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

Dylan Saeed

Supervisor: Associate Professor Susann Beier
Co-supervisor: Dr. Sonit Singh

School of Computer Science and Engineering

University of New South Wales

Term 2, 2025

Abstract

We present a deep learning framework for detecting coronary artery calcification (CAC) from synthetically generated chest X-rays, or digitally reconstructed radiographs (DRRs), derived from gated cardiac and non-gated chest CT scans. Our pipeline integrates super-resolution (SR) enhancement of volumetric CT data prior to DRR generation, followed by tailored image-enhancement preprocessing to improve feature visibility. We evaluate multiple convolutional architectures, introducing a dual-view feature fusion model that jointly leverages posteroanterior (PA) and lateral (LA) projections, as well as curriculum learning and SimCLR-style contrastive pretraining. Across 5 seeds $\times$ 5 folds ($n = 25$), SR-enhanced DRRs consistently achieved higher mean AUCs than native-resolution counterparts, and fusion models frequently outperformed single-view baselines, with the best configuration achieving a mean AUC of 0.754. While some differences did not reach statistical significance, trends were consistently positive across experimental families. These results support the feasibility of SR-enhanced DRRs, multi-view fusion, and curriculum-driven training as complementary strategies for CAC detection in settings where paired chest X-rays and calcium scores are scarce, and highlight the potential of DRRs as a scalable proxy training modality to improve deep learning–based cardiovascular risk assessment.

Contents

1 Introduction **1**

2 Background **3**
2.1 Coronary Artery Disease . . . 3
2.2 CT Imaging for CAC scoring . . . 3
2.3 Convolutional Neural Networks . . . 4
2.4 Attention Mechanisms and Interpretability . . . 5
2.5 Curriculum Learning . . . 5
2.6 Contrastive Learning . . . 6

3 Literature Review **7**
3.1 Automated CT-Based CAC Scoring . . . 7
3.2 Deep Learning-Based CAC Detection from Chest X-rays . . . 9
3.3 Digitally Reconstructed Radiographs . . . 12
3.4 Super Resolution for Medical Imaging . . . 14

4 Methodology **15**
4.1 COCA Dataset . . . 15
4.2 DRR Generation from CT Scans . . . 17
4.3 Preprocessing to Enhance DRR Fidelity . . . 17
4.4 Model Architectures . . . 19
4.5 Evaluation . . . 20

5 Experiments **22**
5.1 Training Configuration . . . 22
5.2 Analysis of Nongated Dataset . . . 22
5.3 Gated Dataset Experiments . . . 23
5.3.1 Architectural Baselines on Native DRRs . . . 23
5.3.2 Effect of Super-Resolution (SR) . . . 24
5.3.3 Impact of Preprocessing Modes . . . 25
5.3.4 Projection Fusion . . . 26
5.3.5 Curriculum Learning . . . 27
5.3.6 SimCLR Pretraining . . . 27
5.4 Summary of Experimental Findings . . . 28

A Statisitcal Testing Methods **33**
A.1 Nongated Dataset . . . 33
A.2 Gated Dataset . . . 34

A.3 Fusion Analysis 35
A.3.1 Lateral-only Performance 35
A.3.2 Fusion and Training Strategy Comparisons 35

Bibliography **36**

List of Figures

2.1 Core Convolutional Neural Network Architecture 4

3.1 U-Net architecture used for automatic CAC scoring from non-contrast CT scans[1] 8
3.2 Calcification on a patient with a score of 6056 (highly severe) with faint features in PA and lateral CXR 9
3.3 Grad-CAM attention heatmaps revealing model activation around the cardiac silhouette 10
3.4 Visualisation of RS with the original image (left), the RS map (middle) and overlapped images (right). 11
3.5 Overview of DRR generation 12
3.6 Digitally Reconstructed Radiograph (left) of a real patient and computed radiography image (right) of a real patient 13

4.1 Examples of generated DRRs. (a) Accepted projection. (b–c) Excluded due to FOV or radiographic obfuscation. 15
4.2 Distribution of Agatston scores in the nongated subset. 16
4.3 Distribution of Agatston scores in the gated subset. 16
4.4 Example DRRs generated from nongated (left) and gated (right) CT volumes. 17
4.5 Comparison of preprocessing variants applied to Native DRR (Top) and Super-Resolved DRR (bottom): (a) Original, (b) CLAHE-enhanced, and (c) Calcium-focused enhancement. 18
4.6 CNN5_GAP architecture with 5 Convolutional blocks, batch normalisation and max pooling followed by a global adaptive pooling layer and a classifier head. 19
4.7 Example receiver operating characteristic (ROC) curve for a single fold and seed of the CNN5_GAP on PA DRRs. The area under the curve (AUC) for this example is 0.782. 20

List of Tables

23table.caption.25
5.2 Validation AUCs for baseline architectures on original, native-resolution DRRs in the gated dataset. 24
5.3 Effect of super-resolution on gated dataset native-resolution DRRs. Results shown as mean AUC ± standard deviation over 25 runs. 24
5.4 Effect of preprocessing mode on native-resolution DRRs. 25
26table.caption.29
5.6 Effect of curriculum learning on cross-attention fusion model. Curriculum learning improves mean AUC and stabilizes performance across seeds. . . 27
5.7 Final fusion model performance across training strategies. SimCLR and curriculum learning offer complementary benefits, with the combined approach achieving the best result. 28

A.1 Nongated dataset – Wilcoxon signed-rank test results within DenseNet model. q-values are Benjamini–Hochberg adjusted across all comparisons in this table; significant results at $q < 0.05$ are in bold. 33
A.2 Nongated CNN5_GAP Wilcoxon signed-rank test results. Bold indicates $q < 0.05$ after Benjamini–Hochberg correction. 34
A.3 Nongated inter-model (CNN5_GAP vs DenseNet) Wilcoxon signed-rank test results. 34
A.4 Gated dataset – Statistical significance testing within CNN5_GAP. Wilcoxon signed-rank test with Benjamini–Hochberg FDR correction applied within this family of comparisons. 34
A.5 Gated dataset – Statistical significance testing within DenseNet. Wilcoxon signed-rank test with Benjamini–Hochberg FDR correction applied within this family of comparisons. 35
A.6 Wilcoxon signed-rank test results for LA-only performance across preprocessing modes using CNN5_GAP. Statistically significant results at $p < 0.05$ are marked with *. 35
A.7 Pairwise Wilcoxon signed-rank test results comparing fusion baselines, curriculum learning, and SimCLR variants. 35

Chapter 1

Introduction

Within the domain of cardiovascular imaging, the quantification of coronary artery calcification (CAC) remains vital for the early detection and risk stratification of coronary artery disease (CAD). The extent of calcification provides a direct, quantifiable marker of subclinical coronary artery disease and has been shown to correlate strongly with the risk of major adverse cardiovascular events[2]. As such, CAC scoring is routinely employed in clinical settings to guide preventative interventions and long-term management strategies.

The most widely adopted method for CAC quantification is the Agatston score, computed from non-contrast computed tomography (CT) scans by measuring the area and intensity of calcified lesions. CT-based CAC scoring is clinically validated, reproducible, and remains the gold standard for coronary calcium assessment[3]. However, routine CT screening is hindered by monetary costs, high radiation exposure[4]—necessitating a careful balance between diagnostic benefit and radiation risk—and limited accessibility——particularly in primary care and low-resource settings. These constraints pose significant barriers to the adoption of CAC screening at population scale.

Chest X-rays, in contrast, are inexpensive, widely available, and expose patients to minimal radiation—representing a promising alternative modality However, the faint radiographic signatures of coronary calcium deposits are often occluded by overlying anatomical structures, rendering manual interpretation challenging——even for experienced radiologists[5]. While deep learning models have demonstrated promise in detecting the presence of CAC from CXRs, their development is fundamentally constrained by the acquisition of data. Acquiring CXR-CAC pairs typically requires identifying patients who underwent both a chest X-ray and a corresponding CT scan within a clinically relevant timeframe. This dependency on dual-modality acquisition severely limits dataset size, thereby inhibiting the training of robust, generalisable models.

One potential remedy lies in the synthetic generation of chest X-rays from existing CT volumes. Known as digitally reconstructed radiographs (DRRs), these projections emulate the geometry and attenuation characteristics of standard posteroanterior chest radiographs, while preserving the calcium scoring ground truth intrinsic to the CT scan. Prior work has demonstrated substantial visual and anatomical concordance between DRRs and true radiographs, with expert radiologists deeming them clinically acceptable[6]. This motivates the central inquiry of this research: *Can digitally reconstructed radiographs (DRRs), synthetically generated from CT scans annotated with Agatston-based CAC scores,*

serve as a viable proxy training modality for deep learning models tasked with CAC detection?

We hypothesise that DRRs—particularly when enhanced through super-resolution (SR) techniques—can bridge the gap between the structured, high-fidelity domain of CT and the low-cost, low-resolution nature of CXR. To evaluate this, we design an experimental framework that:

1. Generates DRRs from both full-field (nongated) and cardiac-focused (gated) CT datasets.
2. Applies SR and image preprocessing to enhance DRR fidelity.
3. Benchmarks multiple deep learning architectures and fusion strategies for CAC detection.
4. Investigates training-time strategies such as curriculum learning and contrastive pretraining to improve generalisation in low-data regimes.

If successful, this approach could provide a scalable pathway for training CXR-based CAC detection models without requiring large paired datasets, enabling population-level cardiovascular risk screening with minimal radiation exposure and infrastructure requirements.

Chapter 2

Background

This chapter outlines the foundational concepts necessary to contextualize the methods explored in this thesis. We begin by introducing the clinical motivation for coronary artery calcium (CAC) detection, followed by an overview of the imaging modality used for CAC quantification—non-contrast computed tomography (CT). We then present key machine learning concepts employed throughout this work, including convolutional neural networks (CNNs), attention mechanisms, curriculum learning, and contrastive pretraining. These components form the methodological backbone of the proposed pipeline and are referenced extensively in subsequent chapters.

2.1 Coronary Artery Disease

Coronary artery disease (CAD) remains the leading cause of mortality worldwide, accounting for a significant proportion of cardiovascular deaths each year. It is primarily driven by the progressive accumulation of atherosclerotic plaque within the coronary arteries, leading to luminal narrowing and impaired myocardial perfusion. While clinical manifestations such as angina and myocardial infarction typically occur in advanced stages, subclinical disease can remain asymptomatic for years, highlighting the importance of early detection.

Coronary artery calcification is a marker of atherosclerosis and serves as a direct, quantifiable indicator of plaque burden. As opposed to traditional risk factors—such as cholesterol and blood pressure—CAC reflects cumulative vascular injury and correlates strongly with future cardiac events. Its presence has been shown to improve risk stratification beyond standard clinical models, particularly in asymptomatic individuals.

2.2 CT Imaging for CAC scoring

Non-contrast cardiac computed tomography is the current clinical gold standard for assessing coronary artery calcification. As a result of its high spatial resolution and calcium sensitivity, CT enables precise localisation and quantification of calcified atherosclerotic plaque within the coronary arteries. During acquisition, axial slices spanning the heart are captured, typically with submillimetre resolution. These images are reconstructed into volumetric data from which calcific lesions are segmented and quantified, typically by a radiologist with specialsied imaging software.

There are two primary types of cardiac CT: gated and nongated scans. Gated CTs are synchronised with the cardiac cycle via electrocardiogram (ECG) triggering, which reduces motion artifacts and provides clearer delineation of coronary structures. This makes them the preferred modality for clinical calcium scoring. Nongated CTs, by contrast, are not ECG-synchronised and are often acquired for other diagnostic purposes (e.g., lung screening). While they offer a broader field of view and are more widely available, their utility for CAC assessment is limited by motion blur and reduced cardiac specificity.

