Machine-Learning Based Detection of Coronary Artery Calcification Using Synthetic Chest X-Rays

Dylan Saeed[1] D.SAEED@STUDENT.UNSW.EDU.AU
Ramtin Gharleghi[3] R.GHARLEGHI@UNSW.EDU.AU
Susann Beier[2] S.BEIER@UNSW.EDU.AU
Sonit Singh[1] SONIT.SINGH@UNSW.EDU.AU

[1] *School of Computer Science and Engineering, University of New South Wales, Australia*
[2] *School of Mechanical and Manufacturing Engineering, University of New South Wales, Australia*
[3] *Independent Researcher, Sydney, Australia*

Editors: Under Review for MIDL 2026

Abstract

Coronary artery calcification (CAC) is a strong predictor of cardiovascular events, with CT-based Agatston scoring widely regarded as the clinical gold standard. However, CT is costly and impractical for large-scale screening, while chest X-rays (CXRs) are inexpensive but lack reliable ground truth labels, constraining deep learning development. Digitally reconstructed radiographs (DRRs) offer a scalable alternative by projecting CT volumes into CXR-like images while inheriting precise labels. In this work, we provide the first systematic evaluation of DRRs as a surrogate training domain for CAC detection. Using 667 CT scans from the COCA dataset, we generate synthetic DRRs and assess model capacity, super-resolution fidelity enhancement, preprocessing, and training strategies. Lightweight CNNs trained from scratch outperform large pretrained networks; pairing super-resolution with contrast enhancement yields significant gains; and curriculum learning stabilises training under weak supervision. Our best configuration achieves a mean AUC of 0.754, comparable to or exceeding prior CXR-based studies. These results establish DRRs as a scalable, label-rich foundation for CAC detection, while laying the foundation for future transfer learning and domain adaptation to real CXRs.

Keywords: Coronary artery calcification, digitally reconstructed radiographs, deep learning, super-resolution, domain adaptation, medical imaging, chest X-ray

1. Introduction

Coronary artery calcification (CAC) is a well-established marker of atherosclerotic burden and a strong predictor of cardiovascular events and all-cause mortality (Budoff et al., 2007). Quantification using ECG-gated cardiac CT via the Agatston method remains the clinical gold standard for risk stratification (Agatston et al., 1990; Grundy et al., 2019). However, CT screening is costly, resource-intensive, and impractical for large-scale population use (Søgaard et al., 2022), motivating alternative modalities that are low-cost, scalable, and suitable for population-level risk assessment.

In contrast, chest X-rays (CXRs) are inexpensive and widely accessible, making them an attractive modality for opportunistic CAC risk assessment (D'Ancona et al., 2023).

Yet, CXRs have limited sensitivity for calcium depiction and lack standardised annotation protocols, leaving reliable ground truth labels scarce (Kamel et al., 2021). Existing CXR–CT datasets are typically assembled retrospectively, requiring patients to have undergone both imaging modalities within a clinically acceptable window (often 6 months). Such dual-modality acquisitions are uncommon in routine care and, when available, can introduce temporal misalignment and label noise due to disease progression between scans. This bottleneck constrains the development of deep learning methods, and prior studies (Kamel et al., 2021; D'Ancona et al., 2023; Jeong et al., 2024) remain limited in scale and generalisability.

In such data-scarce settings, an emerging strategy is to train models within surrogate domains: synthetic or simulated approximations of the target modality that preserve key imaging physics while offering abundant labels. In clinical domains, Digitally reconstructed radiographs (DRRs) are synthetic 2D projections of CT volumes that approximate real CXRs while being able to inherit precise CT-derived labels. Prior work in their generation has validated their clinical fidelity (Moore et al., 2011) suggesting that DRRs may serve as a scalable proxy training domain where methodological feasibility can be established before transferring to real CXRs. Although direct transfer to real CXRs is not evaluated here, recent work has shown DRR-trained models to generalize effectively in fluoroscopic and interventional settings, supporting the plausibility of similar transferability for CAC detection(Unberath et al., 2018).

To our knowledge, this work provides the first systematic evaluation of DRRs as a surrogate training domain for CAC detection. We evaluate feasibility across model capacity, fidelity enhancement, preprocessing, and training strategies, and show that DRRs constitute a scalable and precisely labelled domain for developing models prior to transfer to real CXRs, laying the foundation for low-cost, population-scale cardiovascular risk screening using existing radiography infrastructure.

2. Related Work

2.1. CT-based CAC detection

ECG-gated cardiac CT is the clinical gold standard for CAC quantification via the Agatston method (Agatston et al., 1990), and numerous studies have leveraged CT directly for automated CAC detection and scoring using deep learning(Eng et al., 2021). These approaches benefit from precise attenuation-based ground truth but remain limited by the cost, radiation dose, and infeasibility of CT screening at the population level (Søgaard et al., 2022).

