

Multimodal Weakly Supervised Learning of Epicardial and Pericoronary Adipose Tissue for CCTA-Based Coronary Disease Phenotype Prediction: A Methodological Feasibility Study

Hanene Guesmi

hs.guesmi@uoh.edu.sa

University of Ha'il

Haifa Harrouch

University of Ha'il

Muna O. Alamoudi

University of Ha'il

Karim Gasmi

Jouf University

Research Article

Keywords:

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11217205/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Multimodal Weakly Supervised Learning of Epicardial and Pericoronary Adipose Tissue for CCTA-Based Coronary Disease Phenotype Prediction: A Methodological Feasibility Study

Hanene Guesmi^{1*}, Haifa Harrouch¹, Muna O. Alamoudi²,
Karim Gasmi³

¹Department of Computer Science, Applied College, University of Ha'il, Ha'il, 81411, Saudi Arabia.

²Department of Biology, Faculty of Science, University of Ha'il, Ha'il, 81411, Saudi Arabia.

³Department of Computer Science, College of Arts and Sciences at Tabarjal, Jouf University, Jouf, Saudi Arabia.

*Corresponding author(s). E-mail(s): hs.guesmi@uoh.edu.sa;

Contributing authors: h.harrouch@uoh.edu.sa; m.alamodi@uoh.edu.sa;
karim.gasmi@jouf.edu.sa;

Abstract

Purpose: Epicardial (EAT) and pericoronary (PAT) adipose tissue on cardiac CT angiography (CCTA) are associated with coronary artery disease (CAD), but most computational approaches analyse a single depot using hand-crafted radiomic features. We evaluated whether multimodal integration of EAT, PAT, and clinical variables adds value over single-source models, and whether specific architectural choices meaningfully affect performance, using a dataset-defined CAD phenotype based on expert plaque assessment rather than invasive angiographic confirmation.

Methods: We developed a weakly supervised multiple instance learning (MIL) framework combining paired EAT/PAT volumes with structured clinical variables. Using the MultiD4CAD dataset (117 patients, 39 CAD-positive), we compared single-depot imaging, multimodal fusion, clinical-only models, and MIL design variants (attention aggregation, independent versus shared-weight encoders) under stratified five-fold cross-validation with three-seed ensembling and strict out-of-fold evaluation.

Results: The full multimodal model achieved an AUC-ROC of 83.3% (95% CI, 74.3–91.1), significantly outperforming the clinical-only baseline (64.4%; Δ AUC = 18.9%, 95% CI, 7.6–30.6). No significant differences were observed across imaging configurations or architectural variants; a parameter-efficient shared-backbone configuration (11.4M parameters) matched the default dual-backbone model (22.6M parameters) and is proposed as the preferred configuration for small cardiac imaging cohorts. A simple volumetric baseline using only depot volume and attenuation statistics performed at chance level (AUC-ROC 49.1%), indicating that the imaging models capture spatial or textural information beyond gross tissue quantity. Calibration, decision curve analysis, and post-hoc error analysis further characterised model behaviour and failure modes, including threshold-dependent metrics whose derivation from validation-fold data is disclosed as a methodological limitation.

Conclusion: Multimodal integration of adipose tissue imaging and clinical data improved discrimination of a CCTA-defined CAD phenotype compared with clinical variables alone, while architectural complexity did not provide measurable benefit in this small cohort. These findings represent a methodological feasibility benchmark for imaging informatics pipelines rather than evidence of clinical readiness, pending external multicentre validation.

Keywords: Imaging informatics, Cardiac CT angiography, Multiple instance learning, Weakly supervised learning, Multimodal learning, Coronary artery disease phenotype, Clinical decision support

Acknowledgments. This research has been funded by the Scientific Research Deanship at University of Ha'il, Saudi Arabia, through project number IRP-24 014.

1 Introduction

Coronary artery disease (CAD) remains one of the leading causes of mortality worldwide, accounting for approximately nine million deaths each year [9]. Cardiac CT angiography (CCTA) provides information beyond the coronary lumen, including the epicardial (EAT) and pericoronary (PAT) adipose tissue compartments. These depots are closely linked to local inflammatory processes associated with coronary atherosclerosis and have been investigated as imaging biomarkers of CAD risk [2, 4, 7]. PAT represents a spatially focused component of EAT that directly surrounds the coronary arteries. Both depots can therefore provide non-invasive information about the local vascular environment. Importantly, they are already captured during routine CCTA, making their automated analysis potentially useful for integration into existing reporting workflows without requiring additional image acquisition.

