FULLY-AUTOMATIC DEEP LEARNING BASED SYSTEM FOR AGATSTON SCORE PREDICTION FROM ANY NON-CONTRAST CHEST CT

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

Zebra Medical Vision, Shfaim, Israel

ABSTRACT

The amount of calcium deposits in the coronary arteries is a powerful predictor of cardiovascular events and mortality. Agatson score measured on cardiac CT is the routine method to identify subjects at high risk who might benefit from proactive treatment. We present an automatic method based on fully-convolutional deep neural network to segment coronary calcium and predict Agatston score based on general-indication 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 is highly heterogeneous in terms of originating institutions (15), different acquisition devices (7) and manufacturers (3). We show excellent results on this type of data in terms of Pearson correlation coefficient (0.98), Bland-Altman analysis (bias 0.4 with 95% limits of agreement of [-189.9-190.7]) and linearly weighted Kappa of 0.89 for the risk category assignment.

We also show results on a very large (14,365 scans) cohort from the NLST study, showing correlation with cardiovascular-related clinical outcome as well as the system's ability to proactively detect individuals at high-risk.

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 calcifications (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 using commercially available software. 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], despite the possible motion artifacts.

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 CC in Chest CT, using both local image features and location information (based on a-priory probability map). Shahzad et at [6] detected CC in CSCT scans using a KNN classifier based on local image descriptors and location features based on CCTA atlases. Wolterink et at [7] extended [5] for CSCT, where they used randomized decision trees for vessel-specific CAC classifiers and included human-in-the-loop for ambiguous detections.

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

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 induced by pooling layers, FCNNs introduce skip connections between their downsampling and upsampling paths[11]. The fully-convolutional DenseNet (FC-DenseNet)[14] introduces additional skip connections within the blocks, to simultaneously exploit features from different scales.

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 against a set of 200+ Chest CTs manually annotated by a committee of three radiologists using a standard PACS workstation and 3. We show how the algorithm predictions correlate with clinical outcome based on large-scale data from NLST [15].

2. DATASETS

2.1. Internal

Through partnership with a large health provider in Israel, we've collected 848 adult non-contrast Chest CT scans with

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

high prevalence of significant amount of CC, and used it for training (512) and validation (336). For testing, we have selected an additional set of 203 general indication scans reflecting the distribution of age and gender in the general population. We have removed scans with metallic artifacts and/or signs of deforming pathology such as lung amputation and cardiac surgery. Table 1 presents Agatston score distribution in the datasets.

The train and validation sets were manually annotated using internally developed interactive application where the radiologists could efficiently annotate large number of candidate CC lesions. Only voxels with HU values above the clinical standard threshold for CCS of 130 were considered. Additionally, we had 3 individual experts annotating the test-set data using a standard CCS module on a commercial PACS workstation (Kodak Carestream). Their Agatston scores were averaged to establish ground-truth for reference. Scans with very high level of disagreement between experts were discarded.

2.2. National Lung Screening Trial (NLST)

The National Lung Screening Trial [15] was (2002-2007) a large-scale clinical trial aiming to compare the efficacy of CT screening and standard chest X-ray as methods of lung cancer screening. 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 6.5 years follow-up, 3468 (24.14%) had reported CVD history, 1838 (12.79%) had their coronary calcium reported during the screening (not quantitatively) and 8607 (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 to detect the lungs, trachea and the carina (trachea bifurcation point). Based on the location of the lungs and the carina, we apply as set of empirical rules to detect a bounding box around the heart. The bounding box's shorter dimension in the axial plane is extended isotropically so as to end up with square x-y slices. A soft tissue window (level 40, width 350), which is clinically considered optimal for detecting CC, is applied to the resulting volume. To provide 3d context, the volume is divided into overlapping groups of 3 consecutive slices that will later be fed into the segmentation network to provide pixel-wise prediction for the central slice in the group.

3.2. Fully convolutional neural network

The main idea underlying FCNNs 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" path, respectively. Both in U-net [11] and in the DenseNet [14] 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 figure 1) in order to provide fine-grain localization while propagating 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 [14] architectures. U-Net consists of *2n+1* blocks of 3x3 convolution-dropout-batch normalization-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 about 1/16 of its original size. Every step in the upsampling path consists of a transposed convolution which doubles the feature map 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 equals the number of classes.