2.2. CXR-based CAC detection

Given the wide availability of chest radiographs, several groups have explored their utility for CAC risk assessment. (Kamel et al., 2021) trained an attention-augmented VGG16 on 1,689 CXRs paired with CTs, achieving an AUC of 0.73 for CAC prediction. (D'Ancona et al., 2023) demonstrated that deep learning on CXRs could refine pretest probability estimation in suspected angina patients, validated against invasive coronary angiography.(Jeong et al., 2024) proposed a radiomics-based approach requiring manual cardiac segmentation, reporting an AUC of 0.808 for detecting moderate-to-severe CAC (>100 Agatston units). Collectively, these approaches demonstrate the potential of CXRs for CAC screening, yet despite their promising results, progress remains constrained by the difficulty of assembling large paired CXR–CT datasets, the weak sensitivity of CXRs to calcification, and dependence on manual segmentation—factors that hinder scalability and reproducibility.

2.3. Synthetic imaging with DRRs

DRRs have been widely used in radiotherapy and orthopaedics to simulate radiographs from CT volumes, and prior work has validated their anatomical and clinical fidelity (Moore et al., 2011). Recent studies have begun to explore DRRs as surrogate datasets for training deep learning models in scenarios where paired real-world imaging and labels are scarce (Unberath et al., 2018). However, these efforts have primarily focused on registration or dose optimisation rather than disease detection. To date, no systematic evaluation has been conducted on their feasibility as a surrogate training domain for CAC detection.

3. Methods

3.1. Dataset and Labels

We use the publicly available Coronary Calcium and Chest CT (COCA) dataset (Center, 2022), which contains 790 ECG-gated cardiac CT scans paired with coronary artery calcium (CAC) segmentations in `xml` format. For each patient, a total Agatston score was computed by summing across per-artery calcium masks. Following clinical convention, we binarise labels at a threshold of 100: patients with scores $\leq$100 are classified as negative (no/mild CAC), and those >100 as positive (moderate/severe CAC). This threshold was not tuned, but chosen because it reflects the established clinical cutoff between non-actionable and clinically significant CAC (Grundy et al., 2019).

To ensure geometric consistency, CT volumes were resampled to isotropic δx mm spacing, where δx was matched to the in-plane pixel spacing of the source DICOM. Patients with insufficient slice coverage ($s < 30$) were excluded for quality control, yielding 667 usable scans. Although gated CTs do not capture full thoracic coverage, they do represent the heart-centred regions most relevant for CAC detection. Comparable regions could be isolated in real CXRs with coarse heart localisation, suggesting that this dataset remains a reasonable surrogate for feasibility testing.

3.2. Digitally Reconstructed Radiographs (DRRs)

Synthetic radiographs were generated directly from CT volumes using the Siddon ray-tracing algorithm, implemented in the open-source `DiffDRR` framework (Gopalakrishnan and Golland, 2022). Siddon projection computes exact line integrals through the CT volume, avoiding interpolation artefacts and preserving small, high-density structures such as calcifications. For each scan, we simulated posterior–anterior (PA) and lateral (LA) projections under a fan-beam geometry with fixed source–detector distance (1085.6 mm) and detector width of 512 pixels at 1 mm spacing. Both PA and LA projections used identical source–detector parameters, with the LA view generated by a 90° rotation around the cranio-caudal axis to approximate orthogonal orientation. DRRs were normalised to

$[0, 1]$ and resized to 512×512. This pipeline yields synthetic radiographs that approximate clinical CXRs while retaining precise CAC labels inherited from the source CTs.

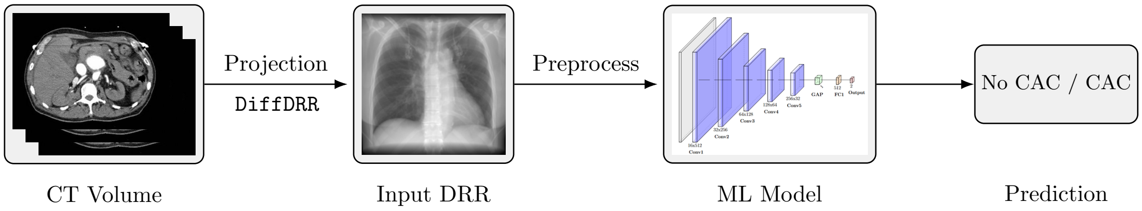


Figure 1: A CT volume (left) is projected into a DRR using Siddon's algorithm (DiffDRR) and fed to a classifier to predict a binary CAC label (Agatston > 100).

