Machine Vision Based Scoring of Coronary Calcium

Ibrahim Kecoglu
Biomedical Physics - Stanford School of Medicine
291 Campus Drive, Stanford, CA 94305
ikecoglu@stanford.edu

Abstract

Coronary artery calcium (CAC) scores are great non-invasive indicators of future cardiovascular problems, yet their routine use is restricted by the need for electrocardiogram-gated CT. We investigated whether vessel-specific CAC scoring can be performed opportunistically on the far more common non-gated chest CT scans. Three regression architectures were utilized: a compact 2-D convolutional neural network (CNN), the same CNN augmented with a squeeze-and-excite block (SE-CNN), and a vision transformer (ViT) that processes 32×32 patches as tokens. All networks are trained on 212 non-gated CT scans with 73 slices each (from the public COCA dataset), using expert Agatston scores for the left coronary artery (LCA), left anterior descending (LAD), left circumflex (LCX), and right coronary artery (RCA) as ground truth. Five-fold cross-validation was applied during hyperparameter tuning, and the models were evaluated on a held-out test set. Test mean-squared errors less than 0.0088 were achieved across all models. These results demonstrate that accurate, vessel-level calcium scoring is feasible without gating, paving the way for widespread, cost-free cardiovascular risk assessment from routine thoracic imaging.

1. Introduction

Coronary artery calcium (CAC) refers to the accumulation of calcified plaque within the coronary arteries. Its scoring, commonly expressed as the Agatston score, provides a measure of atherosclerotic burden. It is shown in the literature that higher CAC scores correlate strongly with one's risk of future cardiovascular events, making them valuable in guiding preventive therapies. [5, 3, 10] Traditionally, CAC is measured using electrocardiogram-gated computed tomography (CT), which synchronizes CT image acquisition with the cardiac cycle to minimize motion artifacts. However, many patients already undergo non-gated chest CT scans for other clinical indications, such as lung cancer screening. Therefore, if CAC could be accurately assessed from these examinations, it would offer a cost-effective and opportunistic means of cardiovascular risk assessment. Unfortunately, the lack of cardiac gating introduces motion blur, variable image quality, and inconsistent slice positioning, all of which can complicate coronary calcium quantification.

The input to our algorithm is a set of non-gated thoracic CT volumes (three-dimensional stacks of axial slices). We then compare three different regression-based machine vision architectures. A standard convolutional neural network (CNN), the same CNN augmented with a squeeze-and-excite layer (SE-CNN) which learns to weight each CT slice's contribution, and also a vision transformer (ViT). The models output four vessel-specific calcium scores (LCA = left coronary artery, LAD = left anterior descending, LCX = left circumflex, and RCA = right coronary artery). Each model is trained on a dataset of non-gated chest CT scans paired with ground-truth LCA, LAD, LCX, and RCA scores determined by experts. Consequently, the models learn to regress directly from images to the numeric measures of calcified plaque burden in each coronary segment. After carefully tuning our models, we were able to achieve a test set (20% of the patients) mean squared error (MSE) of 0.0084-0.0087.

2. Related Work

There is already an extensive literature on the use of different machine learning approaches for CAC scoring tasks on gated CT datasets. [14, 13, 9, 15] However, as discussed in the introduction, the nongated case is a harder problem as it brings additional complications. On that front, Eng *et al*. [4] has utilized a U-net type CNN model to determine CAC score levels (into 5 levels, split by score thresholds) for both gated and nongated CT scans. They opted for a classification approach rather than a regression approach. Their encoder is a 50-layer SE-ResNeXt 2D CNN, which was pretrained on ImageNet. Utilizing the power of such a large dataset like ImageNet for this task is an interesting approach. Nevertheless, the CT scan dataset and ImageNet are quite different from each other, and successfully trans-

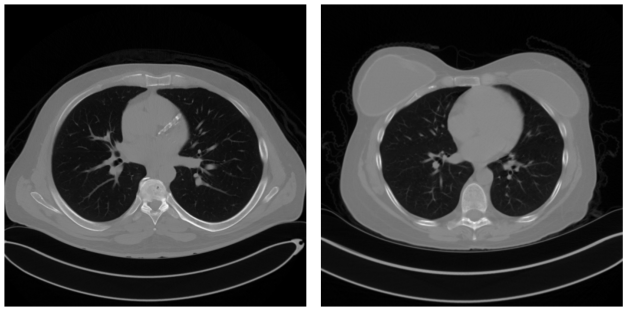


Figure 1: Example slices from patients with and without coronary artery calcification. The patient on the left has calcification shown as bright structures at the heart while the other is clear.

ferring feature extraction capability from one dataset to the other is a challenge on its own. In our project, we utilized simpler and shallower architectures. Moreover, we focused on directly estimating the scores for different coronary arteries (LCA, LAD, LCX, RCA).