Most computational studies of cardiac adipose tissue and CAD rely on hand-crafted radiomic features extracted from a single adipose depot [3, 6, 11–13]. Although these approaches support the predictive value of adipose tissue biomarkers, they generally do not model EAT and PAT jointly and are often difficult to reproduce or inspect at the implementation level. From a computer vision standpoint, the problem is naturally suited to multimodal weakly supervised learning. The available supervision is defined at the patient level, while the potentially informative patterns are distributed across the image volume and are not annotated at the slice level. Multiple instance learning (MIL), particularly attention-based MIL, provides a natural way to address this setting by learning which instances within a patient contribute to the final prediction. MIL has been widely adopted in computational pathology and whole-volume medical image analysis [5], but its use for joint EAT/PAT representation learning in CAD detection has not been established.

Several methodological questions therefore remain open. It is not clear whether EAT and PAT provide complementary information when learned jointly under weak supervision, or whether one depot largely captures the information contained in the other. It is also difficult to determine how much of the reported performance of multimodal models comes from the additional information itself and how much comes from architectural choices such as attention-based aggregation or encoder sharing. Evaluation has further tended to focus on discrimination, with less attention given to calibration and decision-analytic properties that are important when model outputs are intended to support clinical decisions. Finally, overall performance measures do not indicate whether errors are concentrated in particular clinical or procedural subgroups, which may be important when assessing how such systems should be monitored in practice.

We therefore conducted a pilot feasibility study focused on systematic methodological evaluation rather than claims of generalisation to unseen populations. The objective was not to introduce a new architecture, but to examine under a controlled and reproducible protocol which aspects of a multimodal imaging pipeline contribute to CAD detection performance. Three main analyses were undertaken. First, we compared single-depot imaging, multimodal imaging, and multimodal imaging combined with clinical variables to assess the contribution of data integration. Second, we examined key MIL design choices, including attention versus pooling aggregation and

independent versus shared-weight encoders. Third, we extended the evaluation beyond discrimination to include calibration, decision curve analysis, robustness checks, and post-hoc analysis of misclassified patients.

1.1 Dataset

We used the controlled-access MultiD4CAD dataset [8], a peer-reviewed data descriptor available through Zenodo. The cohort was recruited at the University Hospital Policlinico “P. Giaccone” (Palermo, Italy) from patients with suspected chronic CAD undergoing CCTA. Ethics approval and the informed-consent waiver are described in the original data descriptor. After excluding one patient because of insufficient adipose-tissue mask coverage, 117 patients remained, including 39 CAD-positive and 78 CAD-negative cases. CAD-positive patients were significantly younger than CAD-negative patients (53.2 vs 63.8 years, $p < 0.001$), whereas no other baseline characteristic differed significantly between the groups (Table 1).

For each patient, the dataset provides a 3D CCTA volume, radiologist-validated EAT and PAT segmentation masks, 11 structured clinical variables covering demographic characteristics and cardiovascular risk factors, and a binary CAD label. The label was assigned by expert visual assessment of high-risk plaque biomarkers [8]. For this study, CAD status refers to this dataset-defined imaging phenotype based on expert assessment of high-risk plaque biomarkers, rather than to angiographically confirmed obstructive CAD; this distinction is discussed further under Limitations.

Table 1 Baseline characteristics stratified by CAD status. Continuous variables: mean $\pm$ SD, t -test. Categorical: n (%), chi-squared test.

Characteristic	CAD+ (n=39)	CAD− (n=78)	p
Age, years	53.2 $\pm$ 14.5	63.8 $\pm$ 11.0	<0.001
BMI, kg/m ²	26.5 $\pm$ 4.0	26.4 $\pm$ 4.0	0.901
Male, n (%)	27 (69.2)	57 (73.1)	0.828
Smoking, n (%)	21 (53.8)	38 (48.7)	0.744
Hypertension, n (%)	14 (35.9)	43 (55.1)	0.077
Diabetes, n (%)	6 (15.4)	15 (19.2)	0.798

1.2 Preprocessing

Clinical variables were standardised or one-hot encoded separately within each training fold to avoid information leakage. For imaging, we selected $N = 24$ informative axial slices for each adipose depot and patient using a deterministic, patient-seeded mask-guided sampling procedure. For each selected slice, three adjacent axial slices were combined to provide a 2.5D context. The resulting images were multiplied by the corresponding tissue mask, z-score normalised over masked voxels, and resized to 224×224 .