3.3. Image Enhancement Strategies

To test whether DRRs provide sufficient fidelity for CAC detection—or whether additional enhancements are needed—we evaluate two complementary strategies: **(i)** pre-projection super-resolution and **(ii)** post-projection preprocessing.

3.3.1. Pre-projection super-resolution

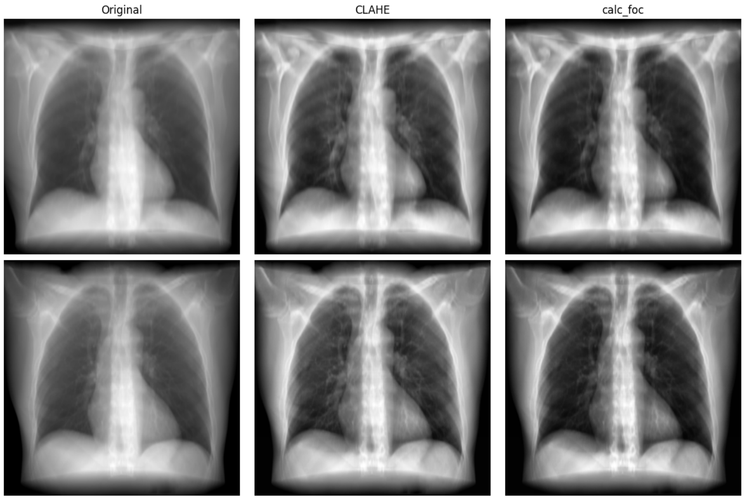
Native CTs often have anisotropic voxels, where coarse in-plane resolution, from a small number of axial slices, can obscure small calcifications once projected. To test whether resolution recovery enhances CAC depiction, we apply a 4× SRResNet (Ledig et al., 2016) to sagittal slices prior to projection. The super-resolved slices were then reassembled into a volume, resampled isotropically, and projected identically to the native-resolution pipeline. This comparison probes whether fine detail restoration materially improves DRR-based detection.

3.3.2. Post-projection image preprocessing

We further test whether DRR adjustments can aid detection by making CAC visually more salient. Three variants were compared: **(1) Original:** unaltered projection. **(2) CLAHE:** contrast-limited adaptive histogram equalisation to locally enhance soft tissue, implemented using an 8×8 tile grid and clip-limit of 2.0, while the unsharp mask used a 5×5 Gaussian kernel (σ=1.0) with a gain of 1.5. **(3) Calc-focused:** a composite filter designed to highlight calcifications and suppress irrelevant anatomy. Comprised of gamma correction ($\gamma = 1.5$), CLAHE and unsharp masking with a 5×5 Gaussian kernel (σ=1.0). This axis examines whether heuristic contrast enhancement improves learnability, or whether native DRRs are sufficient.

3.4. Model Architectures

To assess the robustness across model capacity and inductive bias, we evaluate three representative networks: **(i) CNN5_GAP:** a lightweight custom CNN with five convolutional blocks, global average pooling, and a two-layer classifier. Each block comprised a 3×3 kernel with LeakyReLU activations and batch normalization, followed by dropout (p=0.2–0.3) in the classifier head for regularisation. **(ii) ResNet18**(He et al., 2015) a moderate-capacity residual network serving as a standard baseline.



(a) Original (b) CLAHE (c) Calc-focused

Figure 2: Preprocessing for Native DRR (Top) and Super-resolved DRR (bottom)

(iii) DenseNet121(Huang et al., 2016) pretrained on CheXpert and augmented with a lightweight self-gating spatial attention mechanism derived from CBAM. More specifically, given feature maps $F \in \mathbb{R}^{B\times C\times H\times W}$, we compute a spatial attention mask

$$M = \sigma\left(\frac{1}{C}\sum_{c=1}^{C} F_c\right)$$

where $\sigma(\cdot)$ denotes the sigmoid function. The attended feature map is then $\tilde{F} = F \odot M$, which reweights activations by their spatial salience prior to global pooling.

For multi-view experiments, we use dual-encoder variants incorporating postero-anterior (PA) and lateral (LA) projections. Fusion is tested at three levels: (i) early image concatenation, (ii) latent feature concatenation, and (iii) cross-attention. These designs examine whether DRRs encode complementary view information analogous to real CXR studies.

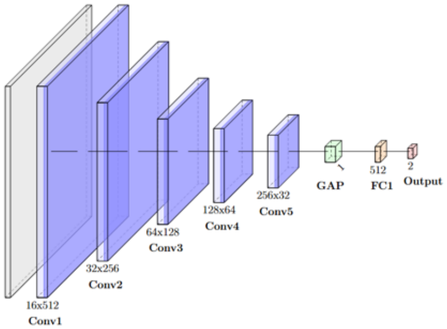


Figure 3: Architecture of the custom CNN5_GAP model, consisting of five convolutional blocks and a 2-layer classifier.