3. Dataset and Features

We used the dataset COCA - Coronary Calcium and chest CT's [1] hosted by Stanford AIMI Center. This dataset contains both gated and nongated CT scans with their corresponding CAC scores. Since our motivation was to develop a model that would work on nongated CT, we only focused on that portion of the dataset. There were 212 patient CT scans and matched scores in the dataset. Each CT scan is composed of a number of slices in the form of DICOM images (512 x 512 px), which constitute a 3D volume. The number of images for each patient slightly varies; therefore, we clipped or padded the 3D volumes around the center, which includes the heart (our main focus), to normalize the depth of scans across patients. Thus, in their final form, each CT scan is composed of 73, 512 x 512 images. We also normalized each scan within itself before feeding it to the models. Two sample slices, one patient with visible CAC and one without any, are shown in Figure 1. Before feeding the data to the models, we patient-wise split the dataset into training (80% - 169 patients) and testing (20% - 43 patients) sets. During the development and hyperparameter tuning of the models, the training set was further split into training and validation sets under a 5-fold cross-validation scheme.

4. Methods

Lets represent the dataset as $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ with $x_i \in \mathbb{R}^{C\times H\times W}$ being the CT scans and $y_i \in \mathbb{R}^4$ being the scores. We have taken the depth axis (different layers) as different channels. Each volume is resampled to $C = 73$ axial slices of size $H = W = 512$ and then stacked along the channel axis. The target vector $y_i = (\mathrm{LCA}, \mathrm{LAD}, \mathrm{LCX}, \mathrm{RCA})$ contains Agatston-like calcium scores for the four coronary arteries.

All models learn a mapping $f_\theta : \mathbb{R}^{73\times512\times512} \to \mathbb{R}^4$ by minimising the mean–squared error (MSE) loss

$$\mathcal{L}(\theta) \;=\; \frac{1}{N}\sum_{i=1}^{N}\big\|f_\theta(x_i) \;-\; y_i\big\|_2^2 \tag{1}$$

For optimization, we used the AdamW algorithm [2]:

Input :

γ (lr), β_1, β_2 (betas),

θ_0 (params), $f(\theta)$ (objective),

ϵ (epsilon), λ (weight decay),

amsgrad, *maximize*

Initialize :

$m_0 \leftarrow 0$ (first moment),

$v_0 \leftarrow 0$ (second moment),

$v_0^{\max} \leftarrow 0$

For $t = 1$ **to** T **do**

if *maximize* :

$g_t \leftarrow -\nabla_\theta f_t\big(\theta_{t-1}\big)$

else :

$g_t \leftarrow \nabla_\theta f_t\big(\theta_{t-1}\big)$

$\theta_t \leftarrow \theta_{t-1} \;-\; \gamma\,\lambda\,\theta_{t-1}$

$m_t \leftarrow \beta_1\, m_{t-1} + (1-\beta_1)\, g_t$

$v_t \leftarrow \beta_2\, v_{t-1} + (1-\beta_2)\, g_t^2$

$\widehat{m}_t \leftarrow \dfrac{m_t}{1-\beta_1^t}$

if *amsgrad* :

$v_t^{\max} \leftarrow \max\big(v_{t-1}^{\max},\, v_t\big)$

$\widehat{v}_t \leftarrow \dfrac{v_t^{\max}}{1-\beta_2^t}$

else :

$\widehat{v}_t \leftarrow \dfrac{v_t}{1-\beta_2^t}$

$\theta_t \leftarrow \theta_t \;-\; \gamma\,\dfrac{\widehat{m}_t}{\sqrt{\widehat{v}_t}+\epsilon}$

---

return θ_t

AdamW is a variant of the Adam optimizer that decouples the weight decay term, applying the regularization directly rather than through the moment estimates. This separation ensures more stable and effective regularization. Which usually leads to better generalization compared to the original Adam.

