Margin requirements for first page
Paper size this page US Letter
72 pt
1 in
25.4 mm

FULLY-CONVOLUTIONAL DEEP-LEARNING BASED SYSTEM FOR CORONARY CALCIUM SCORE PREDICTION FROM NON-CONTRAST CHEST CT

Ran Shadmi, Victoria Mazo, Orna Bregman-Amitai, Eldad Elnekave

Zebra Medical Vision, Shefayim, Israel

ABSTRACT

The amount of calcium deposits in the coronary arteries is an important biomarker of cardiovascular disease. Coronary calcium has traditionally been quantified as an Agatston score using ECG-synchronized **cardiac CT**. Coronary calcium is rarely quantified from general **chest CT** scans, of which nearly 10 Million are performed in the US annually. We present an automatic method based on fully-convolutional deep neural network to segment coronary calcium and predict Agatston score from any non-contrast chest CT.

We experimented with an internal dataset acquired through partnership with a large health organization in Israel. The dataset is composed of 1054 Chest CTs and reflects a variety of originating institutions, acquisition devices and manufacturers. In comparison to expert manual annotations, our algorithm achieved a Pearson correlation coefficient of 0.98. Bland-Altman analysis demonstrated a bias of 0.4 with 95% limits of agreement of [-189.9–190.7]). Our linearly weighted Kappa results are 0.89 for Agatston risk category assignment.

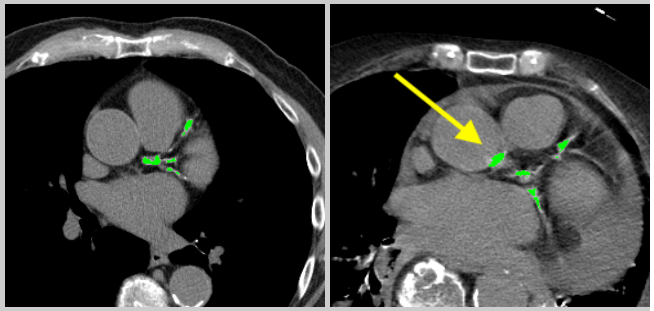
We also applied our method on a very large (14,365 subjects) cohort from the National Lung Screening Trial (NLST). We demonstrate correlation of the algorithm predictions with cardiovascular-related clinical outcomes.

Index Terms— coronary calcium, deep learning, Agatston score, segmentation, chest CT, computer aided diagnosis

1. INTRODUCTION

Cardiovascular disease (CVD) is the leading cause of death in the United States [1]. The amount of coronary artery calcification (CAC) is a powerful predictor of cardiovascular events and mortality [2]. In clinical practice, CAC is identified and quantified in terms of the Agatston score [3] using a dedicated ECG-synchronized cardiac Calcium Scoring CT (CSCT), followed by a human operator manually identifying CAC lesions. In recent years it has been shown that the much more widely performed non-ECG-synchronized chest CT (Chest CT) provides similar diagnostic information across categories of CAC scores [4].

Several works in the last decade have attempted to provide automatic methods to estimate Agatston score from CT scans, mostly CSCT. Išgum et al. [5] presented a comprehensive work for labeling data and training combination of KNN and SVM classifiers to automatically identify CAC in Chest CT, using both local image features and location information (based on *apriory* probability map). Shahzad et al. [6] detected CAC in CSCT scans using a KNN classifier based on local image descriptors and location features based on CCTA atlases. Wolterink et al. [7] extended Išgum's work [5] for CSCT, where they used randomized decision trees for vessel-specific CAC classifiers and included human-in-the-loop for ambiguous detections.



(a) (b)

Fig. 1: Examples of segmentation results (better viewed in color): (a) Accurate segmentation of calcifications in the LM coronary artery, LCX and LAD, (b) additional false-positive in the aorta wall (marked with a yellow arrow)

Advances in deep learning have enabled the extension of convolutional neural networks (CNN) to the task of segmentation of arbitrarily-shaped objects in images [8, 9, 10], including applications in the medical domain [11, 12]. Lessmann et al. [13] presented patch-based CNN segmentation approach for the identification of CAC lesions in Chest CTs, where each voxel is represented by three centered orthogonal slices and classified using three concurrent CNNs. Santini et al. [14] has recently presented another patch-based deep-learning approach for CCS in CSCT scans.