The Agatston score [3] is the most widely adopted method for quantifying CAC. It is computed by identifying regions of interest (ROIs) above a threshold of 130 Hounsfield Units (HU), multiplying their area by a weighting factor that increases with peak attenuation (e.g., 130–199 HU $\rightarrow$ weight 1, 200–299 HU $\rightarrow$ weight 2, etc.), and summing across all coronary arteries. The total Agatston score is then used to classify CAC severity into ordinal categories: 0 (none), 1–99 (mild), 100–399 (moderate), and $\geq$400 (severe). Clinically, a score above 100 is associated with a substantially increased risk of adverse cardiovascular outcomes, and is often used as a threshold for intervention.

However, despite its clinical utility, CT-based calcium scoring is underutilized in general populations due to factors such as radiation dose, cost, and the requirement of specialized equipment and expertise. This has motivated interest in developing alternative, more accessible imaging-based tools for CAC assessment.

2.3 Convolutional Neural Networks

Convolutional neural networks (CNNs) are a class of deep learning architectures particularly well-suited for visual data. Originally inspired by the organisation of the visual cortex, CNNs have become foundational to modern computer vision due to their ability to learn spatial hierarchies of features directly from raw image inputs. A typical CNN consists of a sequence of convolutional layers followed by nonlinear activations and pooling operations as shown in Figure 2.1. These layers progressively extract local features such as edges, textures, and higher-order patterns, culminating in a compressed latent representation suitable for classification, detection, or segmentation.

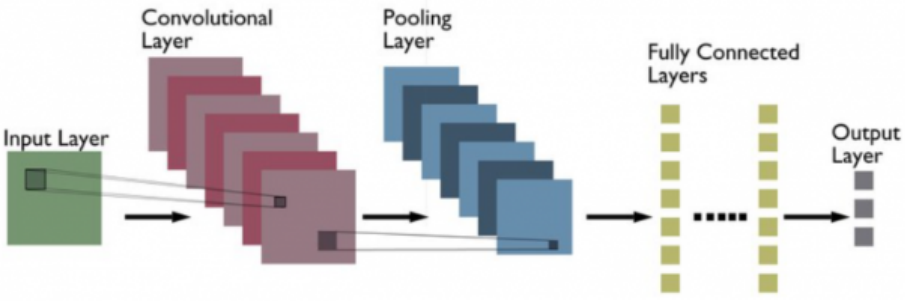


Figure 2.1: Core Convolutional Neural Network Architecture

Unlike fully connected networks, CNNs exploit spatial locality through shared weights (convolutional kernels), dramatically reducing the number of parameters and enabling translation invariance. This inductive bias has made CNNs particularly successful in medical imaging tasks, where anatomical structures may vary in position but retain consistent morphological patterns. Standard architectures such as U-Nets, VGG16,

ResNet, and DenseNet differ primarily in their use of depth and advanced features such as skip connections and dense connectivity, but all share the fundamental convolutional structure.

In this work, all classification models—custom and pretrained—are built upon CNN backbones due to their proven effectiveness in extracting salient features from radiographic images.

2.4 Attention Mechanisms and Interpretability

It is rarely the case that an input is entirely free of redundancy, or that every part of it is equally relevant to solving a task. For instance, when presented with an image of an animal and asked to classify it, we naturally focus on the animal itself rather than on irrelevant background details. Attention mechanisms formalise this intuition by enabling models to dynamically focus on the most informative regions of an input when making predictions. Originally introduced in the context of sequence-to-sequence models for machine translation[7], attention has since been widely adopted in computer vision, where it helps mitigate the limitations of fixed receptive fields in convolutional architectures.

At its core, attention computes a weighted combination of features, where the weights reflect the importance of each element relative to a query. In vision tasks, this allows the model to selectively amplify spatial regions or channels that are most informative for the task at hand. For example, in chest radiographs, attention mechanisms can prioritise regions likely to contain pathological signals—such as calcifications—while suppressing irrelevant anatomical noise.

2.5 Curriculum Learning

Curriculum learning is a training paradigm inspired by the way humans acquire knowledge—starting from simpler concepts and gradually progressing to more complex ones. In contrast to traditional deep learning workflows where models are exposed to randomly ordered training data, curriculum learning proposes that performance and generalization can improve when the model is first trained on "easier" examples before encountering harder, more ambiguous cases.

The core idea, introduced by Bengio et al.[8], is that the optimisation landscape may become smoother and more navigable when the model's capacity is built progressively, akin to educational curricula in human learning. This strategy has been particularly effective in scenarios involving noisy labels, class imbalance, or ambiguous data points, where presenting all samples uniformly can overwhelm the model early in training. In medical imaging, curriculum learning has shown promise in guiding models to first internalise clearer patterns—such as obvious pathology—before attempting borderline or subtle cases.

In the context of CAC detection, where inter-rater variability and anatomical overlap can introduce ambiguity near decision boundaries (e.g. around Agatston score thresholds), curriculum learning serves as a regularization strategy that aligns model focus with clinical intuition.

2.6 Contrastive Learning

SimCLR is a self-supervised learning framework that enables the training of deep neural networks without manual labels by leveraging contrastive objectives. The central idea is to learn an embedding space where representations of different augmented views of the same input (positive pairs) are pulled together, while representations of different inputs (negative pairs) are pushed apart[9]. This forces the model to capture semantically meaningful features invariant to common perturbations.

In SimCLR, each image undergoes two independent stochastic augmentations (e.g., crop, rotation, color jitter), and the network is trained to minimise the following normalized temperature-scaled cross-entropy (NT-Xent) loss:

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{z_i}, \mathbf{z_j})/\tau)}{\sum_{k=1}^{2N} 1_{[k \neq i]} \exp(\text{sim}(\mathbf{z_i}, \mathbf{z_k})/\tau)} \tag{2.1}$$

Where $\mathbf{z_i}, \mathbf{z_j}$ are normalised embeddings of a positive pair (two augmentations of the same image), $\text{sim}(\mathbf{z_i}, \mathbf{z_j})$ is the cosine similarity between vectors, τ is a temperature parameter and N is the batch size. A projection head is then used during training to improve contrastive alignment, but discarded during downstream fine-tuning.

This approach has been particularly effective in domains where labeled data is scarce or expensive to obtain, including medical imaging. By pretraining an encoder on large volumes of unlabeled images, the learned features can serve as a strong initialisation for supervised fine-tuning, often outperforming random or ImageNet-based initialisation.

In the context of CAC detection from DRRs, SimCLR enables the model to benefit from large sets of synthetic radiographs—even without explicit labels—by learning structural regularities in anatomical and calcification patterns prior to downstream classification.

Chapter 3

Literature Review

This chapter assesses existing literature in the context of our guiding question: Can DRRs serve as a viable proxy training modality for CAC detection? The literature review will comprise four sections, each corresponding to a foundational pillar of the proposed research. First, an examination of clinical approaches to CAC scoring, with a focus on CT-based quantification methods. Second, a review of existing applications of deep learning for CAC detection from chest radiographs. Third, the generation and validation of DRRs as a means of synthetic data augmentation. Finally, an assessment of super-resolution techniques within the domain of medical imaging for the purpose of enhancing DRR fidelity for downstream ML.

3.1 Automated CT-Based CAC Scoring

As previously discussed, non-contrast computed tomography (CT) remains the clinical gold standard for quantifying coronary artery calcium (CAC) burden. As the most widely adopted and validated method in current cardiovascular practice, it serves both as a diagnostic benchmark and a reference modality against which alternative approaches—particularly those leveraging more accessible imaging modalities—are evaluated.. Understanding the evolution and capabilities of automated CAC scoring in CT imaging provides essential context for the motivation and feasibility of this project.

Traditionally, the quantification of coronary artery calcium relied on manual interpretation by trained clinicians operating within accredited 3D imaging laboratories. This process, while accurate, incurs significant personnel overhead and remains inherently time-consuming—particularly in large-scale screening scenarios. In such contexts, the manual identification and annotation of calcified lesions across coronary arteries constitutes a substantial bottleneck to clinical throughput, often resulting in underreporting of coronary calcification in routine CT exams not specifically ordered for CAC evaluation.

To address this limitation, Eng et al.[1] propose a deep learning framework for fully automated CAC scoring from non-contrast CT. Their method employs a convolutional neural network (CNN)-based pipeline (Figure 3.1) that ingests axial CT slices and segments calcified lesions contributing to the Agatston score. Each lesion is subsequently classified as belonging to one of the four major coronary arteries: left main (LCA), left anterior descending (LAD), left circumflex (LCX), or right coronary artery (RCA). Connected

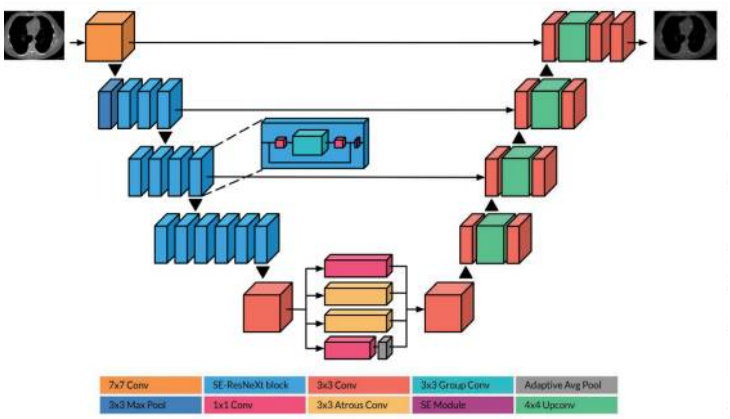


Figure 3.1: U-Net architecture used for automatic CAC scoring from non-contrast CT scans[1]

component analysis is then used to calculate lesion area and maximum Hounsfield Unit (HU) intensity. Slice-wise Agatston scores are aggregated to yield a total calcium score for the scan. This fully automated, end-to-end pipeline entirely eliminates the need for manual review or human interpretation.

The model was trained on over 6,000 CT scans and externally validated across multiple clinical sites. It demonstrated strong concordance with expert annotations, achieving a mean absolute error of –2.86 and positive predictive values ranging from 80–100% across all coronary arteries. These results suggest that deep learning models can replicate——and in some cases surpass—the consistency of manual scoring while significantly reducing clinician workload.

Notably, the high accuracy and full automation of this pipeline enable the large-scale acquisition of paired CT volumes and CAC scores, facilitating population-level studies without the burden of manual annotation. However, its reliance on CT imaging—a costly modality with non-trivial radiation exposure—limits its feasibility for routine screening, especially in low-resource or primary care settings. This motivates interest in exploring automated CAC scoring pipelines using alternative, low-cost imaging modalities that are more scalable and accessible for population health applications.

Notably, the high accuracy and full automation of this pipeline enable the large-scale acquisition of paired CT volumes and CAC scores, facilitating population-level studies without the burden of manual annotation. However, the reliance on CT imaging——an expensive modality associated with non-trivial radiation exposure——limits its feasibility in routine screening, especially in low-resource or primary care settings. This motivates interest in exploring automated CAC scoring pipelines using alternative, low-cost imaging modalities that are more scalable and accessible.

3.2 Deep Learning-Based CAC Detection from Chest X-rays

Chest X-rays (CXRs) represent an attractive alternative to CT imaging for coronary artery calcium (CAC) detection due to their wide availability, low cost, and minimal radiation exposure. However, the projectional nature of CXRs inherently complicates the task: calcifications are often faint, small, and obscured by overlapping anatomical structures such as the heart and spine as illustrated in Figure 3.2, where calcifications remain faint even in highly severe cases. As a result, manual identification of CAC on CXR is challenging—even for expert radiologists. Recent studies have sought to overcome these limitations using deep learning (DL) models trained to detect CAC from frontal chest radiographs.