We utilized cross-validation (5-fold) during the development and hyperparameter tuning. We experimented with the general structures of our models until we could overfit a smaller subset of our data. Then we moved on to the larger dataset and tuning to get the best outcome from the model for the entire dataset. Batch sizes, learning rates, and weight decays for the optimizers were all determined by a grid search during hyperparameter tuning. Best performances were achieved by a batch size of 32 for all the models, but the other parameters varied across the models. All models are developed and optimized using the PyTorch environment[2].

4.1. CNN

One of the models we tried was a very conventional convolutional neural network (CNN) architecture. It consisted of three $\text{Conv2d}-\text{BN2d}-\text{ReLU}-\text{MaxPool2}d$ blocks with $\{32, 64, 128\}$ filters. Filter sizes were 3×3. Then, a global average pooling and a two-layer fully connected network $(128\rightarrow128\rightarrow4)$ was applied.

The convolutional layers learn the local structures and patterns. As the model gets deeper, the receptive field enlarges, allowing the network to aggregate spatial cues across the thorax. Global average pooling collapses the feature map to a single vector, which is linearly regressed to the four coronary artery calcium scores. Even though the model seems simple for the task, it also captures cross-slice context since the channels correspond to the anatomical z-axis.

4.2. SE-CNN

For the second model, we attached a squeeze-and-excite (SE) module [7] in front of the same convolutional network. Given an input $x \in \mathbb{R}^{C\times H\times W}$, the SE block computes

$$s_c \;=\; \sigma\big(W_2\,\phi(W_1\,g_c)\big), \qquad g_c = \tfrac{1}{HW}\textstyle\sum_{h,w} x_{c,h,w}, \;\; (2)$$

where ϕ and σ are ReLU and sigmoid, $W_1 \in \mathbb{R}^{C/r\times C}$, $W_2 \in \mathbb{R}^{C\times C/r}$ and $r = 16$. The activation after the block is $\tilde{x}_{c,h,w} = s_c x_{c,h,w}$.

The SE mechanism learns slice-level importance weights s_c that emphasise relevant slices and suppress background. The downstream CNN thus receives a re-weighted signal. Since the chest CTs have regions above and below the heart, which are not really important for our task. By this approach, we aimed to make the model learn a weighting scheme to compensate for this.

4.3. ViT

Finally, we also experimented with a vision transformer (ViT) architecture. Each input is divided into non-overlapping 32×32 patches. A 32×32 convolution with stride 32 projects every patch to a vector, resulting in $P = 256$ tokens. We append a learnable class token in the front and apply positional embedding. The sequence then passes through the transformer layers with depth 3, multi-head self-attention (4 heads), and a feed-forward dimension of 512. The output of the class token is finally fed to an MLP head to produce the prediction.

Self-attention allows every patch to exchange information with every other patch, capturing long-range dependencies. The learnable class token acts as a global aggregator whose final embedding summarizes the entire scan for the regression.

5. Results & Discussion

As described under the methods section, we have chosen the MSE loss as the loss metric for our regression task. After finalizing the model structures and using cross-validation on the larger dataset for determining the best hyperparameters (learning rate, weight decay, batch size, etc.) Table 1 compares the optimal performance of the models we were able to achieve. It presents the mean training and validation loss across cross-validation folds, and their final test performances. Despite having significant differences in their structures and complexities, all the models showed a similar performance.

It is always important to investigate the loss curves for any signs of training pathologies such as overfitting or underfitting. In Figure 2, we present the mean loss curves and their standard deviations. When investigated, the training performances seem to be consistent across models, and there is no clear sign of over or underfitting.