For each depot, adipose-tissue area was calculated on every axial slice from the corresponding segmentation mask. Slices containing fewer than 10 tissue pixels were excluded. The remaining slices were ranked by mask area, and the top $3N$ candidates were retained. From these candidates, $N = 24$ slice indices were sampled using a deterministic, patient-seeded generator. The selected indices were cached so that the same instances were used consistently during training and evaluation.

Each selected slice was paired with its two axial neighbours to form a three-channel 2.5D pseudo-RGB input. The image was then multiplied by the tissue mask and z-score normalised using masked voxels only, followed by bilinear resizing to 224×224 . No Hounsfield-unit windowing or through-plane resampling was applied. Native in-plane spacing (0.383 mm) and slice thickness (0.75 mm) were retained. The resulting representation consisted of two 24-instance bags per patient, one for EAT and one for PAT.

1.3 Architecture

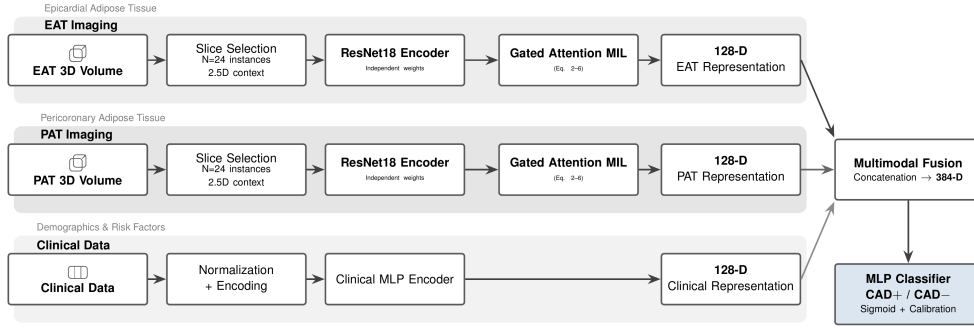


Fig. 1 Overview of the multimodal framework. EAT and PAT imaging streams operate on CCTA-derived 2.5D slice triplets, followed by ResNet18 encoders and gated attention-based MIL aggregation producing 128-D patient-level representations. Clinical data are processed through an MLP encoder into a 128-D embedding. The three modality-specific embeddings are concatenated (384-D) and passed to an MLP classifier.

EAT and PAT streams each use a ResNet18 encoder (ImageNet-pretrained) and a dedicated projection head (512-D to 128-D). Each ResNet18 backbone (approximately 11M parameters; 22.6M total in the default dual-backbone configuration) outputs a 512-D feature vector per instance, projected to 128-D via a modality-specific head (linear-GELU-LayerNorm-Dropout). The resulting instance representations were aggregated using gated-attention MIL [5]. For each bag of $K = 24$ projected instance features $\mathbf{h}_k \in \mathbb{R}^{128}$, a learned pre-activation was computed as

$$e_k = \mathbf{w}^\top (\tanh(\mathbf{V}\mathbf{h}_k) \odot \sigma(\mathbf{U}\mathbf{h}_k)), \quad (1)$$

where $\mathbf{V}, \mathbf{U} \in \mathbb{R}^{L \times 128}$, $\mathbf{w} \in \mathbb{R}^L$ ($L = 128$) are learnable and σ is the logistic sigmoid. Softmax-normalised weights $a_k = \exp(e_k) / \sum_j \exp(e_j)$ yield the bag-level embedding

$\mathbf{M} = \sum_k a_k \mathbf{h}_k \in \mathbb{R}^{128}$, obtained independently for the EAT and PAT bags. Clinical variables are encoded by an MLP into a 128-D embedding. The three 128-D embeddings (EAT, PAT, clinical) were concatenated (384-D) and passed to an MLP classification head with sigmoid output, followed by post-hoc Platt-scaling calibration. Independent versus shared-weight backbones were evaluated as a design variable. The overall architecture is summarised in Figure 1.