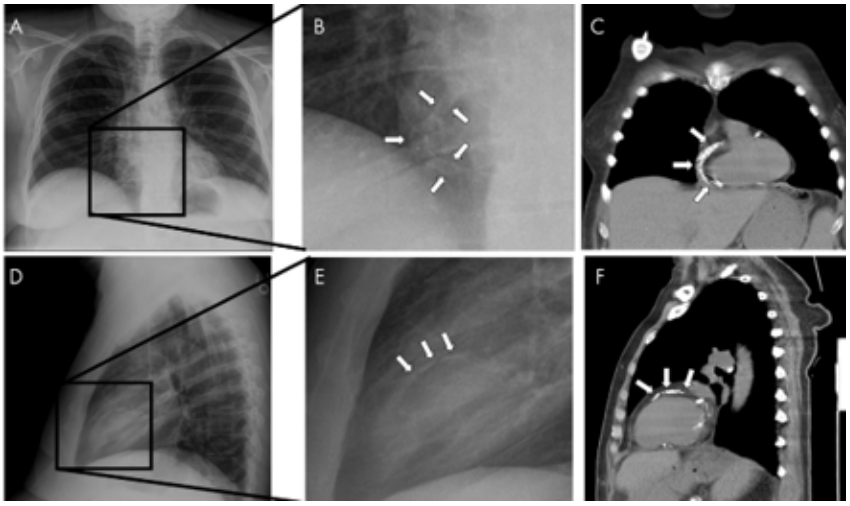


Figure 3.2: Calcification on a patient with a score of 6056 (highly severe) with faint features in PA and lateral CXR

Kamel et al.[5] present one of the earliest and most influential approaches in this space. The authors employ a VGG16 backbone, pretrained on ImageNet, and augment it with an attention mechanism to detect the presence of CAC from paired CXR and CT studies. Given the localised nature of CAC, the introduction of attention serves to enhance interpretability by guiding the model's focus toward anatomically relevant regions—simultaneously allowing for visual validation of its learned representations. Using a dataset of 1,689 image pairs, the model was trained to perform binary classification (CAC present vs. absent at different thresholds), achieving an area under the ROC curve (AUC) of 0.73 on frontal radiographs. Attention heatmaps were used to visualise model focus, revealing regions of activation around the cardiac silhouette as seen in Figure 3.3.

Interestingly, the model occasionally placed disproportionate importance on implanted cardiac devices such as defibrillators rather than directly visualising CAC—raising the possibility that it may have associated such artifacts with elevated risk. However, due to the black-box nature of deep learning, it remains difficult to ascertain whether the model inferred clinical risk contextually or simply relied on confounding visual features.

While the study offers a compelling proof-of-concept, several limitations constrain its broader applicability. First, the use of a VGG16 architecture pretrained on natural RGB

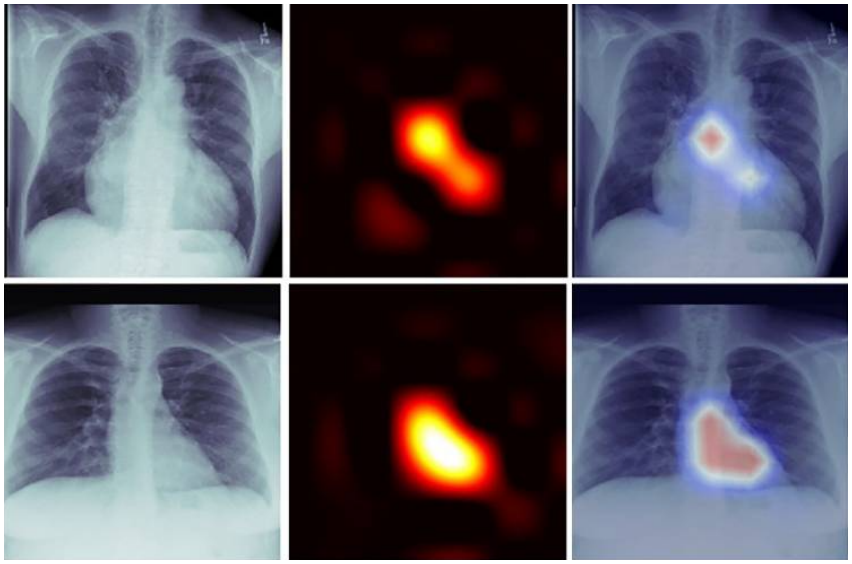


Figure 3.3: Grad-CAM attention heatmaps revealing model activation around the cardiac silhouette

images may not be optimal for grayscale medical images, potentially limiting feature relevance and domain adaptation. Second, the dataset was relatively small by deep learning standards and lacked external validation——raising concerns about model generalisability. This stems from the broader challenge of acquiring temporally aligned CXR–CT pairs with corresponding calcium scores, a process that both limits dataset scalability and introduces potential label noise.

D'Ancona et al.[10] extended the use of deep learning on chest radiographs from detecting coronary artery calcification (CAC) to predicting significant coronary artery disease (CAD), using invasive coronary angiography as the ground truth. Their model, based on a ResNet50 architecture, was trained to classify the presence or absence of severe CAD across a cohort of over 7,700 patients. Importantly, their use of pretraining on large chest X-ray datasets——rather than ImageNet——likely contributed to improved domain-specific feature extraction and convergence stability. This addressed a key limitation of Kamel et al, and the improved alignment between pretraining domain and target modality may also have enabled the model to leverage more diffuse radiographic patterns across the thoracic anatomy, beyond the calcific cues emphasised in prior work.

Notably, rather than relying on visual calcification cues, the model learned distributed radiographic features—including extra-cardiac regions—suggesting that chest radiographs may encode broader systemic indicators of cardiovascular disease. The model achieved an AUC of 0.73, which increased to 0.77 when combined with clinical angina status, outperforming the traditional Diamond-Forrester score. Grad-CAM visualisations showed model attention across diverse anatomical structures, and rigorous exclusion of confounders (e.g., defibrillators, prior interventions) aimed to ensure unbiased learning. While this study does not predict CAC directly, it demonstrates the viability of CXR-based AI tools in cardiovascular risk stratification. The authors caution that due to referral bias and high disease prevalence in the cohort, the model cannot yet be generalised to asymptomatic populations or screening contexts without further external validation.

Jeong et al[11] explored a complementary approach to deep learning by employing a

radiomics-based pipeline for CAC detection from chest radiographs. Rather than using end-to-end convolutional neural networks, they extracted 455 handcrafted radiomic features from manually segmented cardiac contours on PA chest X-rays and used LASSO regression to derive a radiomics score (RS). Figure 3.4 illustrates a visualisation of the feature maps generated from an input CXR.

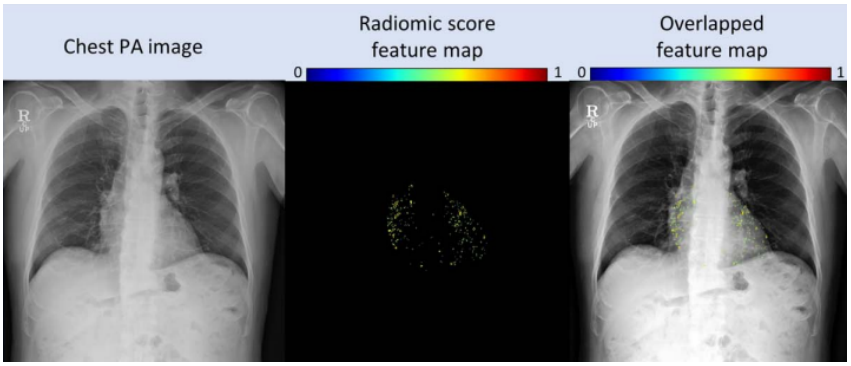


Figure 3.4: Visualisation of RS with the original image (left), the RS map (middle) and overlapped images (right).

This RS was then combined with basic clinical factors (age, sex, BMI) to train a random forest classifier to identify patients with moderate-to-severe CAC (Agatston score ≥ 100). Their model achieved an AUC of 0.808 in the validation cohort—outperforming both clinical-factor-only (AUC 0.633) and RS-only (AUC 0.748) models—and demonstrated balanced sensitivity (0.85) and specificity (0.82). Notably, the RS heatmaps revealed focused activation around the heart and exclusion of irrelevant anatomical structures, suggesting that the radiomics features captured spatially coherent markers of CAC.

This work builds on Kamel et al. by avoiding end-to-end black-box models and enhancing interpretability through engineered features. The inclusion of clinical variables and emphasis on explainable feature attribution align with the broader clinical push toward transparent AI systems. However, several limitations remain. The need for manual cardiac contouring requires expert domain knowledge, limits scalability, and introduces subjectivity that could impact reproducibility. Furthermore, their model was trained on a relatively small cohort (n=559) with no external validation. Additionally, the model excluded images containing common confounders such as ECG leads or devices——raising questions about performance in more typical clinical settings. Despite these caveats, the study reinforces the premise that chest radiographs contain meaningful, quantifiable signals related to CAC burden, even without deep learning——a finding that supports the broader goal of developing accessible, interpretable tools for cardiovascular risk stratification from 2D projections.

While these studies collectively demonstrate that chest radiographs contain meaningful signal for CAC detection, they remain constrained by a fundamental bottleneck: the availability of large, high-quality datasets containing paired CXRs and reliable calcium scores. Label scarcity, annotation costs, and cohort biases limit the development of robust, generalisable models. To overcome this, recent work has explored the use of synthetically generated chest radiographs—digitally reconstructed from CT volumes—as a scalable alternative for model training and evaluation. We explore this next.

3.3 Digitally Reconstructed Radiographs

While deep learning and radiomics models have demonstrated the feasibility of CAC detection from chest radiographs, their broader applicability consistently remains constrained by the limited availability of labelled data—particularly paired CXRs and calcium scores. As previously mentioned, acquiring such datasets typically requires identifying patients who have undergone both a chest X-ray and a CT scan within a clinically relevant timeframe, a co-occurrence that is rare in routine imaging workflows. Furthermore, even when available, these datasets often suffer from temporal misalignment, label noise, and heterogeneity in imaging protocols. To address this bottleneck, recent efforts have turned towards repurposing existing CT datasets through synthetic generation of digitally reconstructed radiographs (DRRs).

Digitally reconstructed radiographs (DRRs) are generated by casting rays from a virtual X-ray source through a CT volume to a simulated detector plane. As each ray traverses the volume, it accumulates voxel-wise attenuation values based on the length of intersection with each voxel—a process typically implemented using Siddon's algorithm, a classic ray-tracing method from medical physics. The result is a synthetic projection image that approximates the attenuation profile of a real radiograph, capturing both anatomical structure and radiometric contrast. Figure 3.5 illustrates this process.

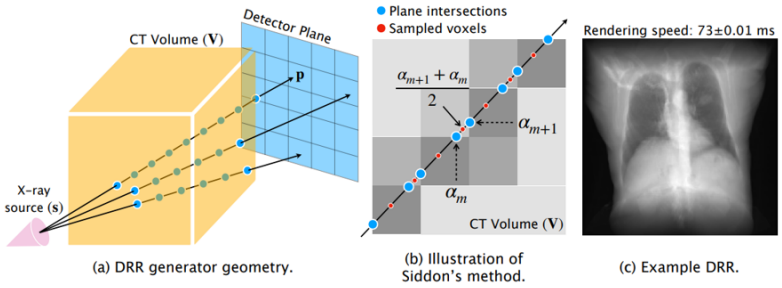


Figure 3.5: Overview of DRR generation

In doing so, DRRs provide a controllable, anatomically accurate, and label-rich alternative to real CXRs for training machine learning models.