3.5. Training & Evaluation

Models were trained with binary cross-entropy loss and label smoothing ($\varepsilon = 0.1$) using Adam (Kingma and Ba, 2017) (learning rate 10^{-4}, weight decay 10^{-5}, batch size 32). Augmentations consisted of random affine transformations: rotations ($\pm 5°$), translations ($\pm 5\%$), scaling ($[0.9, 1.1]$) and shearing ($10°$). Horizontal flips were excluded to preserve anatomical context. Early stopping was applied on validation AUC (being the criterion for model selection), with dropout/batch normalisation for regularisation. Multi-view fusion models used a curriculum schedule(Bengio et al., 2009) (extremes first, then borderline), intended as a heuristic to reduce instability from thresholding.

Performance was evaluated with stratified 5-fold Cross Validation (CV) across 5 seeds (25 runs) with approximately 530 training and 130 validation DRRs. Metrics included AUC-ROC (primary), with accuracy, precision, recall, and F1 as secondary. All splits were patient-level and stratified by CAC label to avoid data leakage across folds. To reduce optimism bias from volatile early epochs, we discarded the first five epochs and averaged performance over the top five subsequent epochs. This conservative reporting may slightly underestimate peak metrics, but ensures robustness. Statistical significance was assessed using paired Wilcoxon signed-rank tests(Hollander et al., 2013) across validation folds. Results were considered significant at $p < 0.05$.

All preprocessing, training, and evaluation code is implemented in PyTorch and will be made publicly available upon publication for reproducibility.

4. Results

We present results in a structured manner, beginning with single-view baselines and ablations on preprocessing and super-resolution, followed by multi-view fusion strategies, and finally training variants such as curriculum learning. Unless otherwise noted, all results are reported as mean validation AUC (± standard deviation) across 25 runs, with statistical significance assessed using the Wilcoxon signed-rank test.

4.1. Baselines, Preprocessing, and Super-resolution

To assess the impact of preprocessing and super-resolution (SR) on model performance, we evaluated three architectures (CNN5_GAP, DenseNet121, and ResNet18) across three DRR preprocessing modes (Original, CLAHE, Calc-focused) under both native and super-resolved inputs (Table 1). For CNN5_GAP, mean AUCs improved by +0.015 ($0.720 \rightarrow 0.735$) from native-original to SR-calc-focused, with SR yielding significant gains under calc-focused and CLAHE preprocessing ($p < 0.05$). DenseNet121 performance degraded ($0.718 \rightarrow 0.688$; $p < 0.05$) under contrast-enhancing preprocessing, with SR offering only minor recovery. ResNet18 remained stable across most modes, with no significant differences observed.

This divergence highlights that pretrained backbones, while strong under natural-image statistics, may be brittle to synthetic-domain contrast shifts. Thus, given its consistent and competitive performance, CNN5_GAP was selected as the primary backbone for subsequent experiments, suggesting that compact task-specific CNNs trained from scratch may generalise better within synthetic domains where texture and contrast statistics could differ to real-world pretraining distributions.

Table 1: Effect of preprocessing and super-resolution on validation AUC for CNN5_GAP, augmented DenseNet121 and ResNet18. [1]

Preprocessing	CNN5_GAP		DenseNet121		ResNet18	
	Native	SR	Native	SR	Native	SR
Original	0.720 ± 0.048	0.727 ± 0.050	**0.718 ± 0.031**†	0.708 ± 0.026	0.714 ± 0.031	0.717 ± 0.031
CLAHE	0.728 ± 0.049	**0.733 ± 0.043***	0.688 ± 0.053	0.695 ± 0.028	0.716 ± 0.068	0.702 ± 0.022
Calc-foc	0.729 ± 0.053	**0.735 ± 0.043***	0.683 ± 0.058	0.688 ± 0.042	0.722 ± 0.070	0.723 ± 0.032

4.2. Fusion & Training Strategies

We next evaluated whether combining posterior–anterior (PA) and lateral (LA) projections could improve performance over best performing single-view (PA) baseline of 0.735 ± 0.043. Three fusion strategies were tested: early fusion (image concatenation), intermediary fusion (latent feature concatenation), and cross-attention fusion. Each was implemented with either shared or unshared encoders and a learnable scalar-gated mechanism to weigh each view's contribution.

Table 2: Performance of PA+LA fusion strategies, varying by level of interaction (early, intermediary, or attention) and whether encoders are shared or unshared.

