector row CT–assessment of effects of different thresholds and quantification methods. Radiology 2003;227:795–801. https://doi.org/10.1148/radiol.2273020369.
[19] Rumberger JA, Brundage BH, Rader DJ, Kondos G. Electron beam computed tomographic coronary calcium scanning: a review and guidelines for use in asymptomatic persons. Mayo Clin Proc 1999;74:243–52. https://doi.org/10.4065/74.3.243.
[20] Arjmand SA. Coronary artery calcium score: a review. Iran Red Crescent Med J 2013;15:. https://doi.org/10.5812/ircmj.16616e16616.
[21] Ming. Imbalanced Dataset Sampler. 2021. Retrieved from https://github.com/ufoym/imbalanced-dataset-sampler.
[22] Isgum I, Rutten A, Prokop M, van Ginneken B. Detection of coronary calcifications from computed tomography scans for automated risk assessment of coronary artery disease. Med Phys 2007;34:1450–61. https://doi.org/10.1118/1.2710548.
[23] 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;26:829–38. https://doi.org/10.1007/s10554-010-9608-1.
[24] Shahzad R, van Walsum T, Schaap M, Rossi A, Klein S, Weustink AC, et al. Vessel specific coronary artery calcium scoring: an automatic system. Acad Radiol 2013;20:1–9. https://doi.org/10.1016/j.acra.2012.07.018.
[25] Wolterink JM, Leiner T, Takx RA, Viergever MA, Išgum I. An automatic machine learning system for coronary calcium scoring in clinical non-contrast enhanced, ECG-triggered cardiac CT. vol. 9035, International Society for Optics and Photonics; 2014, p. 90350E.
[26] de Vos BD, Wolterink JM, Leiner T, de Jong PA, Lessmann N, Isgum I. Direct automatic coronary calcium scoring in cardiac and chest CT. IEEE Trans Med Imaging 2019;38:2127–38. https://doi.org/10.1109/TMI.2019.2899534.
[27] Sandstedt M, Henriksson L, Janzon M, Nyberg G, Engvall J, De Geer J, et al. Evaluation of an AI-based, automatic coronary artery calcium scoring software. Eur Radiol 2020;30:1671–8. https://doi.org/10.1007/s00330-019-06489-x.
[28] Wang W, Wang H, Chen Q, Zhou Z, Wang R, Zhang N, et al. Coronary artery calcium score quantification using a deep-learning algorithm. Clin Radiol 2020;75:237. e11-237. e16.
[29] Greenland P, Blaha MJ, Budoff MJ, Erbel R, Watson KE. Coronary calcium score and cardiovascular risk. J Am Coll Cardiol 2018;72:434–47. https://doi.org/10.1016/j.jacc.2018.05.027.
[30] Mautner GC, Mautner SL, Froehlich J, Feuerstein IM, Proschan MA, Roberts WC, et al. Coronary artery calcification: assessment with electron beam CT and histomorphometric correlation. Radiology 1994;192:619–23.
[31] Shields JP, Mielke Jr CH, Watson P. Inter-rater reliability of electron beam computed tomography to detect coronary artery calcification. Am J Card Imaging 1996;10:91–6.
[32] Budoff MJ, McClelland RL, Chung H, Wong ND, Carr' JJ, Gray MM, et al. Reproducibility of coronary artery calcified plaque with cardiac 64-MDCT: the multi-ethnic study of atherosclerosis. Am J Roentgenol 2009;192:613–7.
[33] Mao SS, Pal RS, McKay CR, Gao YG, Gopal A, Ahmadi N, et al. Comparison of coronary artery calcium scores between electron beam computed tomography and 64-multidetector computed tomographic scanner. J Comput Assist Tomogr 2009;33:175–8.
[34] Nasir K, Cainzos-Achirica M. Role of coronary artery calcium score in the primary prevention of cardiovascular disease. BMJ 2021;373:. https://doi.org/10.1136/bmj.n776n776.
[35] Arnett DK, Blumenthal RS, Albert MA, Buroker AB, Goldberger ZD, Hahn EJ, et al. 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: a report of the american college of cardiology/american heart association task force on clinical practice guidelines. Circulation 2019;140: e596–646. https://doi.org/10.1161/CIR.0000000000000678.
[36] Qian Z, Anderson H, Marvasty I, Akram K, Vazquez G, Rinehart S, et al. Lesion- and vessel-specific coronary artery calcium scores are superior to whole-heart Agatston and volume scores in the diagnosis of obstructive coronary artery disease. J Cardiovasc Comput Tomogr 2010;4:391–9. https://doi.org/10.1016/j.jcct.2010.09.001.
[37] Lahti SJ, Feldman DI, Dardari Z, Mirbolouk M, Orimoloye OA, Osei AD, et al. The association between left main coronary artery calcium and cardiovascular-specific and total mortality: the coronary artery calcium consortium. Atherosclerosis 2019;286:172–8.
[38] Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C. A survey on deep transfer learning, Springer; 2018, p. 270–9.