One of the earliest and most influential contributions to DRR simulation in the context of chest radiography is Moore et al.[6], who developed a physically grounded, ray-casting-based model for generating DRRs from patient CT data. Their simulator aimed to optimise image quality and dose in computed radiography (CR) systems by producing anatomically realistic projections validated against both phantom and patient data. The pipeline incorporated corrections for scatter, frequency-dependent quantum noise modeling, and conversion of CT values to linear attenuation coefficients via Gammex phantom calibration.

As shown in Figure 3.6, the resulting DRRs demonstrated close agreement with real radiographs. Expert radiologists assessed the images as clinically acceptable, and quantitative metrics such as signal-to-noise ratio (SNR) and tissue-to-rib contrast ratio (TRR) reflected high radiometric fidelity. While the simulator was originally developed to optimise chest CR acquisition parameters, this work validated the anatomical plausibility of DRRs and laid a foundation for their use in downstream machine learning applications.

Crucially, the anatomical and radiometric realism of DRRs makes them a compelling proxy

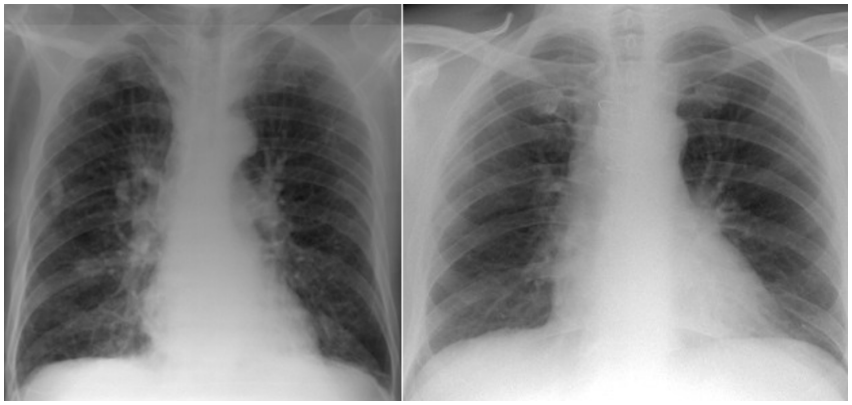


Figure 3.6: Digitally Reconstructed Radiograph (left) of a real patient and computed radiography image (right) of a real patient

for real CXRs in ML pipelines. In the context of CAC detection, this represents a paradigm shift: instead of requiring paired CXR–CT datasets—an inherently scarce and logistically burdensome resource—one can synthetically generate CXR-like projections directly from annotated CT volumes. This enables scalable model training using existing CAC-labelled CT data, entirely circumventing the need for dual-modality image acquisition. However, this approach is not without limitations. The fidelity of DRRs is fundamentally bounded by the spatial resolution of the source CT, and efforts to achieve realism—such as more accurate scatter modeling—incur significant computational cost and can result in DRR generations taking upwards of 1 hour. These trade-offs must be carefully considered when designing DRR pipelines for learning tasks.

To address the computational limitations of traditional DRR generation, Gopalakrishnan and Golland introduced DiffDRR, an open-source, GPU-accelerated DRR generation library built in PyTorch[12]. The core contribution of DiffDRR is its vectorised reformulation of Siddon's algorithm as a series of vectorised tensor operations. This design allows the forward projection process to be executed efficiently on modern GPUs, reducing the generation time of 512×512 DRRs to under 100 ms, compared to several minutes or even hours for non-optimised CPU-based implementations. While the method was initially developed to support real-time slice-to-volume registration in interventional settings, its ease of installation, open-source availability, and compatibility with PyTorch's autograd system make it readily deployable within modern ML workflows. For CAC detection and other DRR-based classification tasks, DiffDRR provides a highly accessible tool to synthetically generate large-scale, high-fidelity projection datasets from annotated CT volumes-removing both the computational barrier and engineering overhead associated with traditional ray-tracing pipelines.

Together, these works demonstrate both the viability and practicality of DRRs as a proxy for chest radiographs in ML applications. The former established the potential for DRRs to circumvent the need for acquiring physical CXRs entirely when paired with annotated CT data. The latter then operationalised this insight by providing an efficient implementation capable of generating DRRs at scale. However, the resolution of the resulting DRR is ultimately bounded by the spatial fidelity of the input CT scan. This limitation has motivated growing interest in enhancing CT resolution through super-resolution techniques prior to forward projection, enabling the synthesis of higher-quality DRRs for training downstream models.

3.4 Super Resolution for Medical Imaging

Despite the anatomical fidelity of digitally reconstructed radiographs (DRRs), their visual quality and texture realism are inherently limited by the resolution of the source CT scans. Clinical multidetector CT (MDCT) protocols often prioritize dose efficiency and acquisition speed, resulting in coarse voxel grids and reduced high-frequency detail—particularly in low-dose or rapid-acquisition scenarios. This resolution bottleneck propagates into the DRR domain, where low-fidelity inputs lead to smoothed projections that may obscure subtle radiographic cues relevant for tasks such as CAC detection.

While a number of super-resolution (SR) techniques have been proposed in the medical imaging literature, most require access to paired low- and high-resolution CT volumes for supervised training—data that is rarely available in clinical repositories. For example, Fok et al.[13] introduce a physics-aware super-resolution framework trained on cone-beam CT (CBCT) datasets, where paired reconstructions are generated from the same raw HR data using distinct modulation transfer function (MTF) kernels. Their U-Net-based architecture demonstrates marked improvements in structural similarity and edge preservation, and the method is explicitly motivated by the goal of generating higher-quality DRRs. While elegant in design, this approach remains limited by the need for paired CT data and the absence of publicly available code or pretrained models, hindering its direct applicability in the scope of this project.

Thus a pivot to traditional computer vision techniques is necessary. Here, the works of Ledig et al[14] on SRGAN, a widely used generative adversarial network (GAN) for single image super-resolution may be employed due to its accessibility, perceptual realism, and modular integration. While originally developed for natural RGB images, its architecture and open-source availability render it a promising off-the-shelf baseline for medical imaging contexts, particularly when realistic high-resolution CT volumes are unavailable.

In contrast, traditional methods such as SRGAN[14], a widely used generative adversarial network for single image super-resolution, may be employed due to its accessibility, perceptual realism, and plug-and-play integration. SRGAN replaces traditional mean squared error (MSE) losses with a perceptual loss that combines adversarial training and VGG-based content similarity—encouraging the generation of sharper, more photo-realistic textures. While originally developed for natural images, its open-source implementation and architectural simplicity make it a practical baseline for CT super-resolution in settings where high-resolution training data is unavailable. However, its use outside of its intended domain warrants caution: GAN-based models can hallucinate unrealistic textures or amplify non-physical features, potentially introducing artifacts into downstream DRRs that could mislead classification models. This risk necessitates a thorough analysis of the downstream deep learning model's interpretability—such as inspecting attention heatmaps or saliency maps—to ensure that any improvements in performance are not driven by spurious textures or non-anatomical cues introduced during the super-resolution process.

Chapter 4

Methodology

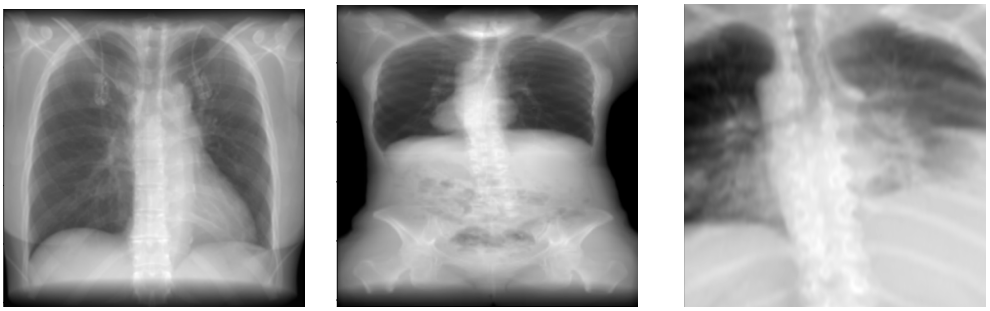
4.1 COCA Dataset

Our experiments are based on Stanford's open-source Coronary Calcium and Chest CT (COCA) dataset[15] which consists of two subsets: (1) 790 gated coronary CT scans with corresponding coronary artery calcium segmentations and scores in XML format, and (2) 213 nongated full chest CTs with total and per-vessel Agatston scores.

CT Image Acquisition and Metadata

Each patient study comprises a volumetric stack of axial DICOM slices with in-plane resolution of 512×512 pixels. Scans were acquired using heterogeneous imaging protocols, with variations in slice thickness (typically 2.5-5mm), inter-slice spacing, and field of view (FOV). The number of slices per scan ranged from 6 to 180.

To ensure data quality, volumes with insufficient anatomical coverage (e.g., < 30 slices), excessive FOV, or obfuscated anatomy (see Figure 4.1) were excluded. This resulted in 667 usable gated CTs and 211 usable nongated CTs.



(a) Accepted DRR (b) Large FOV (c) Obfuscated anatomy

Figure 4.1: Examples of generated DRRs. (a) Accepted projection. (b–c) Excluded due to FOV or radiographic obfuscation.

CAC Score Labels

For each patient, we extracted total Agatston scores—a clinical metric quantifying the extent of coronary artery calcification—computed as the area of each calcified lesion multiplied by a weight based on peak Hounsfield Unit (HU) intensity[3].

In the nongated subset, total scores were provided directly. In the gated subset, scores were reconstructed from per-lesion ROI data by applying the standard area $\times$ HU weighting logic to each segmented lesion in the XML annotations.

To reduce the problem to classification, Binary CAC labels were derived by thresholding total Agatston scores at ≥ 10, a threshold commonly used to distinguish any measurable calcification from a score of zero. Patients with scores above this threshold were labeled as CAC-positive (class 1), and those below as CAC-negative (class 0).

To reduce the problem to classification, binary CAC labels were derived by thresholding total Agatston scores at ≥ 10, a threshold commonly used to distinguish any measurable calcification from a score of zero. Patients with scores at or above this threshold were labeled as CAC-positive (class 1), and those below as CAC-negative (class 0). Figures 4.2 and 4.3 show the class distributions across the full Agatston severity range. The green segment corresponds to CAC-negative cases and serves as the reference for binary label balance. Both subsets exhibit a moderate class imbalance, particularly at higher severity levels.

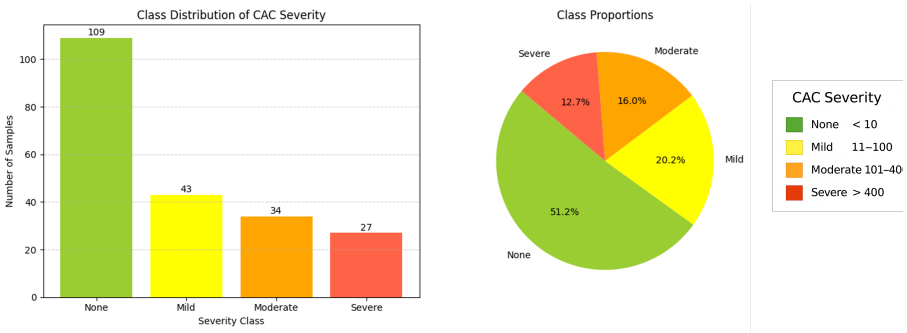


Figure 4.2: Distribution of Agatston scores in the nongated subset.