Fusion Strategy	Shared Encoder	Unshared Encoder
PA-Only (Baseline)	0.735 ± 0.043	
Early Fusion	0.702 ± 0.044	–
Intermediate Fusion	0.736 ± 0.059	0.739 ± 0.065
Cross-Attention Fusion	**0.740 ± 0.043**	0.729 ± 0.045

As shown in Table 2, early fusion was significantly worse ($p < 0.05$) relative to the PA-only baseline (0.735 ± 0.043). Intermediary fusion achieved mean AUCs of 0.736 (shared encoders) and 0.739 (unshared), while cross-attention fusion reached 0.740 (shared) and 0.729 (unshared). Differences among fusion variants were within one standard deviation, and none reached statistical significance *relative* to the PA-only baseline. These results suggest that PA projections carry the dominant discriminative signal and that complementary anatomical information from orthogonal views may already be implicitly captured under limited-data conditions, rather than indicating methodological failure. Consistent with Appendix C, LA-only performance (0.695) trailed PA substantiating this claim.

We also assessed the effect of alternative training strategies on the best-performing fusion model (cross-attention/shared). Specifically, we evaluated curriculum learning, SimCLR-based self-supervised pretraining(Chen et al., 2020), and their combination.

Table 3 shows that curriculum learning yielded a higher mean AUC from 0.740 to 0.750. SimCLR pretraining yielded 0.742, close to the baseline. SimCLR combined with curriculum achieved the highest observed mean AUC (0.754), however this was comparable within

[1] † $p < 0.05$ vs CLAHE, Calc_foc $*$ $p < 0.05$ vs Original

Table 3: Final fusion model performance across training strategies. SimCLR and curriculum learning offer complementary benefits.

Training Strategy	Mean AUC
Fusion (standard)	0.740 ± 0.043
Curriculum Learning	0.750 ± 0.046
SimCLR Pretraining	0.742 ± 0.045
SimCLR + Curriculum	**0.754 ± 0.055**

variance. The consistent direction of improvement across folds suggests potential value, but larger datasets are required for confirmation (See Appendix C, Table 7 for pairwise Wilcoxon statistics). These results motivate examining how view complementarity and training stability interact under limited data, as discussed next

5. Discussion

Our experiments evaluated the feasibility of using DRRs as a proxy domain for CAC detection, focusing on: **(i)** fidelity enhancement via super-resolution, **(ii)** signal optimisation through preprocessing, and **(iii)** fusion of complementary projections.

5.1. Super-resolution and Preprocessing

Calcified lesions are high-frequency features and their visibility is diminished when CT volumes are reconstructed at low axial resolutions. On its own, SR modestly improved fidelity, yielding consistent but non-significant gains in CNN5 GAP ($+0.007; p = 0.107$). When paired with preprocessing, SR-enabled pipelines achieved improvements over their native baselines ($p < 0.05$), with Calc_foc + SR yielding the strongest single-view performance (0.735). This suggests that SR may play an enabling role: by partially restoring high-frequency structure, it could provide the substrate that preprocessing methods can then amplify into a more discriminative signal.

Preprocessing alone showed model-dependent behaviour. CNN5_GAP adapted well, whereas DenseNet121 degraded significantly ($p < 0.05$). This highlights a trade-off: pretrained backbones encode strong priors but are brittle to contrast shifts, while lightweight task-specific models are more flexible and can exploit enhanced contrast once fidelity is recovered. However, real CXRs are already acquired at high resolutions; SR's benefit is unique to CT-derived projections. In contrast, preprocessing strategies transfer rather trivially. In clinical CXRs, where acquisition noise and patient variability are greater, local contrast enhancement or normalisation may provide more value by homogenising inputs while highlighting subtle calcifications.

Taken together, these results suggest that fidelity restoration and contrast enhancement jointly determine the effective signal-to-noise ratio for CAC depiction in DRRs

5.2. Fusion & Training Strategies

Given calcifications may be obfuscated in a single projection, we evaluated fusion of PA and LA DRRs. Early image-level fusion performed significantly worse ($p < 0.05$), suggesting destructive interference across modalities. Intermediary feature fusion and cross-attention achieved the best observed gains, though improvements did not reach statistical significance, likely reflecting fold-level variance under limited data. Notably, LA-only classification underperformed (AUC 0.695, see Table 6), indicating that PA projections provide most discriminative information, while the marginal gain from incorporating LA views may be limited under current dataset size and coverage. Curriculum learning stabilised training (+0.010 AUC), SimCLR offered marginal gains, and their combination reached the highest mean AUC (0.754), consistent across folds but not statistically significant. Their complementary effects likely stem from stabilising optimisation (curriculum) and improving representation initialisation (SimCLR). Ultimately, these observations highlight the need for larger, more diverse datasets to fully exploit multi-view complementarity and self-supervised pretraining.

Overall, SR enhances DRR fidelity for lightweight CNNs, preprocessing benefits depend on architecture, and fusion trends are promising but data-limited. DRRs reproduce key behaviours seen in real CXR studies while providing label-rich CT-derived data, substantiating their value as a proxy domain and motivating future work in domain-adaptation.