Unlike patch-based segmentation, fully convolutional neural networks (FCNNs) add upsampling layers to standard CNNs to recover the spatial resolution of the input at the output layer. In order to compensate for the resolution loss

54 pt
0.75 in
19.1 mm
54 pt
0.75 in
19.1 mm
54 pt
0.75 in
19.1 mm

Margin requirements for the other pages
Paper size this page US Letter
54 pt
0.75 in
19.1 mm

Risk Category		Train	Validation	Test
0	(Zero)	15%	29%	32%
1-10	(Minimal)	3%	4%	10%
11-100	(Mild)	10%	17%	19%
101-400	(Moderate)	21%	19%	17%
>400	(Severe)	51%	31%	21%

Table 1: Agatston risk category distribution for each dataset

induced by pooling layers, FCNNs introduce skip connections between their downsampling and upsampling paths[11]. The fully-convolutional DenseNet (FC-DenseNet)[15] introduces additional connections between each layer within a block, encouraging re-use of features from previous layers and better convergence during training.

The presented work contributions are threefold: (1) It is the first (to our knowledge) application of FCNN-based system to measure the CAC score from general-indication non-contrast Chest CT; (2) Our method was rigorously validated in accordance with FDA guidelines; and (3) We demonstrate how our method can assist in early detection of people with high risk of CVD.

2. DATASETS

2.1. Internal

Through partnership with a large health provider in Israel, we collected 848 adult non-contrast Chest CT scans with high prevalence of significant amount of CAC, and used it for training (512 samples) and validation (336). For testing, we selected an additional set of 203 general indication scans reflecting the distribution of age and gender in the general population. We excluded scans with metallic artifacts or signs of deforming pathology such as cardiac surgery. Table 1 presents Agatston score distribution in the datasets.

For the train and validation sets, candidate CAC lesions were manually annotated by radiologists using internally developed interactive application. Only voxels with values above the clinical standard threshold for Coronary Calcium Scoring (CCS) of 130 Hounsfield Units (HU) [3] were considered. Additionally, 3 individual experts annotated the test-set data using a standard CCS module on a commercial radiology workstation (Kodak Carestream PACS). Their Agatston scores were averaged to establish ground-truth for reference.

2.2. National Lung Screening Trial (NLST)

The National Lung Screening Trial [16] was a large-scale clinical trial aiming to compare the efficacy of CT screening and standard chest X-ray as methods of lung cancer screening, with a 6.5 years follow-up. Data from 14,365 patients was made available to us by the NIH for this analysis, of whom 452 (3.15%) had CVD related death in the follow-up, 3468 (24.14%) had reported CVD history, 1,838 (12.79%) were flagged as having significant CAC by the clinician during the study, and 8,607 (59.92%) had no documented CVD.

3. METHODS

3.1. Preprocessing

Firstly we apply a sequence of thresholding, connected components analysis and morphological operations to detect the lungs, trachea and the carina (trachea bifurcation point). Based on the location of the lungs and the carina, we apply a set of heuristics to detect a bounding box around the heart. A soft tissue window (level 40HU, width 350HU), which is routinely utilized for detecting CAC, is applied to the resulting volume. To provide 3D context, the volume is divided into overlapping sequences of 3 consecutive slices which are later fed into the segmentation network to provide pixel-wise prediction for the middle slice in the sequence.

3.2. Fully convolutional neural network for segmentation

The main idea underlying FCNNs for segmentation is extending a regular CNN in which a sequence of pooling operators progressively reduces the spatial size of the network, by adding successive layers where pooling operators are replaced by upsampling operators. The two paths are commonly called ”contracting” and ”expanding”, respectively. Both in U-net [11] and in the DenseNet [15] variation of FCNN, which we explore in this paper, high resolution features from the contracting path are combined with the upsampled output via the so-called ”skip” connections (see Fig. 2). This enables propagation of fine-grain localization together with context information. A successive convolution layer can learn to combine both low and high resolution information.

In this paper we consider both U-Net [11] and FC-DenseNet [15] architectures. The U-Net consists of $2n+1$ blocks of 3x3 convolution-dropout-batchnorm-ReLU sequences. Every time the spatial dimensions are halved, the number of feature channels doubles. Between the downsampling and upsampling paths there is a bottleneck block, in which the image spatial resolution reaches its smallest spatial dimension. Every step in the upsampling path consists of a transposed convolution which doubles the feature map spatial dimensions, a concatenation with the correspondingly feature map from the contracting path and a U-net block. At the end of the expanding path, a 1x1 convolution is applied followed by a softmax, which results in a probability map of the same spatial dimensions as the input image, and a 3rd dimension equaling the number of classes in the segmentation task.