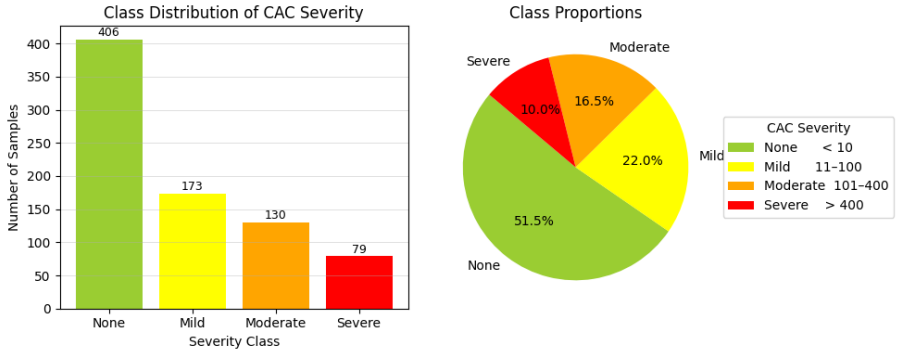


Figure 4.3: Distribution of Agatston scores in the gated subset.

4.2 DRR Generation from CT Scans

To generate digitally reconstructed radiographs (DRRs), axial DICOM slices were first stacked into a 3D volume and resampled to isotropic voxel spacing using the patient's native in-plane resolution (δx) via B-spline interpolation. This ensured uniform spatial resolution across all three dimensions, which is a prerequisite for consistent forward projection.

The resulting isotropic volumes were projected using `DiffDRR`, a GPU-accelerated differentiable ray-tracing library that implements the Siddon algorithm in Python. Each volume was projected onto a virtual detector of size 512×512 pixels, with an in-plane pixel spacing of 1 mm and a fixed source-to-detector distance (SDD) of 1085.60 mm. This produced synthetic DRRs that approximate the geometry and resolution of standard posteroanterior (PA) chest radiographs.

Both PA and lateral (LA) views were simulated by rotating the 3D volume by $\frac{\pi}{2}$ radians around the superior-inferior axis prior to projection.

Differences in nongated and gated DRRs are illustrated in Figure 4.4 where gated DRRs are derived from cardiac-focused CT scans with a restricted field-of-view centred on the heart, and nongated DRRs originate from full chest CT scans with broader anatomical coverage but potentially noisier CAC signals.

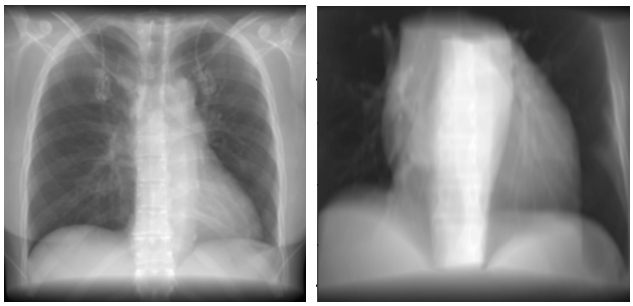


Figure 4.4: Example DRRs generated from nongated (left) and gated (right) CT volumes.

4.3 Preprocessing to Enhance DRR Fidelity

CT Volume Enhancement via Super-Resolution

For each CT scan orientation, two DRR variants were generated: a native-resolution baseline and a super-resolved (SR) version. The SR pipeline was designed to mitigate the degradation in axial resolution inherent to many low-dose CT protocols, particularly in the nongated subset.

Specifically, sagittal slices were extracted from the original 3D volume and upsampled by a factor of 4× using a pretrained SRGAN model. These enhanced slices were then reassembled into a higher-resolution intermediate volume, which was subsequently downsampled along the coronal axis to a fixed width of 512 pixels—ensuring compatibility with the downstream projection pipeline. Finally, the super-resolved volume was resampled to isotropic voxel spacing via B-spline interpolation and projected identically to the native-resolution variant.

This super-resolution procedure improved anatomical coherence and edge sharpness in the resulting DRRs while preserving the geometric fidelity of the original CT volume. Quantitative validation using SSIM, PSNR, and Laplacian variance confirmed that SR-enhanced DRRs exhibited superior perceptual quality, particularly in rib contours and cardiac borders.

Post-Projection Contrast Enhancement

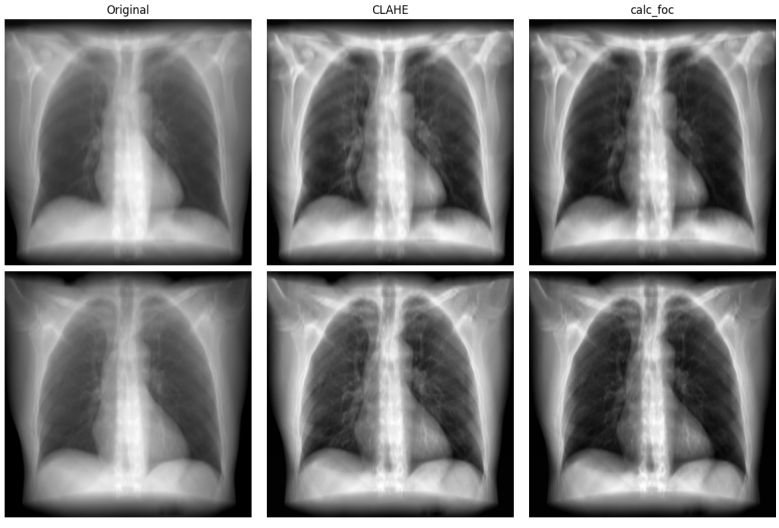
In addition to volumetric enhancement, we explored two DRR-level preprocessing strategies aimed at improving visual contrast and calcification salience.

The first employed contrast-limited adaptive histogram equalization (CLAHE), which enhances local contrast in soft tissue regions while avoiding global saturation. CLAHE was applied independently to each DRR and yielded sharper anatomical delineation, especially in lung fields and mediastinum.

The second mode, referred to as *calcium-focused* enhancement (`calc_foc`), was a custom composite transformation tailored to amplify radiodense calcifications while suppressing irrelevant structures.Parameters were empirically tuned to attenuate background structures such as ribs and soft tissue while preserving high-intensity calcium signals. This pipeline comprised of the following:

1. Gamma correction ($\gamma = 1.5$) to adjust image luminance and emphasise mid-range intensities.
2. CLAHE (Contrast Limited Adaptive Histogram Equalisation) to locally enhance contrast while mitigating over-amplification of noise.
3. An unsharping mask to sharpen edges and accentuate fine structural details

These preprocessing modes are illustrated in Figure 4.5.



(a) (b) (c)

Figure 4.5: Comparison of preprocessing variants applied to Native DRR (Top) and Super-Resolved DRR (bottom): (a) Original, (b) CLAHE-enhanced, and (c) Calcium-focused enhancement.

4.4 Model Architectures

We evaluated a range of convolutional neural network (CNN) architectures for binary classification of CAC presence from DRRs. All models output a probability score corresponding to the CAC-positive class and were trained under consistent preprocessing settings. Three primary architecture families were explored:

CNN5_GAP: Custom Baseline

A lightweight four-block convolutional neural network (CNN5_GAP) was developed as a baseline model. As shown in Figure 4.6, it consists of five convolutional layers with batch normalization, leaky ReLU activations, and max-pooling, followed by a global average pooling (GAP) layer and a two-layer fully connected classifier. Designed for efficiency on small datasets, this model contains approximately 100k trainable parameters and offers fast convergence with low risk of overfitting.

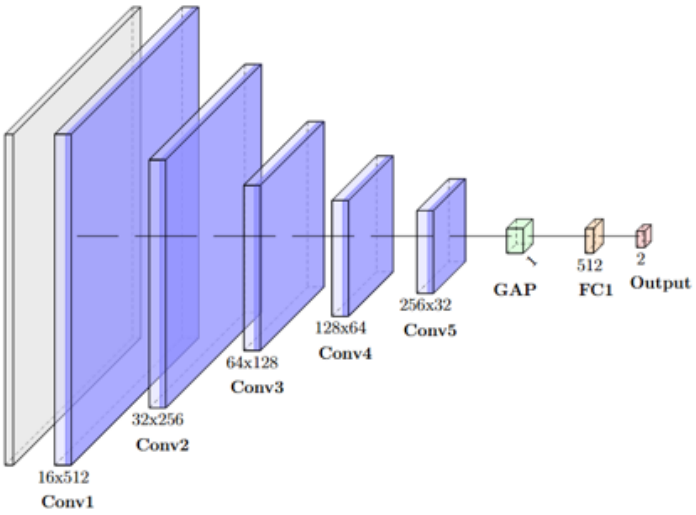


Figure 4.6: CNN5_GAP architecture with 5 Convolutional blocks, batch normalisation and max pooling followed by a global adaptive pooling layer and a classifier head.

DenseNet121 with Attention

To leverage pretrained radiographic features, we adapted DenseNet121 pretrained on the CheXpert chest X-ray dataset. As inspired by prior work, the final classification head was replaced with an attention mechanism operating on the convolutional feature maps, followed by fully connected layers. This spatial attention highlights salient anatomical regions—such as cardiac borders and coronary arteries—while suppressing irrelevant background features. The added interpretability also facilitates Grad-CAM visualizations for downstream attribution analysis.

Fusion Models: Complementary Projections

Dual-view models were implemented to explore whether combining posteroanterior (PA) and lateral (LA) projections enhances classification performance. Three fusion strategies were explored:

- **Early Fusion**: PA and LA views were concatenated along the channel dimension and passed through a single CNN.
- **Intermediate Fusion**: PA and LA views were passed through independent or shared CNN encoders, followed by global average pooling. The resulting feature vectors were combined using a learned scalar gate, producing a weighted latent embedding for classification.
- **Cross-Attention Fusion**: PA and LA views were passed through shared encoders. The LA feature map received cross-attention from the PA map before pooling and classification, enabling the model to align complementary spatial features across views.

These models aimed to capture orthogonal anatomical perspectives from the same CT volume. Results for all architectures are reported in Chapter 5.

4.5 Evaluation

Area under the receiver operating characteristic curve (AUC-ROC) was adopted as the primary evaluation metric throughout all experiments due to its threshold-independent nature. Figure 4.7 shows a representative ROC curve for one fold of the CNN5_GAP models on posteroanterior DRRs. Secondary metrics—including accuracy, precision, recall, and F1 score—were also recorded, but comparative performance reporting and statistical testing were based on validation AUC unless otherwise specified.

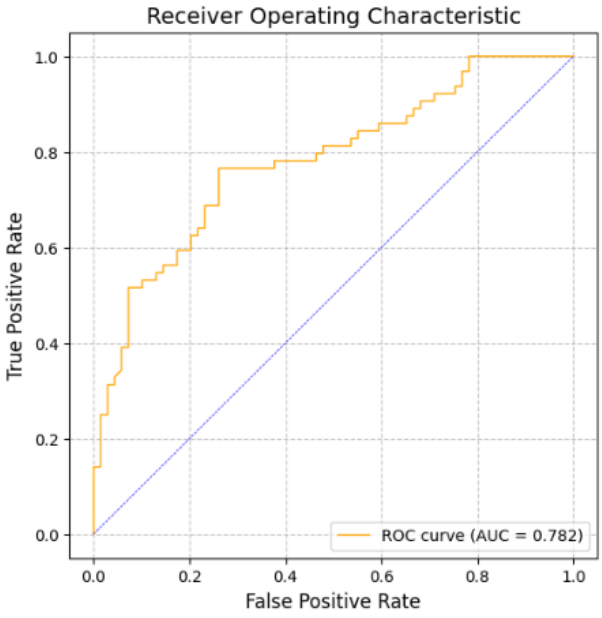


Figure 4.7: Example receiver operating characteristic (ROC) curve for a single fold and seed of the CNN5_GAP on PA DRRs. The area under the curve (AUC) for this example is 0.782.