5.3. Comparison to Prior Work

Our best-performing configuration (mean AUC = 0.754) is comparable to, and in some cases slightly exceeds, results from prior CXR-based CAC detection studies. (Kamel et al., 2021) reported an AUC of 0.73 using an attention-augmented VGG16 trained on 1,689 paired CXRs and (D'Ancona et al., 2023) achieved ≈0.71 AUC when predicting significant coronary artery disease from radiographs. In contrast to (Jeong et al., 2024) whose radiomics-based approach required manual cardiac segmentation, our pipeline is fully end-to-end and does not rely on handcrafted features or manual delineation. Additionally, our approach offers two further methodological advantages: **(i)** precise CT-derived labels free from annotation noise and **(ii)** scalability unconstrained by the scarcity of dual-modality data.

Our best-performing configuration (mean AUC = 0.754 ± 0.055) is comparable to, and in some cases slightly exceeds, previously reported results from real CXR–based CAC detection studies. (Kamel et al., 2021) achieved an AUC of 0.73 using an attention-augmented VGG16 trained on 1,689 paired CXRs and CTs, while (D'Ancona et al., 2023) reported approximately 0.71 for detecting significant coronary artery disease from radiographs. (Jeong et al., 2024) reached an AUC of 0.808 using a radiomics-based pipeline requiring manual cardiac segmentation.

In contrast, our pipeline is **fully automated and end-to-end**, requiring no manual segmentation or handcrafted features, and achieves comparable discriminative performance despite being trained on a smaller dataset of synthetic DRRs. Crucially, our approach inherits precise, noise-free labels directly from CT-derived Agatston scores, avoiding the temporal misalignment and label uncertainty that affect retrospective CXR–CT pairing studies.

Moreover, the proposed DRR-based framework is **intrinsically scalable and reproducible**. Once a labelled CT repository is available, thousands of perfectly aligned DRRs can be generated automatically under controlled imaging geometry. This removes the bottleneck of acquiring dual-modality CXR–CT pairs and establishes a standardized, label-rich training domain that can be shared and benchmarked across studies.

These characteristics enable our synthetic DRR pipeline to achieve near state-of-the-art AUC performance within a controlled and reproducible framework, offering a pragmatic foundation for developing and validating CAC detectors prior to transfer to real CXRs.

6. Limitations

Although results are promising, several limitations should be acknowledged. Since Agatston scores are aggregated across coronary arteries, they provide only image-level supervision. While appropriate for weakly-supervised feasibility, future work should explore region-localized DRR labels derived from per-artery calcium masks to better capture spatial heterogeneity. Further, the dataset remains small by deep-learning standards with validation folds containing about 130 DRRs, contributing to fold-level variance and modest effect sizes. More extensive cohorts and higher-fold cross-validation would clarify the validity of observed trends. Importantly, the synthetic domain may not fully replicate the characteristics of clinical CXRs. Differences in noise statistics, detector response, and cardiac field-of-view mean that models trained purely on DRRs may not transfer directly to real radiographs. Thus, the present study demonstrates methodological feasibility rather than clinical readiness. Future work should therefore prioritise domain-adaptation strategies and real-CXR validation to establish relevance.

7. Conclusion

This study demonstrates the feasibility of using CT-derived DRRs as a proxy training domain for DL–based CAC detection, offering a scalable alternative in the face of scarce labelled CXR datasets. We showed that lightweight architectures trained from scratch can outperform larger pretrained models, and that pairing SR with contrast-enhancing preprocessing yields statistically significant performance gains, supporting the value of fidelity recovery in DRRs. Fusion of orthogonal projections produced the best observed performance despite not reaching significance under data-limited evaluation. Structured training strategies such as curriculum learning improved robustness under label noise, with SimCLR pretraining offering only marginal benefit. Beyond these methodological insights, our findings underscore the broader potential of DRRs as a stable, label-rich surrogate for CXRs. By establishing a viable synthetic environment, this framework lowers the barrier for reproducible CAC detection research and could serve as a pretraining resource for future CXR-based screening pipelines. With further work in domain adaptation, dataset expansion, and multimodal integration, DRR-trained models could in principle form the basis of clinically deployable screening tools for population-level cardiovascular risk assessment.

Acknowledgments

The authors are grateful for the support by the 2024 (Cardiac, Vascular, and Metabolic Medicine) CVMM THEME COLLABORATIVE GRANT SCHEME.