Fully Convolutional DenseNet is similar in its block structure to U-Net but has a higher connectivity within each block, called dense block [17] and additional skip connections in the contracting path (shown on Fig. 2). There is also a

54 pt
0.75 in
19.1 mm
54 pt
0.75 in
19.1 mm
54 pt
0.75 in
19.1 mm

Margin requirements for the other pages
Paper size this page US Letter
54 pt
0.75 in
19.1 mm

1x1 convolution-dropout-batchnorm-ReLU block before each max pooling, leading to additional *n+1* convolutional layers compared to U-Net with the same block structure. Within a dense block, each layer is connected to all the following layers via concatenation of their activation maps. The number of filters in all the convolutional layers in a dense block is the same and it is denoted by g (growth rate). The number of filters in all the 1x1 convolutional layers (except the last one), equals the number of input channels.

3.3. Post-processing

After feeding all the cropped axial slices through the network, a "prediction" volume is assembled where each voxel's intensity represents its probability of being a CAC voxel. Then we identify 2D candidate blobs on each axial slice by thresholding with 130 HU and performing connected-components analysis. We characterize each blob by θ, its 95% percentile probability value. Each blob is classified as CAC if its θ is greater than a predefined threshold. Calculation of the final Agatston score for the whole volume is done following the clinically accepted protocol described in [3].

To determine the optimal threshold, we used the validation set to exhaustively search the best threshold in terms of the smallest standard deviation of the differences between the predicted Agatston scores and the references, while limiting the search to small bias values (less than 3).

54 pt
0.75 in
19.1 mm

4. EXPERIMENTS AND RESULTS

We trained both FC-DenseNet and U-net. FC-DenseNet was composed of 5 blocks on each path with 4 layers in each block and growth rate g=12, a total of 56 convolutional layers. U-Net design followed closely that of [11]; 23 layers with 4 blocks on each path. Both architectures were trained using weighted cross-entropy loss applied on the prediction map and L2 weight regularization of 1e-4, for 40,000 iterations (roughly 13 epochs) and optimized using RMSProp. We used batches of 12 randomly cropped images (up to 75% of the original size). We started with a learning-rate of 1e-4 which exponentially decayed with rate of 0.5 every 8,000 iterations. Both networks were implemented using TensorFlow and were trained for 12-14 hours on 3 NVIDIA K80s.

Model selection was based on the best performing checkpoint in terms of Agatston score prediction with regard to the validation dataset. The same model and parameters were used for all of the experiments.

4.1. Internal Dataset

All results reported in this section are based on our test dataset, which was neither used during training nor validation. Unless mentioned otherwise, all the reported results were achieved using our FC-DenseNet implementation. We only report results with regard to the Agatston score, since it is the most clinically relevant measure. We present a scatter plot showing correlation between the 203 reference and predicted scores in Fig. 3, and a Bland-Altman analysis plot in Fig. 4. Fig. 1 shows examples of segmentation results.

In terms of the Pearson correlation coefficient, we achieve r=0.98 (p <0.0001) between the reference and predicted Agatson score. Bland-Altman analysis shows a very small bias of 0.4 Agatston units with 95% limits of agreement of [-189.9–190.7]. Only 3.9% of the samples lie outside the region of agreement. For comparison, the U-Net architecture achieves Pearson 0.97 but introduces bias of -14.5 with limits of agreement of [-252.0–222.9].

We present a risk category agreement in Table 2. The linearly weighted Kappa score is 0.89, with 84.7% of the scans placed in the correct category. Another 11.3% are placed only one category away, where almost 55% of the mistakes are confusing the first (zero) and second (1-10) categories, which are both difficult to differentiate and highly sensitive to small prediction mistakes. The U-Net architecture achieves a Kappa of 0.86 with only 81.8% of the scans placed in the correct category.

		Reference					
		I	II	III	IV	V	Total
Predicted	I	**72**	17	6	0	0	95
	II	0	**3**	2	0	0	5
	III	1	0	**27**	1	1	30
	IV	0	0	1	**34**	1	36
	V	0	0	0	1	**36**	37
	Total	73	20	36	36	38	203

Table 2: Agreement in cardiovascular risk categorization (I: 0, II: 1-10, III: 11-100, IV: 101-400 and V: >400) based on Agatston score predicted by the algorithm vs. the reference.

4.2. National Lung Screening Trial (NLST) Dataset

We applied the algorithm on 14,365 subjects selected from the National Lung Screening Trial (NLST), for which a median of 6.5 years follow-up was available. Some relevant insights are provided below, as the full analysis is the focus of a separate clinical publication.

1. 1% of the patients with a predicted CCS of zero died due to CVD during the follow-up. 4% of those with scores above 400 and 6% of those with scores above 1000 died in the same interval. We see a correlation of the predicted CCS with the risk of cardiovascular-related death.
2. There's an exponential correlation of CVD with increasing Agatston score: for risk categories I-III (0-100), 13% reported having heart attack or stroke in the past. That doubles (22%) for category IV (101-400), and doubles again (45%) for the highest category (>400).