Fully Convolutional DenseNet is similar in its block structure to U-Net but has a higher connectivity within each block, called dense block [16], and additional skip connections in the contracting path (shown on Fig. 1). There is also a 1x1 convolution-dropout-batch normalization-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 denoted as *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 represent its probability of being a coronary calcium. To

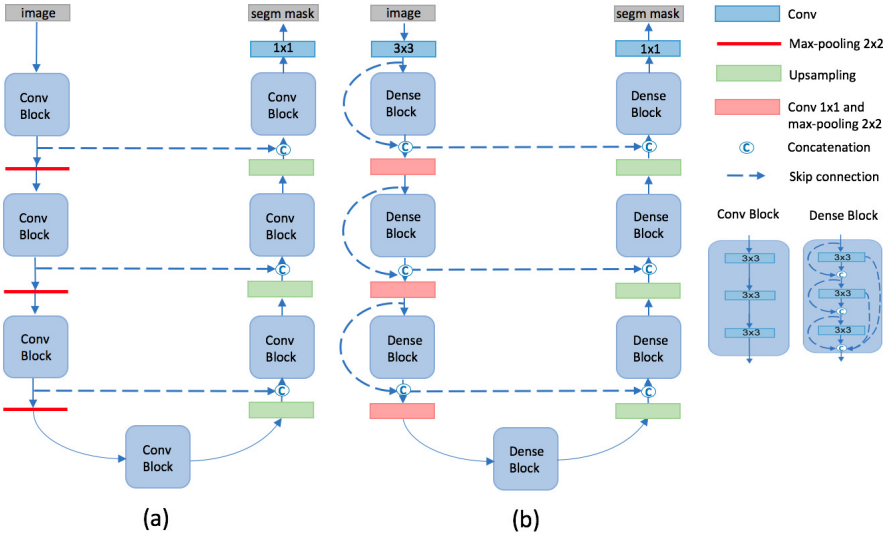


Fig. 1: (a) Generic (for brevity) architecture of U-Net, (b) Generic architecture of a Fully Convolutional DenseNet.

determine the correct threshold for classifying a voxel as CC, we used the validation dataset. First we identify 2d candidate blobs on each axial slice by thresholding with 130 HU and conducting connected-components analysis. Then we characterize each blob by its 95% percentile probability value. Then we exhaustively search for the best threshold in terms of smallest standard deviation of the difference between the predicted Agatston score and the reference, while limiting the search to small bias values (less than 3). Finally, calculation of the Agatston score is done following the clinically accepted protocol described in [3].

4. EXPERIMENTS AND RESULTS

We have 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, 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 amp and L2 weight regularization of 0.0001, for 40000 iterations (roughly 13 epochs) and optimized using RMSProp. We used batches of 12 randomly cropped (up to 75% of the original size) images. We started with learning-rate of 1e-4 and employed exponential decay policy with decay-rate of 0.5 every 8000 iterations. Both networks were implemented using TensorFlow and were trained for approx. 12-14 hours on 3 NVIDIA K80s.

Model selection was done based on the best performing checkpoint with regard to the validation dataset. The same model & parameters were used for all of the experiments.

4.1. Internal

All results reported in this section are based on our test dataset, which was not used during training nor validation. We only report results with regard to the Agatston score, since it is the most clinically relevant measure. We present both a scatter plot showing correlation between the 203 reference and predicted scores in Fig. 2, and a bland-Altman plot in Fig. 3. Fig. 4 shows an examples of segmented calcifications.

In terms of the Pearson correlation coefficient, we achieved r=0.98 ($p < 0.0001$) between the reference and predicted Agatson score. Bland-Altman analysis shows a very small bias of 0.4 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 achieved Pearson 0.97 but introduced bias of -14.5 with limits of agreement of [-252.0-222.9].

We present a risk category agreement Table 2. Linearly weighted Kappa score was 0.89, with 84.7% of the scans placed in the correct category. Another 11.3% were placed only one category away, with almost 55% of the mistakes contributed by confusing the first (zero) and second (1-10) categories, which are both difficult to tell apart and highly sensitive for even a very small mistake. The U-Net architecture achieved Kappa of 0.86 with only 81.8% of the scans placed in the correct category.