References

Arthur S Agatston, William R Janowitz, Frank J Hildner, Norman R Zusmer, Michael Viamonte, and Robert Detrano. Quantification of coronary artery calcium using ultrafast computed tomography. *Journal of the American College of Cardiology*, 15(4):827–832, 1990.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

Matthew J. Budoff, Leslee J. Shaw, Sandy T. Liu, Steven R. Weinstein, Philip H. Tseng, Ferdinand R. Flores, Tracy Q. Callister, Paolo Raggi, Daniel S. Berman, and Tristen P. Mosler. Long-term prognosis associated with coronary calcification. *JACC*, 49(18):1860–1870, 2007. doi: 10.1016/j.jacc.2006.10.079. URL https://www.jacc.org/doi/abs/10.1016/j.jacc.2006.10.079.

Stanford AIMI Center. Coronary calcium and chest ct (coca). https://doi.org/10.71718/ge5g-ds80, 2022. Stanford Center for Artificial Intelligence in Medicine Imaging.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020. URL https://arxiv.org/abs/2002.05709.

Giuseppe D'Ancona, Matteo Massussi, Matteo Savardi, Alberto Signoroni, Ludovica Di Bacco, Diletta Farina, Marco Metra, Roberto Maroldi, Claudio Muneretto, Hüseyin Ince, Domenico Costabile, Marcello Murero, Giulia Chizzola, Stefano Curello, and Stefano Benussi. Deep learning to detect significant coronary artery disease from plain chest radiographs: Ai4cad. *International Journal of Cardiology*, 370:435–441, Jan 2023. doi: 10.1016/j.ijcard.2022.10.154.

David Eng, Christopher Chute, Nishith Khandwala, Jacob Cho, Nicholas Sheets, Andrew Buckler, Puneet Goel, Pranav Rajpurkar, Robyn L Ball, Irene Ha, et al. Automated coronary calcium scoring using deep learning with multicenter external validation. *NPJ Digital Medicine*, 4(1):88, 2021. doi: 10.1038/s41746-021-00442-9.

Vivek Gopalakrishnan and Polina Golland. Fast auto-differentiable digitally reconstructed radiographs for solving inverse problems in intraoperative imaging. In *Workshop on Clinical Image-Based Procedures*, pages 1–11. Springer, 2022.

Scott M. Grundy, Neil J. Stone, Alison L. Bailey, Craig Beam, Kim K. Birtcher, Roger S. Blumenthal, Lynne T. Braun, Sarah de Ferranti, Joseph Faiella-Tommasino, Daniel E.

Forman, Ronald Goldberg, Paul A. Heidenreich, Mark A. Hlatky, Daniel W. Jones, Donald Lloyd-Jones, Nuria Lopez-Pajares, Chiadi E. Ndumele, Carl E. Orringer, Carmen A. Peralta, Joseph J. Saseen, Sidney C. Smith, Laurence Sperling, Salim S. Virani, and Joseph Yeboah. 2018 aha/acc/aacvpr/aapa/abc/acpm/ada/ags/apha/aspc/nla/pcna guideline on the management of blood cholesterol: A report of the american college of cardiology/american heart association task force on clinical practice guidelines. *Circulation*, 139(25):e1082–e1143, 2019. doi: 10.1161/CIR.0000000000000625. URL https://www.ahajournals.org/doi/abs/10.1161/CIR.0000000000000625.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Myles Hollander, Douglas A. Wolfe, and Eric Chicken. *Nonparametric Statistical Methods*. John Wiley & Sons, Hoboken, NJ, 3rd edition, 2013. ISBN 978-0-470-38737-5.

Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016. URL http://arxiv.org/abs/1608.06993.

Hyeonjun Jeong, Hyun B Park, Jinhee Hong, Jihwan Lee, Seung-Pil Ha, Ran Heo, Jin-Young Jung, Young Jin Hong, and Hyuk-Jae Chang. Identifying coronary artery calcification using chest x-ray radiographs and machine learning: The role of the radiomics score. *Journal of Thoracic Imaging*, 39(2):119–126, Mar 2024. doi: 10.1097/RTI.0000000000000757.

Peter I Kamel, Peter H Yi, Haris I Sair, and Chia-Ti Lin. Prediction of coronary artery calcium and cardiovascular risk on chest radiographs using deep learning. *Radiology: Cardiothoracic Imaging*, 3(3):e200486, Jun 2021. doi: 10.1148/ryct.2021200486.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. *CoRR*, abs/1609.04802, 2016. URL http://arxiv.org/abs/1609.04802.

C. S. Moore, G. P. Liney, A. W. Beavis, and J. R. Saunderson. A method to produce and validate a digitally reconstructed radiograph-based computer simulation for optimisation of chest radiographs acquired with a computed radiography imaging system. *The British Journal of Radiology*, 84(1006):890–902, 2011. doi: 10.1259/bjr/30125639.