All models were trained using a 5-fold stratified cross-validation protocol, with each fold repeated across five randomized seeds to account for variance introduced by stochastic initialization. This yielded 25 independent training runs per experimental configuration,

each with an 80/20 training, validation split. During training, model performance was evaluated after each epoch on the held-out validation fold. Final metrics were computed by averaging the top five validation AUC scores, excluding the first five epochs to allow for stabilization. This approach reduces optimism bias and minimizes the impact of early-epoch noise.

To assess the statistical significance of observed improvements, we applied the Wilcoxon signed-rank test to per-fold AUC scores ($n = 25$). This non-parametric test was chosen for two primary reasons. First, our experimental design involved paired comparisons: each model was evaluated on identical data splits across folds and seeds, allowing for matched observations. Second, the distribution of AUC scores was not assumed to be Gaussian, violating the assumptions of the paired t-test. The Wilcoxon test does not require normality and is robust to small sample sizes and outliers, making it an appropriate and conservative choice for detecting consistent differences across matched folds. Mean AUCs were reported alongside p-values to quantify both effect size and statistical significance.

Chapter 5

Experiments

All experiments were implemented in Python using PyTorch. Training and evaluation were conducted on NVIDIA A100/V100 GPUs using a fixed seed protocol to ensure reproducibility. We now describe the training configuration as well as each experimental group and its corresponding findings.

5.1 Training Configuration

All models were trained using the Adam optimizer with an initial learning rate of 1×10^{-4}, a weight decay of 1×10^{-5}, and a batch size of 32. The primary loss function was cross-entropy with label smoothing ($\varepsilon = 0.1$), which was found to improve generalization over vanilla cross-entropy. Focal Loss was also evaluated in select experiments to address class imbalance, but it yielded no consistent improvements and was ultimately not adopted in final configurations.

To mitigate overfitting, early stopping was applied based on validation AUC, and light regularization was introduced via Dropout ($p = 0.2$–0.3) and Batch Normalization in all convolutional layers. Standard image augmentations—random rotation, shifting, and zoom—were applied during training to increase generalization. Horizontal or vertical flips were not used in order to preserve the anatomical orientation inherent to chest radiographs.

All training metrics, including loss, F1 score, precision, recall, and AUC, were logged per epoch and aggregated using a custom logging utility. Final tables report average performance across seeds, with best-performing configurations highlighted for each experimental group.

5.2 Analysis of Nongated Dataset

Experiments were first conducted on the nongated subset with DRRs that closely resemble full-field chest radiographs and are thus most aligned with the intended domain; these however suffer from coarse axial resolution. The goal was to evaluate how model architecture, DRR preprocessing, and CT super-resolution interact under conditions that most closely mirror the target deployment domain.

Table 5.1: Mean AUCs and Standard Deviations on Nongated Dataset Across Preprocessing Modes and Resolutions[1]

Preprocessing	**DenseNet121**		**CNN5_GAP**	
	Native	SR	Native	SR
Original	**0.674 ± 0.047**†	0.672 ± 0.058	**0.694 ± 0.051**	0.688 ± 0.041
CLAHE	0.648 ± 0.058	0.685 ± 0.060	0.683 ± 0.033	0.669 ± 0.069
Calc_foc	0.648 ± 0.044	**0.697 ± 0.048**	0.692 ± 0.035	**0.696** ± 0.074

Table 5.1 summarizes the mean AUCs across two architectures (CNN5_GAP and DenseNet121), three preprocessing modes (Original, CLAHE, and Calc_foc), and two projection types (native-resolution vs super-resolved). While certain configurations—such as DenseNet121 with calc_foc and SR, or CNN5_GAP with original DRRs—showed marginal performance improvements, the results were inconsistent overall. No preprocessing strategy, resolution or architectural configuration demonstrated a statistically or practically consistent advantage.

We attribute this instability to several properties inherent in the nongated dataset: a limited number of samples, weaker localization of coronary calcifications, and significant heterogeneity in scan resolution and field-of-view. These factors likely hindered the model's ability to learn robust and generalisable discriminative features. While this dataset's anatomical coverage aligns well with true CXRs, the poor signal-to-noise ratio and lack of sufficient data made it difficult to establish robust conclusions.

Based on these findings, we pivoted to the gated dataset for all subsequent experiments. Although its field of view is more localized, it offers a higher baseline resolution, reduced projection artifacts, and a substantially larger number of patients—leading to more consistent DRR quality and a greater diversity of calcium scores. The limited FOV's impact on anatomical alignment can be mitigated through heart-localisation methods, making this trade-off acceptable given the gains in data quality and quantity.

5.3 Gated Dataset Experiments

5.3.1 Architectural Baselines on Native DRRs

To establish a reference point for subsequent experiments, we first evaluated three baseline architectures on original, native-resolution DRRs from the gated dataset (see Table 5.2). As aforementioned, this dataset was chosen for its higher intrinsic resolution, reduced projection artifacts, and larger sample size compared to the nongated subset, offering a stronger signal-to-noise ratio for benchmarking architectures. These models were selected based on prior work and their suitability to the task under differing resource constraints.

DenseNet121 with an attention head, pretrained on the CheXpert chest X-ray dataset, served as our high-capacity pretrained baseline. ResNet18 was included as a moderate-capacity alternative, offering a shallower and more lightweight residual architecture to

[1]Daggers (†) indicate statistically significant differences ($q < 0.05$) for within-model, within-resolution comparisons only; the highest mean value is bolded, and full significance results are reported in the appendix.

contrast against DenseNet. Finally, CNN5_GAP, our custom low-parameter convolutional network designed for rapid convergence and reduced overfitting in small data regimes, was evaluated as a domain-tailored architecture.

Table 5.2: Validation AUCs for baseline architectures on original, native-resolution DRRs in the gated dataset.

Architecture	**Mean AUC ± Std**
CNN5_GAP	**0.720 ± 0.048**
ResNet18	0.714 ± 0.059
DenseNet121 (pretrained)	0.718 ± 0.031

As shown in Table 5.2, CNN5_GAP achieved the highest mean AUC (0.720 ± 0.048), slightly outperforming ResNet18 (0.714 ± 0.059) and DenseNet121 (0.718 ± 0.031). While these differences were not large, the consistent competitiveness of CNN5_GAP suggests that overparameterized models like DenseNet121 may offer limited benefit when data is scarce, even with strong pretrained initialization. This supports the hypothesis that lightweight, domain-specific architectures can match or surpass deeper backbones in small-data medical imaging tasks by reducing overfitting risk and converging faster.

Given its strong performance and efficiency, CNN5_GAP was retained as the primary model for subsequent experiments. DenseNet121 was also kept as a secondary baseline to evaluate whether advanced techniques such as fusion, curriculum learning, or self-supervised pretraining, could better exploit its pretrained feature space.

5.3.2 Effect of Super-Resolution (SR)

Following the architectural baselines, we investigated whether input fidelity was a limiting factor—particularly for smaller, custom architectures such as CNN5_GAP. Unlike pretrained networks that leverage large-scale external data, lightweight models rely more heavily on the richness of the training signal. However, DRR quality is fundamentally constrained by the axial resolution of the source CT scans. Applying super-resolution (SR) prior to projection offers a potential means of recovering fine anatomical detail and enhancing subtle calcification cues relevant for CAC detection.

Table 5.3: Effect of super-resolution on gated dataset native-resolution DRRs. Results shown as mean AUC ± standard deviation over 25 runs.

Resolution	**CNN5_GAP**	**DenseNet121**
Native	0.720 ± 0.048	**0.714 ± 0.031**
Super-Resolved	**0.727 ± 0.050**	0.708 ± 0.026

As shown in Table 5.3, SR yielded a small but consistent gain for CNN5_GAP, improving mean AUC from 0.720 ± 0.048 to 0.727 ± 0.050. This trend supports the hypothesis that input fidelity is more critical for smaller models, where improved visual cues can directly translate to better feature learning. In contrast, DenseNet121 saw a slight decline under SR, potentially due to a domain shift between SR-enhanced DRRs and the natural image statistics encountered during pretraining. While the difference is modest, the opposite

trends between models suggest that SR may offer selective benefits depending on model capacity and training regime. These findings motivated further experiments combining SR with alternative preprocessing techniques to test whether these effects could be amplified or mitigated.

5.3.3 Impact of Preprocessing Modes

Having assessed the effect of SR in isolation, we next examined whether post-projection preprocessing could enhance CAC classification performance. To ensure a controlled comparison, these experiments were conducted exclusively on native-resolution DRRs—without SR—so that preprocessing effects could be evaluated independently. (For combinations of SR and preprocessing, see Appendix A.2.)

In clinical chest radiographs, contrast enhancement and masking techniques are often used to highlight anatomical boundaries or suppress irrelevant structures. We explore whether similar strategies applied to synthetic DRRs can improve CAC classification. We evaluated three modes:

- **Original** — unmodified DRRs.
- **CLAHE** — contrast-limited adaptive histogram equalization, with a default clip limit and 8×8 tile grid size, applied directly to the DRR to locally enhance contrast while mitigating over-amplification of noise.
- **calc_foc** — a custom three-stage transformation designed to boost mid-tone contrast and sharpen fine details:
  1. **Gamma correction** with $\gamma = 1.5$ to enhance mid-intensity structures.
  2. **CLAHE** to improve local contrast in calcium-rich and soft-tissue regions.
  3. **Unsharp masking** to emphasise edges and fine structural boundaries, improving calcification visibility.

Table 5.4: Effect of preprocessing mode on native-resolution DRRs.

Preprocessing	CNN5_GAP	DenseNet121
Original	0.720 ± 0.048	$\mathbf{0.714 \pm 0.031}$
CLAHE	0.728 ± 0.049	0.688 ± 0.053
Calc_foc	$\mathbf{0.729 \pm 0.053}$	0.683 ± 0.058

As shown in Table 5.4, both CLAHE and calc_foc produced small but consistent gains over the original DRRs for CNN5_GAP, with calc_foc achieving the highest mean AUC (0.729 ± 0.053). In contrast, DenseNet121's performance declined under both enhanced modes. This is likely due to a distributional shift introduced by preprocessing, diverging from the CheXpert-pretrained feature space. While pretraining on large-scale CXR data offers strong prior representations, its benefit can be attenuated when the input distribution changes significantly.

This highlights a fundamental tradeoff when working with pretrained networks. Pretraining on large-scale chest X-ray datasets (e.g., CheXpert) equips a model with strong prior representations, but these priors are tuned to the original training distribution.

Image preprocessing—such as contrast enhancement or sharpening—can shift the input distribution enough to reduce the benefit of those pretrained features. In the context of CAC detection, where calcifications are subtle and easily obscured, one must weigh whether to retain the pretrained feature space and avoid aggressive preprocessing, or instead train from scratch on preprocessed images that explicitly enhance these faint signals. Without re-pretraining on a dataset like CheXpert using the same preprocessing pipeline, these approaches remain mutually constraining.

In contrast, CNN5_GAP—trained from scratch—appears more adaptable to distributional changes, leveraging its lower parameter count and reduced overfitting risk. This adaptability, coupled with consistently competitive performance, reinforced its selection as the primary architecture for subsequent experiments.

5.3.4 Projection Fusion

Having established CNN5_GAP with SR and calc_foc preprocessing as our most performant configuration (See Appendix A.2), we next explored whether combining multiple projections—specifically lateral (LA) and posteroanterior (PA) DRRs—could improve performance through complementary spatial information. While a single projection captures a compressed 2D representation of volumetric structure, it inherently occludes or flattens anatomical cues. Fusion strategies aim to mitigate this by aggregating multiple views into a richer latent representation, allowing the network to reason over different anatomical perspectives.

As mentioned in Section 4.4, we evaluate three fusion paradigms.