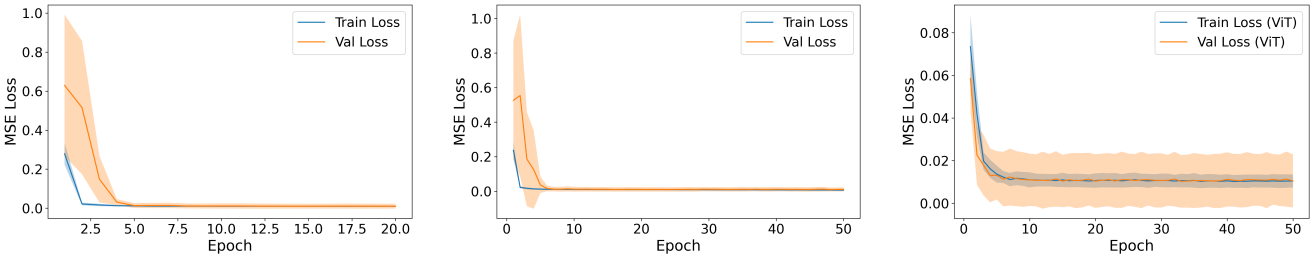
Finally, we also produced saliency maps using the CNN model to get a better qualitative understanding of how our model works. In Figure 3, an example saliency map is shown. We can clearly see that there are hot spots on the heart, the sternum, the ribs, and the spine. The focus on the heart is expected, since we are already training the model for the exact purpose of determining the level of coronary artery calcification. Yet, the bony structures might seem odd at first glance. We believe this is also an expected result as the calcified plaque in the arteries and the bones both contain calcium. Resulting in similar attenuation coefficients. Hence, they both have similar intensities (HU units) in the CT scans.

6. Conclusion & Future Work

All our models showed less than 0.0088 MSE on the test set with very similar performances. The CNN model seems

Model	Mean Train MSE $\pm$ std	Mean Val. MSE $\pm$ std	Test MSE
CNN	0.0098 ± 0.0030	0.0103 ± 0.0122	0.0084
SE-CNN	0.0068 ± 0.0038	0.0108 ± 0.0125	0.0087
ViT	0.0104 ± 0.0030	0.0105 ± 0.0124	0.0086

Table 1: Performance comparison of models. Mean train and validation MSEs are computed over cross-validation folds. "std" stands for standard deviation.



(a) CNN (b) SE-CNN (c) ViT

Figure 2: Shaded loss curves (over 5-folds) of each model.

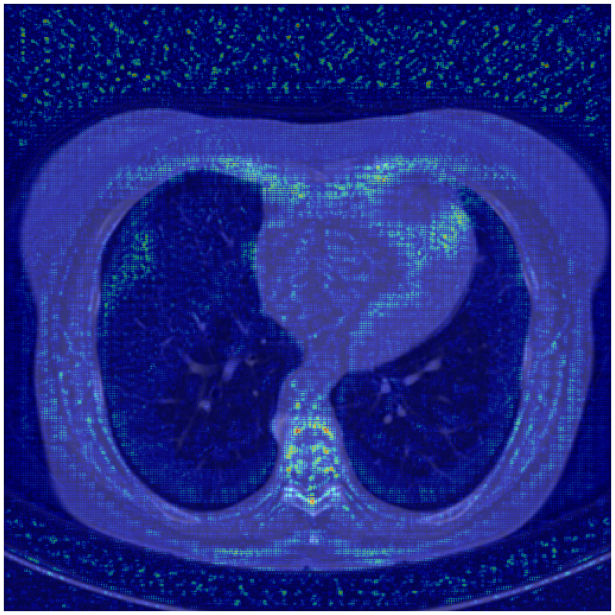


Figure 3: Example saliency map from the CNN model.

to be the best choice due to its simplicity and performance. Though having observed that the SE-CNN and ViT models can also be utilized with similar performances for this task, they make a great candidate to consider for future works where the dataset and complexity of the task might be different.

One downside of our study was the lack of data augmentation techniques due to limited time and compute. Particularly, adding different levels of noise, blur, and shifts to better span possible contributors to the problems with visualization and scoring of coronary calcium. Such effects are of clinical importance in the form of detector efficiencies, device and stabilization differences, patient motion, etc. Given enough resources, we would like to expand our dataset with such augmentations.

7. Contributions & Acknowledgements

Other than the collection of the dataset, which is shared by Stanford AIMI center, this project was carried out by Ibrahim Kecoglu and solely for the CS 231N class.

References

[1] COCA - Coronary Calcium and chest CT's. `https://stanfordaimi.azurewebsites.net/datasets/e8ca74dc-8dd4-4340-815a-60b41f6cb2aa`, July 2021.