4.2. National Lung Screening Trial (NLST)

Applying the algorithm on a large cohort of patients with known CVD history and 6.5 years follow-up provided some very interesting insights. We'll provide a few highlights since a more in depth review of this experiment is pending publication elsewhere:

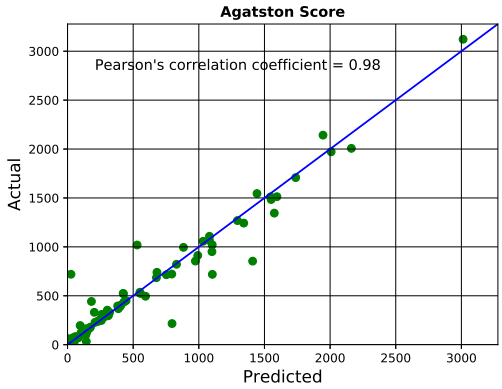


Fig. 2: Reference Agatston score vs. predicted Agatston score

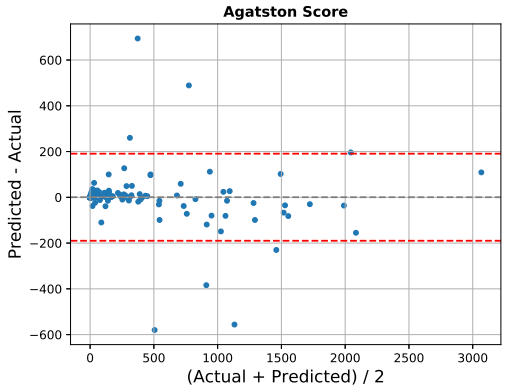
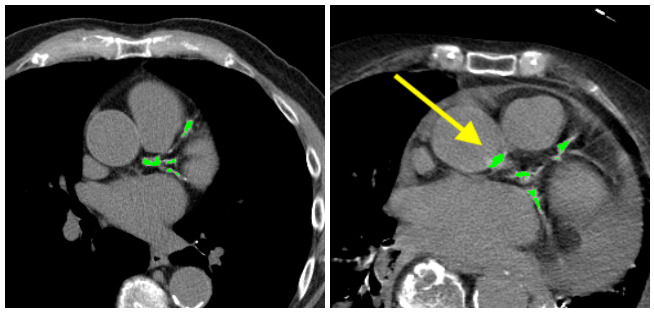


Fig. 3: Bland-Altman plot



(a) (b)

Fig. 4: Examples of segmentation results: (a) Accurate segmentation of calcifications in the LM coronary artery, LCX and LAD, (b) additional false-positive in the aorta wall (marked with an arrow)

		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.

1. 1% of the patients with a calculated CCS of zero died during the follow-up. Comparatively, 4% of those with scores above 400 and 6% of those with scores above 1000 died in the same interval. So we see a correlation of the calculated 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).
3. Radiologists under-report the coronary calcium burden: among patients with scores over 400, 32% were not read as having coronary calcium. Among people whose score is between 100 and 400, 57% were not reported. This alone shows the kind of impact routine application of CCS algorithm on regular Chest CTs can have.

5. SUMMARY AND CONCLUSIONS

We have presented a fully-automated system to estimate the Agatston calcium score in general-indication non-contrast Chest CT scans that yields high Pearson correlation (0.98), high level of agreement based on Bland-Altman analysis and very good risk category stratification (Kappa of 0.89). Additionally, our method was applied to a large cohort (∼15K scans) of NLST data and showed very good correlation with clinical outcome. We have evaluated two network architectures: the popular U-Net [11] and the more recent DenseNet [14] which proved to be superior in terms of CCS prediction.

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.

The orCaScore challenge [17] offers a public dataset to evaluate algorithms for CCS prediction. However, the data in the challenge is cardiac CT while we address the much more common Chest CT, therefore currently the challenge is not applicable to our method.

We plan apply our system it in a real clinical setting soon to enable early detection of individual at risk of CVD.

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

[15] "Reduced lung-cancer mortality with low-dose computed tomographic screening," *New England Journal of Medicine*, 2011.

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

[17] 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.