1.4 Training, Validation, and Statistical Analysis

Model development used stratified five-fold cross-validation with a fixed split seed. Within each fold, three models were trained using different random seeds, resulting in 15 trained models overall. All performance metrics were calculated from strict out-of-fold (OOF) predictions, ensuring that each prediction was generated by a model that had not been trained on the corresponding patient. The main training and optimisation settings are summarised in Table 2.

Table 2 Key training hyperparameters.

Parameter	Value
Instances per bag (N)	24
Batch size	4
Learning rate (η)	2×10^{-4}
Weight decay (λ)	5×10^{-4}
Dropout rate	0.5
Label smoothing (ϵ)	0.03
Early stopping patience	8 epochs
Folds / seeds	5 / {42,43,44}

Models were optimised using binary cross-entropy with label smoothing ($\epsilon = 0.03$) and AdamW, with a learning rate of 2×10^{-4} and weight decay of 5×10^{-4} , combined with gradient clipping (maximum L2 norm 1.0) and learning-rate reduction on validation AUC-ROC plateau (patience 3 epochs, factor 0.5). Weighted random sampling was used during minibatch construction to address the 1:2 CAD-positive to CAD-negative class ratio, avoiding duplication of imaging instances across training batches. Dropout, layer normalisation, decoupled weight decay, label smoothing, early stopping (patience 8 epochs), and multi-seed ensembling were combined to mitigate overfitting given the cohort size. Automatic mixed precision was used to reduce memory usage.

The primary metric was AUC-ROC, with 95% bootstrap confidence intervals (5,000 resamples, stratified). We also report AUC-PR and threshold-dependent metrics at an optimised operating point, probability calibration (Brier score), and decision curve analysis (DCA) [10]. Incremental value was quantified as ΔAUC with paired bootstrap resampling; improvements were considered significant when the 95% CI excluded zero. Robustness was assessed via permutation testing (1,000 permutations)

and Harrell bootstrap optimism correction. Seven configurations were compared (Clinical Only, EAT Only, PAT Only, EAT+Clinical, PAT+Clinical, Imaging Only, Full Multimodal), plus an architectural ablation of aggregation strategy (gated attention vs mean/max pooling) and backbone sharing. A post-hoc error analysis characterised misclassified patients by cross-validation fold and clinical variables (age, sex, smoking, comorbidities), using independent-samples t -tests and Fisher’s exact tests, treated as hypothesis-generating given small subgroup sizes.

2 Results

2.1 Overall Discrimination

The full multimodal model (EAT+PAT+Clinical) achieved an AUC-ROC of 83.3% (95% CI, 74.3–91.1) and an AUC-PR of 72.8%. At the optimised threshold, the model yielded an F1 score of 72.7%, precision of 73.7%, recall of 71.8%, and specificity of 87.2% (Table 5, Figure 2). In comparison, the clinical-only baseline achieved an AUC-ROC of 64.4% (95% CI, 53.4–75.1) and a specificity of only 14.1%. Although its optimised operating point produced high sensitivity (97.4%), this resulted largely from classifying most patients as CAD-positive, limiting its usefulness as a stand-alone triage rule. The full multimodal model therefore provided a substantially more balanced operating point, retaining similar sensitivity while reducing false-positive classifications.

The nonlinear tabular baselines did not improve on logistic regression. Random forest, gradient boosting, and histogram gradient boosting achieved AUC-ROC values of 63.0%, 60.3%, and 57.9%, respectively, compared with 64.4% for logistic regression (Table 3). Thus, within this cohort, increasing the flexibility of the tabular classifier did not provide additional discrimination from the 11 clinical variables. The full multimodal model remained above all tabular baselines, with the same significant AUC difference observed against logistic regression.

Table 3 Nonlinear tabular baselines under the identical five-fold out-of-fold protocol (117 patients).

Tabular model	AUC-ROC (%) [95% CI]
Logistic regression	64.4 [53.4–75.1]
Random forest	63.0 [51.4–73.9]
Gradient boosting	60.3 [48.8–71.0]
Histogram gradient boosting	57.9 [46.7–69.1]

To verify that the imaging models capture information beyond gross tissue quantity, we evaluated a simple volumetric baseline using only depot volume and mean/SD Hounsfield-unit attenuation within each mask (seven features total, no deep learning; Table 4). This baseline performed at chance level (AUC-ROC 49.1%, 95% CI 37.7–60.9), and combining it with clinical variables provided no meaningful improvement

Table 4 Volumetric/Hounsfield-unit baseline using depot volume and mean/SD attenuation only (no deep learning), under the identical five-fold out-of-fold protocol (117 patients).