Table 5.5: Comparison of PA+LA fusion models. Fusion strategies vary by where interaction occurs, and whether encoders are shared or unshared.[2]

Fusion Strategy	**Shared Encoder**	**Unshared Encoder**
Early Fusion	0.702 ± 0.044	–
Intermediate Fusion	0.736 ± 0.059	0.739 ± 0.065
Cross-Attention Fusion	$\mathbf{0.740 \pm 0.043}^{\dagger}$	0.729 ± 0.045

As shown in Table 5.5, cross-attention with shared encoders yielded the highest AUC (0.740), narrowly outperforming intermediary fusion. Early fusion underperformed, likely because merging projections before sufficient feature abstraction entangles spatial signals and limits the network's ability to learn projection-specific representations. The superiority of shared-encoder cross-attention suggests that a unified feature space, enriched by attention-driven interaction between views, can yield a more discriminative representation without overfitting. While the performance gain over single-projection models was modest and not statistically significant, the improvement was consistent across folds, indicating potential for further optimisation.

Taken together, the results from our single-projection experiments established CNN5_GAP with SR and calc_foc preprocessing as the most effective baseline configuration, balancing adaptability, fidelity, and robustness. The fusion experiments then demonstrated that

[2]Daggers (†) indicate statistically significant differences ($p < 0.05$)

integrating complementary PA and LA views—particularly via shared-encoder cross-attention—consistently enhanced performance, albeit modestly. This progression reflects a layered optimisation strategy: first maximising the discriminative power of individual projections through super-resolution and targeted preprocessing, and then enriching the learned feature space by combining projections in a semantically meaningful way. Building on these insights, we next investigate whether training strategies can further amplify these gains by shaping the data presentation to the model during optimisation.

5.3.5 Curriculum Learning

We investigated whether a curriculum-based training strategy could further improve the optimized fusion model's performance. Curriculum learning introduces training examples in order of increasing difficulty, a paradigm inspired by human learning that has been shown to improve convergence and generalization, particularly on noisy or imbalanced datasets.

In our implementation, difficulty was defined based on a sample's proximity to the binary classification threshold (Agatston score = 100). DRRs associated with extreme calcium scores—either clearly positive or clearly negative—were treated as "easy" and introduced early in training. As epochs progressed (epoch ≥ 50), we gradually incorporated more ambiguous samples closer to the decision boundary until the entire training set was included. This approach encourages the model to first learn coarse discriminative features before encountering borderline cases.

Table 5.6: Effect of curriculum learning on cross-attention fusion model. Curriculum learning improves mean AUC and stabilizes performance across seeds.

Training Strategy	**Mean AUC**
Standard Fusion Training	0.740 ± 0.043
Curriculum Fusion Training	$\mathbf{0.750} \pm 0.046$

As shown in Table 5.6, curriculum learning improved the mean AUC by roughly 1 point over the baseline fusion model. This suggests that introducing easier cases first may stabilize early gradients and act as a form of implicit regularization, reducing overfitting on noisy or borderline samples. The curriculum strategy is particularly beneficial in clinical classification tasks where subtle features and low inter-class contrast complicate early learning dynamics.

5.3.6 SimCLR Pretraining

We next investigated whether self-supervised pretraining could further improve the fusion model's performance by enhancing feature representations before supervised fine-tuning. We used a SimCLR-style contrastive learning framework, which encourages the model to learn invariant features by maximizing agreement between differently augmented views of the same image while pushing apart representations of different images. This approach is particularly appealing for medical imaging tasks where labeled data is scarce but unlabeled images are abundant.

For pretraining, we used only the training set DRRs and applied random cropping, rotation, and intensity jittering as positive-pair augmentations. The pretrained encoder weights were then used to initialize the fusion model, which was fine-tuned on the CAC classification task using the same cross-validation splits as before. We evaluated two setups: (1) SimCLR pretraining only, and (2) SimCLR pretraining followed by curriculum learning ("SimCLR + Curriculum").

Table 5.7: Final fusion model performance across training strategies. SimCLR and curriculum learning offer complementary benefits, with the combined approach achieving the best result.

Training Strategy	**Mean AUC**
Fusion (standard)	0.740 ± 0.043
Curriculum Learning	0.750 ± 0.046
SimCLR Pretraining	0.742 ± 0.045
SimCLR + Curriculum	$\mathbf{0.754 \pm 0.055}^{\dagger}$

As shown in Table 5.7, SimCLR alone yielded a modest performance gain over the fusion baseline, indicating that unsupervised representation learning captures useful inductive structure in DRRs. However, the combination of SimCLR and curriculum learning led to the highest overall performance, with a mean AUC of 0.754 ± 0.055. Various degrees of fine-tuning were explored, however the best results were achieved by leveraging the pretrained encoder as a form of weight initialisation and keeping the entire model unfrozen. This result supports the hypothesis that representational pretraining and curriculum-based supervision act as complementary regularizers—the former shaping the encoder's feature space, and the latter structuring the supervised learning process in a more pedagogically effective manner.

These findings suggest that carefully staged supervision, when scaffolded on strong latent features, can help overcome the noise and ambiguity that typically limit performance in low-resource medical imaging settings.

5.4 Summary of Experimental Findings

We began by evaluating the nongated subset, aiming to approximate full-field chest X-rays. Performance across architectures, preprocessing modes, and super-resolution variants proved inconsistent, likely due to weak calcification signals, small sample size, and heterogeneous scan quality that hindered robust feature learning. These limitations motivated a pivot to the gated dataset, which—despite its localized field-of-view—offered higher-quality DRRs and stronger CAC signal, enabling more reliable evaluation.

On this dataset, we first established architectural baselines and found that our lightweight CNN (CNN5_GAP) matched or exceeded larger pretrained models (AUC = 0.720), indicating that overparameterised backbones may not provide a consistent advantage in this domain. Super-resolution modestly improved CNN5_GAP performance but degraded DenseNet121, suggesting that pretrained filters may be tuned to lower-frequency features and disrupted by the high-frequency detail introduced by SR. Post-projection preprocessing, particularly contrast-focused transformations, further improved CNN performance while

degrading DenseNet121, highlighting a tradeoff between leveraging prior representations and adapting to domain-specific transformations.

We then explored dual-view fusion, finding that gated intermediate fusion yielded the best results. The gain likely stems from complementary spatial cues in PA and LA projections—features occluded in one view may appear more distinctly in the other—with gating allowing adaptive weighting of each view's contribution.

Finally, curriculum learning and SimCLR-style contrastive pretraining each provided incremental gains: curriculum learning improved optimization stability by introducing easy, high-confidence samples first, while contrastive pretraining enhanced representation quality in limited-data settings. Their combination achieved our highest overall AUC (0.754), underscoring the value of integrating structured supervision with enriched feature initialization. Collectively, these findings demonstrate that DRR-based CAC detection benefits most from strategies that jointly enhance input fidelity, exploit complementary views, and tailor the training process to the structure of the data.

Limitations

While the results presented here demonstrate promising performance for DRR-based CAC detection, several limitations must be acknowledged. These constraints not only define the scope of the current findings but also highlight important opportunities for future work.

Synthetic-real domain gap

Although DRRs closely resemble clinical chest radiographs in appearance, they remain synthetic images generated from CT volumes via ray-projection algorithms, and should thus be treated as approximations. Consequently, they may not fully capture the distributional complexity of real CXRs, including factors such as variable patient positioning, exposure settings, imaging hardware, and incidental artefacts. This raises legitimate concerns regarding the direct generalisability of our findings to real-world clinical settings without explicit domain adaptation or fine-tuning on real CXR data.

Limited anatomical field of view

The gated CT scans used for DRR generation are cardiac-focused acquisitions, typically encompassing only a small region surrounding the heart. As a result, the DRRs derived from these scans exclude substantial portions of the thoracic cavity that are present in full-field chest radiographs. In the absence of a robust localisation or segmentation stage, models trained on such data may underperform when applied to full-field CXRs, where the proportion of cardiac structures is reduced and additional anatomical variation is present.

Dataset scale and variance

While a dataset of 700 gated DRRs is comparatively large for CAC-specific research, it remains modest by deep learning standards. The small validation set size per fold in our cross-validation setup likely contributes to the high variance observed across seeds and

folds, limiting the stability of reported metrics. Larger datasets would not only improve statistical power but also enable the exploration of more data-hungry architectures without the same degree of overfitting risk.

Weak supervision and absence of spatial localisation

All experiments in this study were conducted using image-level labels derived from patient-level Agatston scores, without explicit localisation of the coronary arteries or individual lesions. This weak supervision setup means that models may rely on indirect visual cues or spurious correlations rather than directly identifying calcium features. In clinical contexts, such behaviour could compromise robustness and interpretability, particularly when applied to data from different acquisition protocols or populations.

Together, these limitations frame the boundary conditions for interpreting our results. Addressing them through: domain adaptation to real CXRs, the use of full-field datasets, larger sample sizes, and spatially localised supervision, represents a clear path for future research to strengthen both the validity and clinical applicability of DRR-based CAC detection systems.

Future Work

Building on the findings and recognising the limitations discussed above, several directions present themselves for advancing DRR-based CAC detection toward greater robustness and clinical relevance.

Bridging the synthetic–real domain gap

To address the distributional differences between DRRs and real chest radiographs, domain adaptation strategies should be explored. Techniques such as adversarial feature alignment or test-time adaptation could be employed to align the feature spaces or pixel-level distributions of synthetic and real CXRs. This may enable models trained primarily on DRRs to generalise effectively to clinical images without requiring extensive real-data retraining.

Scaling datasets and improving statistical robustness

Given the modest dataset size in this work, scaling to larger and more diverse datasets remains a priority. Future efforts should focus on acquiring or curating datasets that pair full-field CXRs with ground-truth calcium scores, ideally across multiple institutions and imaging protocols. Larger sample sizes would reduce variance, improve statistical power for significance testing, and permit exploration of higher-capacity models without substantial overfitting risk.

Extending field-of-view coverage

The limited spatial scope of gated CT-derived DRRs can be mitigated by introducing automated localisation methods for full-field CXRs. Learnable cropping or attention-based localisation modules could dynamically identify heart-centric regions in full CXRs,

providing inputs more comparable to gated DRRs while preserving efficiency. This approach could also serve as a preprocessing step for deployment in settings where gated CT is not available.

Localisation-aware supervision

Another promising avenue is to exploit the existing CT-derived CAC annotations in the COCA dataset for weakly-supervised localisation in the DRR domain. Each annotated calcification in CT space can be represented as a binary mask or centroid and forward-projected into the DRR plane using the same ray-tracing geometry applied during DRR generation. The resulting 2D regions provide coarse localisation cues that could be integrated into training via weakly-supervised attention mechanisms or multiple-instance learning (MIL), enabling the model to focus on anatomically relevant areas without requiring manual DRR segmentation. Furthermore, post hoc interpretability techniques such as Grad-CAM or Integrated Gradients could be compared against these projected regions to assess the degree to which model attributions align with true coronary anatomy. Such methods could strengthen the clinical validity of the learned representations and help disentangle performance gains driven by relevant pathology from those due to spurious correlations.

Incorporating multimodal information

CAC is only one dimension of cardiovascular risk, which is influenced by demographic, clinical, and biochemical factors. Integrating structured patient metadata — such as age, sex, smoking history, and lipid profiles — alongside imaging data could enhance predictive accuracy and enable risk stratification tailored to individual patient profiles. Multimodal fusion approaches could be implemented at either feature or decision level to capture synergistic relationships between visual and non-visual information.

Taken together, these directions aim to move beyond proof-of-concept performance on synthetic data toward robust, generalisable, and clinically meaningful CAC detection systems. The combination of domain adaptation, improved localisation, dataset expansion, and multimodal integration represents a natural progression toward deployment-ready models capable of assisting in population-level cardiovascular risk screening.