54 pt
0.75 in
19.1 mm

Margin requirements for the other pages
Paper size this page US Letter
54 pt
0.75 in
19.1 mm

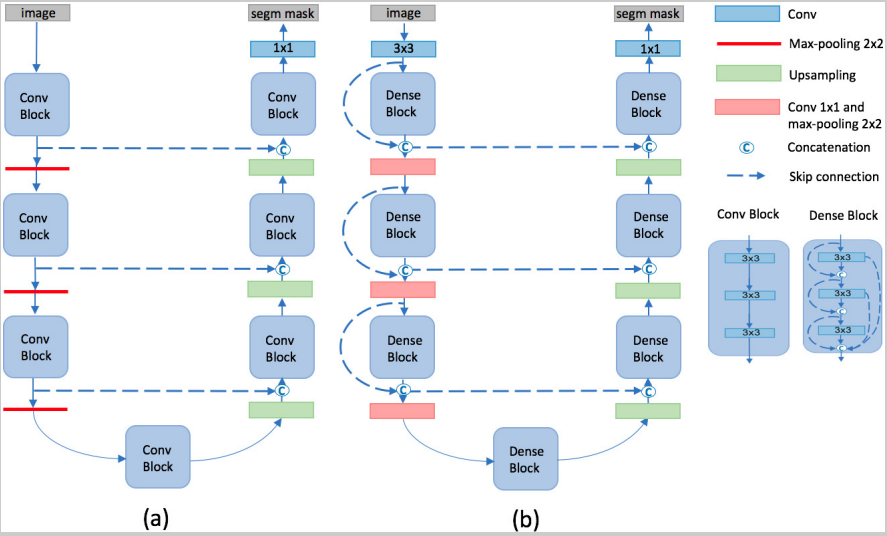


Fig. 2: (a) Generic architecture of U-Net, (b) Generic architecture of a Fully Convolutional DenseNet. Note that for the sake of brevity, the illustration shows fewer blocks than our actual implementation.

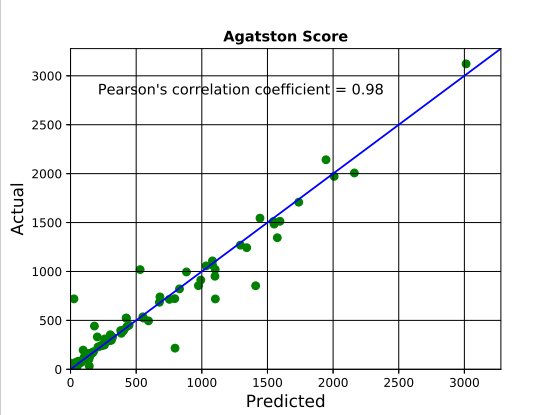


Fig. 3: Reference Agatston score vs. predicted Agatston score

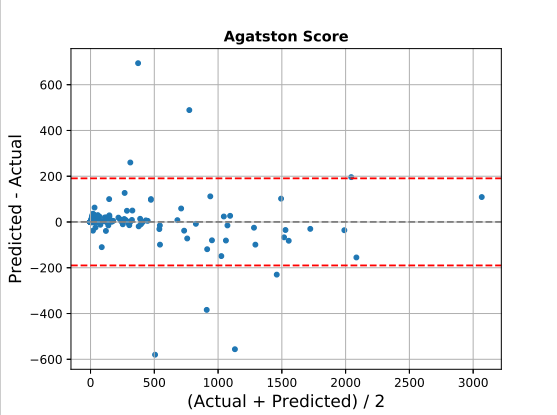


Fig. 4: Bland-Altman plot

3. Radiologists under-report the CAC burden: among patients with scores over 400, 32% were not reported as having significant CAC. For people with scores 100-400, 57% were not reported. This suggests the potential impact of automatic CCS quantification on regular Chest CTs.

5. SUMMARY AND CONCLUSIONS

We present an automated system to estimate the Agatston calcium score in general-indication non-contrast Chest CT scans that yields high performance in comparison to a committee of experts. We evaluated two network architectures: the popular U-Net [11] and the more recent DenseNet [15] which proved to be superior in terms of Agatston score prediction.

When applied to the CT data from 14,365 NLST subjects, the results of our method demonstrated strong correlation to the cardiovascular outcomes.

The orCaScore challenge [18] offers a public dataset to evaluate algorithms for CCS. However, the challenge data is cardiac CT while we address the more common Chest CT, therefore the challenge is currently not applicable to us.