Baseline	AUC-ROC (%) [95% CI]
Volumetric/HU only	49.1 [37.7–60.9]
Volumetric/HU + Clinical	56.1 [44.0–67.7]
Clinical only (logistic regression)	64.4 [53.4–75.1]

over the clinical-only model (56.1%, 95% CI 44.0–67.7). This indicates that the discriminative signal captured by the deep imaging models does not reduce to simple measures of adipose tissue quantity or density, and instead reflects spatial or textural patterns accessible only through the learned representations.

Table 5 Model performance on the MultiD4CAD dataset (117 patients), strict out-of-fold aggregation, five-fold cross-validation with three-seed ensembling.

Model	AUC-ROC (%) [95% CI]	F1 (%)	Prec. (%)	Recall (%)	Spec. (%)
Full Multimodal	83.3 [74.3–91.1]	72.7	73.7	71.8	87.2
PAT Only	80.3 [71.2–88.2]	63.8	48.1	94.9	48.7
EAT Only	75.9 [66.5–84.5]	63.3	52.5	79.5	64.1
Imaging Only	75.7 [66.1–84.3]	60.3	45.5	89.7	46.2
Clinical Only	64.4 [53.4–75.1]	52.8	36.2	97.4	14.1

Full Multimodal = EAT+PAT+Clinical

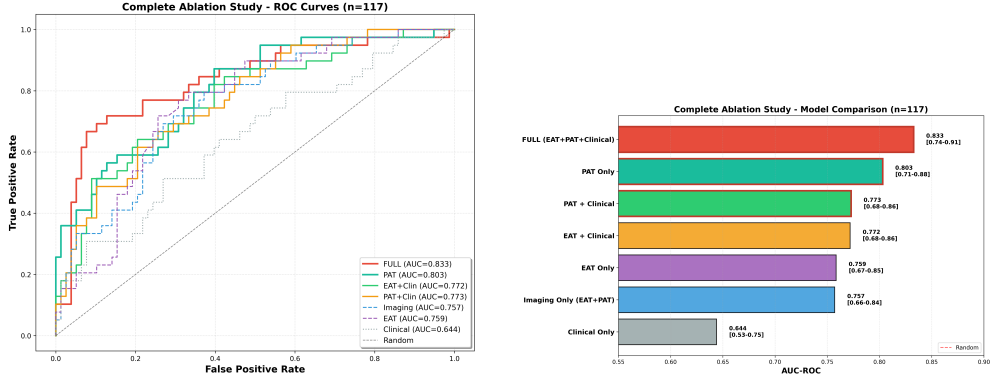


Fig. 2 Ablation study (n=117). Left: ROC curves for all model configurations. Right: AUC-ROC values with 95% bootstrap CIs.

2.2 Depot Comparison and Incremental Value

PAT-only and EAT-only models achieved AUC-ROC values of 80.3% and 75.9%, respectively, with overlapping confidence intervals (Table 5, Figure 3). The difference between the two depots was therefore not statistically significant. The full multimodal model achieved an AUC improvement of 18.9 percentage points over the clinical-only baseline (95% CI, 7.6–30.6), which was the only comparison in which the improvement reached statistical significance under the prespecified analysis. Single-depot and depot-plus-clinical configurations also showed higher point estimates than the clinical-only baseline, but adding a second information source did not produce a statistically significant improvement at this cohort size. These results suggest that the benefit of multimodal integration was not attributable to either depot alone, although the small sample limits the ability to distinguish complementary signal from sampling variability.