Conclusion

This work explored the feasibility of using CT-derived Digitally Reconstructed Radiographs (DRRs) as a proxy training domain for coronary artery calcification (CAC) detection, motivated by the scarcity of large, labeled chest X-ray datasets paired with calcium scores. Through a series of controlled experiments on DRRs generated from gated CT scans, we systematically evaluated architectural choices, image enhancement strategies, multi-view fusion techniques, and advanced training paradigms such as curriculum learning and contrastive pretraining. Our final configuration — combining a lightweight CNN5_GAP architecture with super-resolution inputs, composite calc_foc preprocessing, shared cross-attention fusion, and SimCLR-initialised curriculum training — achieved an AUC of 0.754, surpassing several prior works on prior CXR-based CAC studies while relying only on CT-derived labels.

The results highlight three central findings. First, tailored lightweight architectures can match or exceed the performance of larger pretrained backbones in low-data, domain-specific settings. Second, complementary projections (PA + LA) offer measurable performance gains when combined at semantically meaningful stages through learned fusion mechanisms. Third, the integration of structured supervision (via curriculum learning) and robust feature initialisation (via contrastive pretraining) can stabilise training and improve generalisation in data-scarce regimes.

Beyond these methodological contributions, this study highlights the potential of DRRs as a scalable, label-rich surrogate for real CXR data. While the synthetic nature of DRRs and the limited anatomical coverage of gated scans impose constraints, the demonstrated performance suggests that, with appropriate domain adaptation, DRR-trained models could serve as a foundation for clinically applicable CAC detection systems. The pathway forward lies in bridging the domain gap, expanding datasets, enhancing spatial localisation, and integrating multimodal risk factors — steps that could transform this proof-of-concept into a robust tool for cardiovascular risk screening. If successfully adapted to real CXR data, such models could enable scalable, low-cost screening for subclinical cardiovascular disease, particularly in resource-limited settings.

Appendix A

Statisitcal Testing Methods

For each paired comparison (same folds/seeds), we apply the Wilcoxon signed-rank test on per-run AUC values. Multiple comparisons within an experiment family are corrected using the Benjamini–Hochberg (FDR) method where applicable. We report the test statistic, p-value, and q-value. Significant results are bolded at $q < 0.05$ (FDR-adjusted) or $p < 0.05$ when no adjustment is applied.

A.1 Nongated Dataset

Table A.1: Nongated dataset – Wilcoxon signed-rank test results within DenseNet model. q-values are Benjamini–Hochberg adjusted across all comparisons in this table; significant results at $q < 0.05$ are in bold.

Comparison	Stat	p-value	q-value	Mean AUC A	Mean AUC B
Native: orig vs calc_foc	75.0	0.0173	**0.0346***	0.6738	0.6479
Native: orig vs clahe	70.0	0.0115	**0.0346***	0.6738	0.6477
Native: calc_foc vs clahe	153.0	0.798	0.7980	0.6479	0.6477
SR: orig vs calc_foc	56.5	0.00434	**0.0195***	0.6724	0.6965
SR: orig vs clahe	98.0	0.0826	0.1239	0.6724	0.6850
SR: calc_foc vs clahe	103.0	0.109	0.1395	0.6965	0.6850
native vs sr (orig)	156.0	0.874	0.8740	0.6738	0.6724
native vs sr (calc_foc)	7.0	0.000001	**0.000009***	0.6479	0.6965
native vs sr (clahe)	41.0	0.000556	**0.00250***	0.6477	0.6850

Table A.2: Nongated CNN5_GAP Wilcoxon signed-rank test results. Bold indicates $q < 0.05$ after Benjamini–Hochberg correction.

Comparison	Stat	p-value	q-value	Mean AUC A	Mean AUC B
Native: orig vs calc_foc	161.0	0.979	0.979	0.694	0.692
Native: orig vs clahe	113.0	0.191	0.289	0.694	0.683
Native: calc_foc vs clahe	110.0	0.164	0.269	0.692	0.683
SR: orig vs calc_foc	142.5	0.590	0.689	0.688	0.696
SR: orig vs clahe	81.0	0.027	0.0695	0.688	0.669
SR: calc_foc vs clahe	56.0	**0.00309**	**0.0139***	0.696	0.669
Native vs SR (orig)	124.0	0.300	0.450	0.694	0.688
Native vs SR (calc_foc)	147.0	0.692	0.766	0.692	0.696
Native vs SR (clahe)	127.0	0.353	0.499	0.683	0.669

Table A.3: Nongated inter-model (CNN5_GAP vs DenseNet) Wilcoxon signed-rank test results.

Comparison	Stat	p-value	q-value	Mean AUC A	Mean AUC B
SR - orig	103.0	0.113	0.187	0.688	0.672
SR - calc_foc	155.0	0.853	0.874	0.696	0.697
SR - clahe	145.0	0.653	0.746	0.669	0.685

A.2 Gated Dataset

Table A.4: Gated dataset – Statistical significance testing within CNN5_GAP. Wilcoxon signed-rank test with Benjamini–Hochberg FDR correction applied within this family of comparisons.

Comparison	Stat	p-value	q-value	Mean AUC A	Mean AUC B
Native: orig vs calc_foc	97.0	0.0780	0.195	0.720	0.729
Native: orig vs clahe	104.0	0.120	0.210	0.720	0.728
Native: calc_foc vs clahe	151.0	0.771	0.771	0.729	0.728
SR: orig vs calc_foc	66.0	0.00807	0.0726	0.727	0.735
SR: orig vs clahe	89.0	0.0480	0.196	0.727	0.744
SR: calc_foc vs clahe	123.0	0.300	0.375	0.735	0.744
Native vs SR (orig)	129.0	0.381	0.407	0.720	0.727
Native vs SR (calc_foc)	102.0	0.107	0.195	0.729	0.735
Native vs SR (clahe)	91.0	0.0551	0.196	0.728	0.744

Table A.5: Gated dataset – Statistical significance testing within DenseNet. Wilcoxon signed-rank test with Benjamini–Hochberg FDR correction applied within this family of comparisons.

Comparison	Stat	p-value	q-value	Mean AUC A	Mean AUC B
Native: orig vs calc_foc	78.0	0.0219	0.0618	0.718	0.683
Native: orig vs clahe	81.0	0.0275	0.0618	0.718	0.688
Native: calc_foc vs clahe	122.0	0.276	0.345	0.683	0.688
SR: orig vs calc_foc	78.0	0.0219	0.0618	0.708	0.688
SR: orig vs clahe	80.0	0.0255	0.0618	0.708	0.695
SR: calc_foc vs clahe	128.0	0.367	0.367	0.688	0.695
Native vs SR (orig)	105.0	0.122	0.183	0.718	0.708
Native vs SR (calc_foc)	150.0	0.751	0.751	0.683	0.688
Native vs SR (clahe)	129.0	0.381	0.381	0.688	0.695

A.3 Fusion Analysis

A.3.1 Lateral-only Performance

Here we merely explore the strength of the calcium signal in Lateral (LA) projections and therefore omit any BH corrections as multiple comparisons are minimal.

Table A.6: Wilcoxon signed-rank test results for LA-only performance across preprocessing modes using CNN5_GAP. Statistically significant results at $p < 0.05$ are marked with *.

Comparison	Stat	p-value	Mean AUC A	Mean AUC B
LA Native: orig vs calc_foc	37.0	**0.000329***	0.672	0.695
LA Native: orig vs clahe	57.0	**0.00342***	0.672	0.691
LA Native: calc_foc vs clahe	120.5	0.258	0.695	0.691

A.3.2 Fusion and Training Strategy Comparisons

Table A.7: Pairwise Wilcoxon signed-rank test results comparing fusion baselines, curriculum learning, and SimCLR variants.

Comparison	Stat	p-value	Mean AUC A	Mean AUC B
Early vs Int. Fusion (Shared)	11.0	**0.000046***	0.703	0.736
Early vs Int. Fusion (Unshared)	11.0	**0.000046***	0.703	0.740
Early vs Cross-Attention (Shared)	154.0	0.832509	0.739	0.740
Standard vs Curriculum	132.0	0.426142	0.740	0.750
Standard vs SimCLR	152.0	0.791476	0.740	0.742
Standard vs Curriculum + SimCLR	128.0	0.366585	0.740	0.754

Bibliography

[1] D. Eng, C. Chute, N. Khandwala, J. Cho, N. Sheets, A. Buckler, P. Goel, P. Rajpurkar, R. L. Ball, I. Ha, *et al.*, "Automated coronary calcium scoring using deep learning with multicenter external validation," *NPJ Digital Medicine*, vol. 4, no. 1, p. 88, 2021.

[2] P. P. Parikh, N. Shah, H. Ahmed, P. Schoenhagen, and M. Fares, "Coronary artery calcium scoring: Its practicality and clinical utility in primary care," *Cleveland Clinic Journal of Medicine*, vol. 85, no. 9, pp. 707–716, 2018.

[3] A. S. Agatston, W. R. Janowitz, F. J. Hildner, N. R. Zusmer, M. Viamonte, and R. Detrano, "Quantification of coronary artery calcium using ultrafast computed tomography," *Journal of the American College of Cardiology*, vol. 15, no. 4, pp. 827–832, 1990.

[4] A. Berrington de González, M. Mahesh, K. P. Kim, M. Bhargavan, R. Lewis, F. Mettler, and C. Land, "Projected cancer risks from computed tomographic scans performed in the united states in 2007," *Archives of internal medicine*, vol. 169, no. 22, pp. 2071–2077, 2009.

[5] P. I. Kamel, P. H. Yi, H. I. Sair, and C.-T. Lin, "Prediction of coronary artery calcium and cardiovascular risk on chest radiographs using deep learning," *Radiology: Cardiothoracic Imaging*, vol. 3, p. e200486, Jun 2021.

[6] 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*, vol. 84, no. 1006, pp. 890–902, 2011.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[8] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, (New York, NY, USA), p. 41–48, Association for Computing Machinery, 2009.

[9] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," *CoRR*, vol. abs/2002.05709, 2020.

[10] G. D'Ancona, M. Massussi, M. Savardi, A. Signoroni, L. Di Bacco, D. Farina, M. Metra, R. Maroldi, C. Muneretto, H. Ince, D. Costabile, M. Murero, G. Chizzola, S. Curello, and S. Benussi, "Deep learning to detect significant coronary artery disease from plain chest radiographs: Ai4cad," *International Journal of Cardiology*, vol. 370, pp. 435–441, Jan 2023.

[11] H. Jeong, H. B. Park, J. Hong, J. Lee, S.-P. Ha, R. Heo, J.-Y. Jung, Y. J. Hong, and H.-J. Chang, "Identifying coronary artery calcification using chest x-ray radiographs and machine learning: The role of the radiomics score," *Journal of Thoracic Imaging*, vol. 39, pp. 119–126, Mar 2024.

[12] V. Gopalakrishnan and P. Golland, "Fast auto-differentiable digitally reconstructed radiographs for solving inverse problems in intraoperative imaging," in *Workshop on Clinical Image-Based Procedures*, pp. 1–11, Springer, 2022.

[13] W. Y. R. Fok, A. Fieselmann, M. Herbst, L. Ritschl, S. Kappler, and S. Saalfeld, "Deep learning in computed tomography super resolution using multi-modality data training," *Medical Physics*, vol. 51, no. 4, pp. 2846–2860, 2024.

[14] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," *CoRR*, vol. abs/1609.04802, 2016.

[15] S. A. Center, "Coronary calcium and chest ct (coca)." `https://doi.org/10.71718/ge5g-ds80`, 2022. Stanford Center for Artificial Intelligence in Medicine Imaging.