[2] J. Ansel, E. Yang, H. He, N. Gimelshein, A. Jain, M. Voznesensky, B. Bao, P. Bell, D. Berard, E. Burovski, G. Chauhan, A. Chourdia, W. Constable, A. Desmaison, Z. DeVito, E. Ellison, W. Feng, J. Gong, M. Gschwind, B. Hirsh, S. Huang, K. Kalambarkar, L. Kirsch, M. Lazos, M. Lezcano, Y. Liang, J. Liang, Y. Lu, C. K. Luk, B. Maher, Y. Pan, C. Puhrsch, M. Reso, M. Saroufim, M. Y. Siraichi, H. Suk, S. Zhang, M. Suo, P. Tillet, X. Zhao, E. Wang, K. Zhou, R. Zou, X. Wang, A. Mathews, W. Wen, G. Chanan, P. Wu, and S. Chintala. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, pages 929–947, La Jolla CA USA, Apr. 2024. ACM.

[3] L. Azour, M. A. Kadoch, T. J. Ward, C. D. Eber, and A. H. Jacobi. Estimation of cardiovascular risk on routine chest CT: Ordinal coronary artery calcium scoring as an accurate predictor of Agatston score ranges. *Journal of Cardiovascular Computed Tomography*, 11(1):8–15, Jan. 2017.

[4] D. Eng, C. Chute, N. Khandwala, P. Rajpurkar, J. Long, S. Shleifer, M. H. Khalaf, A. T. Sandhu, F. Rodriguez, D. J. Maron, S. Seyyedi, D. Marin, I. Golub, M. Budoff, F. Kitamura, M. S. Takahashi, R. W. Filice, R. Shah, J. Mongan, K. Kallianos, C. P. Langlotz, M. P. Lungren, A. Y. Ng, and B. N. Patel. Automated coronary calcium scoring using deep learning with multicenter external validation. *npj Digital Medicine*, 4(1):1–13, June 2021. Publisher: Nature Publishing Group.

[5] P. Greenland, M. J. Blaha, M. J. Budoff, R. Erbel, and K. E. Watson. Coronary Calcium Score and Cardiovascular Risk. *JACC*, 72(4):434–447, July 2018. Publisher: American College of Cardiology Foundation.

[6] C. R. Harris, K. J. Millman, S. J. Van Der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. Van Kerkwijk, M. Brett, A. Haldane, J. F. Del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, Sept. 2020.

[7] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu. Squeeze-and-Excitation Networks, May 2019. arXiv:1709.01507 [cs].

[8] J. D. Hunter. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[9] D. Mu, J. Bai, W. Chen, H. Yu, J. Liang, K. Yin, H. Li, Z. Qing, K. He, H.-Y. Yang, J. Zhang, Y. Yin, H. W. McLellan, U. J. Schoepf, and B. Zhang. Calcium Scoring at Coronary CT Angiography Using Deep Learning. *Radiology*, 302(2):309–316, Feb. 2022. Publisher: Radiological Society of North America.

[10] P. O. Neves, J. Andrade, and H. Monção. Coronary artery calcium score: current status. *Radiologia Brasileira*, 50:182–189, June 2017. Publisher: Publicação do Colégio Brasileiro de Radiologia e Diagnóstico por Imagem.

[11] T. pandas development team. pandas-dev/pandas: Pandas, Sept. 2024.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[13] W. Wang, H. Wang, Q. Chen, Z. Zhou, R. Wang, H. Wang, N. Zhang, Y. Chen, Z. Sun, and L. Xu. Coronary artery calcium score quantification using a deep-learning algorithm. *Clinical Radiology*, 75(3):237.e11–237.e16, Mar. 2020.

[14] D. J. Winkel, V. R. Suryanarayana, A. M. Ali, J. Görich, S. J. Buß, A. Mendoza, C. Schwemmer, P. Sharma, U. J. Schoepf, and S. Rapaka. Deep learning for vessel-specific coronary artery calcium scoring: validation on a multi-centre dataset. *European Heart Journal - Cardiovascular Imaging*, 23(6):846–854, June 2022.

[15] N. Zhang, G. Yang, W. Zhang, W. Wang, Z. Zhou, H. Zhang, L. Xu, and Y. Chen. Fully automatic framework for comprehensive coronary artery calcium scores analysis on non-contrast cardiac-gated CT scan: Total and vessel-specific quantifications. *European Journal of Radiology*, 134:109420, Jan. 2021.