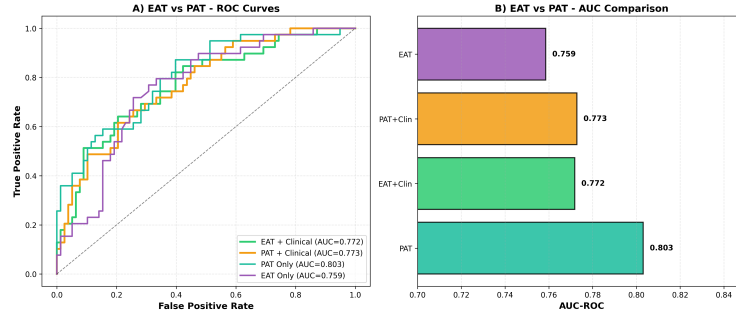


Fig. 3 ROC comparison between EAT and PAT models and summary of AUC-ROC values for related configurations.

2.3 Calibration and Clinical Utility

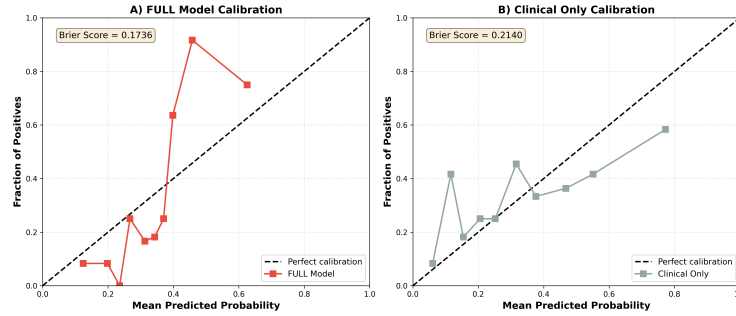


Fig. 4 Calibration curves for the full multimodal model (EAT+PAT+Clinical) and tabular-only baseline. Mean predicted probability (x-axis) against observed event rate (y-axis).

The full multimodal model showed a lower Brier score than the clinical-only baseline (0.174 vs 0.214). This provides preliminary evidence of improved probabilistic calibration, although the reliability diagram showed visible variability across predicted-probability ranges (Figure 4). Because Platt scaling was fitted and evaluated on the same OOF predictions, these calibration results should be interpreted as exploratory rather than as an unbiased estimate of prospective calibration.

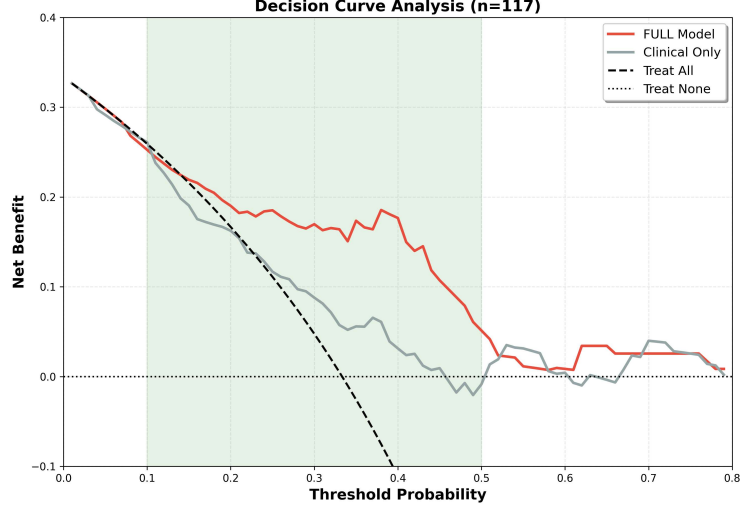


Fig. 5 Decision curve analysis comparing net benefit across threshold probabilities for the full multimodal model versus the clinical-only baseline.

In decision curve analysis, the full model showed higher net benefit than the clinical-only model and the treat-all and treat-none strategies across the 10–50% threshold range (Figure 5). At a 20% threshold, the net benefit was 0.190 for the full model compared with 0.162 for the clinical-only baseline. These findings suggest potential decision-analytic value, although confirmation in an independent cohort is required.

2.4 Robustness and Stability

Performance was relatively consistent across folds, with AUC-ROC values ranging from 77.5% to 90.6% and a mean of $83.4 \pm 4.4\%$ (Table 6). The agreement between the fold-level estimates and the aggregated OOF result suggests that the overall performance was not attributable to a single fold. A permutation test based on 1,000 label shuffles yielded $p < 0.001$, with a maximum permuted AUC of 0.708 compared with the observed AUC of 0.833 (Figure 6A).