Rikke Søgaard, Axel Cosmus Pyndt Diederichsen, Lars M Rasmussen, Jess Lambrechtsen, Flemming H Steffensen, Lars Frost, Kenneth Egstrup, Grazina Urbonaviciene, Martin Busk, and Jes S Lindholt. Cost effectiveness of population screening vs. no screening for cardiovascular disease: the danish cardiovascular screening trial (dancavas) . *European Heart Journal*, 43(41):4392–4402, 08 2022. ISSN 0195-668X. doi: 10.1093/eurheartj/ehac488. URL https://doi.org/10.1093/eurheartj/ehac488.

Mathias Unberath, Jan-Nico Zaech, Sing Chun Lee, Bastian Bier, Javad Fotouhi, Mehran Armand, and Nassir Navab. Deepdrr – a catalyst for machine learning in fluoroscopy-guided procedures, 2018. URL https://arxiv.org/abs/1803.08606.

Appendix A. Statistical Testing Methods

For each paired comparison (same folds/seeds), we apply the Wilcoxon signed-rank test on per-run AUC values. Using 5 randomised seeds across a 5-fold CV, this yields $n = 25$ runs. We report the test statistic and p-value. Significant results are marked as bold at $p < 0.05$ with a $*$.

Appendix B. Gated Dataset

Table 4: Gated dataset – Wilcoxon signed-rank test within CNN5_GAP across super-resolution and preprocessing modes. Statistical significance at $p < 0.05$ is marked with *.

Comparison	Stat	p-value	Mean AUC A	Mean AUC B
Native: orig vs calc_foc	97.0	0.078	0.720	0.729
Native: orig vs clahe	104.0	0.120	0.720	0.728
Native: calc_foc vs clahe	151.0	0.771	0.729	0.728
SR: orig vs calc_foc	66.0	**8.07e-03***	0.727	0.735
SR: orig vs clahe	89.0	**0.048***	0.727	0.733
SR: calc_foc vs clahe	123.0	0.300	0.735	0.733
Native vs SR (orig)	129.0	0.381	0.720	0.727
Native vs SR (calc_foc)	102.0	0.107	0.729	0.735
Native vs SR (clahe)	91.0	0.055	0.728	0.733

Table 5: Gated dataset – Wilcoxon signed-rank test within DenseNet121 across super-resolution and preprocessing modes. Statistical significance at $p < 0.05$ is marked with *.

Comparison	Stat	p-value	Mean AUC A	Mean AUC B
Native: orig vs calc_foc	83.0	**0.032***	0.718	0.683
Native: orig vs clahe	81.0	**0.028***	0.718	0.688
Native: calc_foc vs clahe	122.0	0.276	0.683	0.688
SR: orig vs calc_foc	66.5	**0.010***	0.708	0.688
SR: orig vs clahe	80.0	**0.026***	0.708	0.695
SR: calc_foc vs clahe	128.0	0.367	0.688	0.695
Native vs SR (orig)	105.0	0.122	0.718	0.708
Native vs SR (calc_foc)	150.0	0.751	0.683	0.688
Native vs SR (clahe)	129.0	0.381	0.688	0.695

Appendix C. Fusion Analysis

C.1. Lateral-only Performance

Here we assess the strength of the calcium signal in Lateral (LA) projections. Significance was assessed with Wilcoxon signed-rank tests at $p < 0.05$.

Table 6: Wilcoxon signed-rank test results for LA-only performance across preprocessing modes using CNN5_GAP. Statistical significance at $p < 0.05$ is marked with *.

Comparison	Stat	p-value	Mean AUC A	Mean AUC B
LA Native: orig vs calc_foc	37.0	**3.29e-04***	0.672	0.695
LA Native: orig vs clahe	57.0	**0.003***	0.672	0.691
LA Native: calc_foc vs clahe	120.5	0.258	0.695	0.691

C.2. Fusion and Training Strategy Comparisons

Pairwise Wilcoxon signed-rank test results comparing fusion baselines, curriculum learning, and SimCLR variants. Significant results at $p < 0.05$ are marked with $*$.

Table 7: Pairwise Wilcoxon signed-rank test results comparing fusion baselines, curriculum learning, and SimCLR variants. Statistical significance at $p < 0.05$ is marked with *.

Comparison	Stat	p-value	Mean AUC A	Mean AUC B
Early vs Int. Fusion (Shared)	11.0	**4.6e-05***	0.703	0.736
Early vs Int. Fusion (Unshared)	11.0	**4.6e-05***	0.703	0.740
Early vs Cross-Attention (Shared)	154.0	0.833	0.739	0.740
Standard vs Curriculum	132.0	0.426	0.740	0.750
Standard vs SimCLR	152.0	0.791	0.740	0.742
Standard vs Curriculum + SimCLR	128.0	0.367	0.740	0.754