The main sources of confusion for our algorithm were (a) false-positive calcified mitral valve and (b) confusion between calcifications in the aorta and in the coronary arteries very close to the aortic wall. We consider extending our tagging effort in the future to account for these cases and add them as separate classes during training.

Additional future work may include experimenting with different network architectures, using larger 3D context for classification and adding other loss function, for example direct L2 minimization of the CCS prediction error.

54 pt
0.75 in
19.1 mm
54 pt
0.75 in
19.1 mm
54 pt
0.75 in
19.1 mm

Margin requirements for the other pages
Paper size this page US Letter
54 pt
0.75 in
19.1 mm

6. REFERENCES

[1] "Health, united states, 2016: With chartbook on long-term trends in health," 2017.

[2] R Detrano, AD Guerci, JJ Carr, DE Bild, G Burke, AR Folsom, K Liu, S Shea, M Szklo, DA Bluemke, DH O'Leary, R Tracy, K Watson, ND Wong, and RA Kronmal, "Coronary calcium as a predictor of coronary events in four racial or ethnic groups," *New England Journal of Medicine*, 2008.

[3] AS Agatston, WR Janowitz, FJ Hildner, NR Zusmer, M Viamonte, and R Detrano, "Quantification of coronary artery calcium using ultrafast computed tomography," *American College of Cardiology*, 1990.

[4] X Xie, Y Zhao, GH de Bock, PA de Jong, WP Mali, M Oudkerk, and R Vliegenthart, "Validation and prognosis of coronary artery calcium scoring in non-triggered thoracic computed tomography: Systematic review and meta-analysis," *Circulation: Cardiovascular Imaging*, 2013.

[5] I Išgum, M Prokop, M Niemeijer, MA Viergever, and B van Ginneken, "Automatic coronary calcium scoring in low-dose chest computed tomography," *TMI*, 2012.

[6] R Shahzad, T van Walsum, M Schaap, A Rossi, S Klein, AC Weustink, PJ de Feyter, LJ van Vliet, and WJ Niessen, "Vessel specific coronary artery calcium scoring: an automatic system," *Academic radiology*, 2013.

[7] JM Wolterink, T Leiner, RAP Takx, MA Viergever, and I Išgum, "Automatic coronary calcium scoring in non-contrast-enhanced ecg-triggered cardiac ct with ambiguity detection," *TMI*, 2015.

[8] D Ciresan, A Giusti, LM Gambardella, and J Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *NIPS*, 2012.

[9] J Long, E Shelhamer, and T Darrell, "Fully convolutional networks for semantic segmentation," *CVPR*, 2015.

[10] A Garcia-Garcia, S Orts-Escolano, S Oprea, Victor Villena-Martinez, and Jose Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," *arXiv:1704.06857*, 2017.

[11] O Ronneberger, P Fischer, and T Brox, "U-net: Convolutional networks for biomedical image segmentation," *MICCAI*, 2015.

[12] B Kayalibay, G Jensen, and P van der Smagt, "CNN-based Segmentation of Medical Imaging Data," *arXiv:1701.03056*, 2017.

[13] N Lessmann, I Isgum, AAA Setio AA, BD de Vos, F Ciompi, PA de Jong, M Oudkerk, WPTM Mali, MA Viergever, and B van Ginneken, "Deep convolutional neural networks for automatic coronary calcium scoring in a screening study with low-dose chest ct," in *SPIE Medical Imaging*, 2016.

[14] G Santini, D Della Latta, N Martini, G Valvano, A Gori, A Ripoli, CL Susini, L Landini, and D Chiappino, "An automatic deep learning approach for coronary artery calcium segmentation," in *EMBEC & NBC*. 2017.

[15] S Jégou, M Drozdzal, D Vazquez, A Romero, and Y Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," *CVPR*, 2017.

[16] National Lung Screening Trial Research Team, "Reduced lung-cancer mortality with low-dose computed tomographic screening," *New England Journal of Medicine*, 2011.

[17] G Huang, Z Liu, KQ Weinberger, and L van der Maaten, "Densely connected convolutional networks," *arXiv:1608.06993*, 2016.

[18] JM Wolterink, T Leiner, BD De Vos, JL Coatrieux, BM Kelm, S Kondo, RA Salgado, R Shahzad, H Shu, and M Snoeren, "An evaluation of automatic coronary artery calcium scoring methods with cardiac ct using the orcascore framework," *Medical physics*, 2016.

54 pt
0.75 in
19.1 mm
54 pt
0.75 in
19.1 mm
54 pt
0.75 in
19.1 mm