Harrell bootstrap optimism correction produced negligible estimated optimism (-0.0001) (Figure 6B). This result provides no evidence of substantial optimism under the applied resampling procedure, although it should not be interpreted as proof that overfitting is absent in a 117-patient cohort. The sensitivity analysis across

$N \in \{16, 24, 32\}$ instances per bag produced AUC-ROC values between 78.7% and 83.3% (Table 7), indicating that the choice of $N = 24$ did not determine the observed performance.

Table 6 Per-fold performance of the full multimodal model (out-of-fold predictions).

Metric	F1	F2	F3	F4	F5	Mean $\pm$ SD
AUC-ROC (%)	85.2	90.6	82.1	81.7	77.5	83.4 $\pm$ 4.4
Specificity (%)	93.7	87.5	68.7	93.3	86.7	86.0 $\pm$ 10.2

Table 7 Sensitivity of the full model to the number of sampled instances per bag (N).

N	AUC-ROC (%) [95% CI]
16	79.8 [71.0–87.7]
24	83.3 [74.3–91.1]
32	78.7 [68.9–87.6]

The age-stratified analysis also produced similar discrimination after splitting the cohort at the median age of 60 years, with AUC-ROC values of 82.7% and 83.6% in the two strata. This provides some evidence that the overall result was not explained solely by the age imbalance between the CAD-positive and CAD-negative groups, although the limited number of patients within each stratum warrants cautious interpretation.

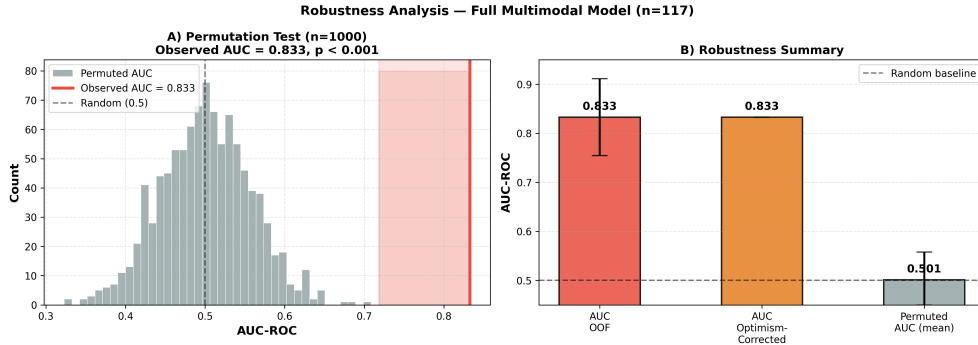


Fig. 6 Robustness analysis. (A) Permutation test: observed AUC of 0.833 exceeds all 1,000 permuted AUC values ($p < 0.001$). (B) Negligible optimism (-0.0001) from bootstrap correction indicates no detectable overfitting despite the limited cohort size.

2.5 Architectural Ablation

Gated attention produced the highest AUC-ROC (83.3%), compared with 79.3% for mean pooling and 76.0% for max pooling; bootstrap testing did not identify a statistically significant difference between gated attention and either pooling variant (Table 8, Figure 2). The learned attention weights were close to uniform, with an inter-slice standard deviation below 0.004. This behaviour is consistent with an aggregation regime that approaches mean pooling, suggesting that the discriminative information was distributed across the sampled instances rather than concentrated in a small number of slices.

The weight-sharing backbone variant used 11.4M parameters, approximately half the parameter count of the default dual-backbone configuration, while achieving the same AUC-ROC of 83.3%. The 22.6M-parameter dual-backbone architecture is large relative to a 117-patient cohort; the fact that an 11.4M-parameter shared-backbone model matched its performance indicates that the additional parameterisation was not associated with a measurable discrimination gain in this dataset, and that the observed gain was more closely related to the combination of imaging and clinical information than to maintaining separate backbone parameters for the two imaging streams. Given this parameter efficiency at no measurable cost to discrimination, we consider the shared-backbone configuration the preferred choice for cardiac imaging cohorts of comparable or smaller size, and recommend it as the default configuration in future applications of this pipeline; the dual-backbone configuration is retained in the remaining analyses only for consistency with the ablation protocol.

Table 8 Architectural ablation of aggregation strategy and backbone sharing (Δ AUC relative to gated attention MIL, based on 5,000 bootstrap iterations).

Model	AUC-ROC (%)	Params	Δ AUC (%)
Gated Attention MIL (ours)	83.3	22.6M	ref.
Mean Pooling MIL	79.3	22.6M	−4.0
Max Pooling MIL	76.0	22.6M	−7.3
Weight-Sharing Backbone	83.3	11.4M	±0.0
None of the pairwise Δ AUC bootstrap confidence intervals excluded zero.			

2.6 Error Analysis

Using the fold-specific decision thresholds, the model produced 30 true positives, 9 false negatives, 67 true negatives, and 11 false positives, corresponding to an overall accuracy of 82.9%. False positives were concentrated in one cross-validation fold, with 5 of 11 occurring in the fold with the lowest specificity (68.7%). They were also more frequently observed among male patients (11/11 vs 68.7% among correctly classified CAD-negative patients, Fisher’s exact test, $p = 0.030$) and current smokers (81.8% vs 43.3%, $p = 0.023$). These comparisons were exploratory and were not

adjusted for multiple testing. The observed association with smoking may reflect differences in inflammatory or adipose-tissue characteristics, but this interpretation remains speculative and would require validation in a larger cohort.

False-negative patients were older on average than correctly classified CAD-positive patients (60.3 vs 51.0 years), although the difference did not reach statistical significance ($p = 0.092$) (Table 9). This pattern is consistent with the age imbalance observed between the CAD groups but does not establish age as a cause of model failure.

One representative false-negative case had diabetes, hypertension, and hypercholesterolaemia and received a moderate clinical-only predicted probability of 0.219, but the lowest full-model probability in the cohort (0.079). This example illustrates that the imaging component can substantially modify a clinically plausible risk estimate rather than simply adding information to the clinical baseline (Table 9).

Table 9 Clinical characteristics of misclassified versus correctly classified patients.

Characteristic	FN (n=9)	Correct CAD+ (n=30)	FP (n=11)	Correct CAD− (n=67)
Age, years	60.3	51.0	64.9	63.6
Male, n (%)	7 (77.8)	20 (66.7)	11 (100)	46 (68.7)
Smoking, n (%)	4 (44.4)	17 (56.7)	9 (81.8)	29 (43.3)

FN: false negatives, FP: false positives.

3 Discussion

This pilot study systematically separates the contribution of multimodal data integration from that of architectural design choices for CAD detection from cardiac adipose tissue. Within this cohort, differences between information configurations were more apparent than differences between the architectural variants examined: every architectural variant (aggregation strategy, encoder sharing) overlapped zero in bootstrap comparison, whereas the full multimodal model was the only configuration to significantly outperform the clinical-only baseline. This observation is consistent with the near-uniform attention distributions observed across instances (inter-slice standard deviation < 0.004), which imply that the gated-attention operator behaved close to a mean-pooling regime in this cohort: discriminative information appears distributed across multiple sampled adipose-tissue regions rather than concentrated in a small number of anatomically localised instances. Whether this reflects a genuine biological property of diffuse peri-vascular inflammation, or simply the limited statistical power to detect a sharper attention pattern at $n = 117$, cannot be resolved with the present cohort size and would benefit from validation against expert anatomical annotations in future, larger studies. This interpretation is further supported by the volumetric baseline (Table 4), which performed at chance level using depot volume and attenuation alone; the imaging models’ discrimination therefore cannot be explained by gross

tissue quantity or density, and must instead derive from spatial or textural patterns learned by the convolutional encoders.

Table 10 situates this study against representative prior approaches. Reported CAD/cardiovascular-event prediction performance in the literature ranges from AUC 72–76% for fat-attenuation-index-based single-depot analysis in a larger single-centre cohort [2], to AUC 88.1% for a machine-learning model combining clinical variables with coronary artery calcium scoring in a much larger multicentre registry [1], to AUC 82.0% for a radiomics-plus-clinical pipeline evaluated on a closely related MultiD4CAD cohort of similar size [6]. Direct comparison across these studies is limited by heterogeneity in datasets, endpoints, and populations, and reported metrics should be read as contextual benchmarks rather than head-to-head comparisons. Within this context, the present multimodal MIL approach reaches comparable discrimination to the closest-matched radiomics baseline [6] while additionally reporting calibration, decision-analytic, and error-level characterisation seldom provided alongside discrimination metrics in this domain, and while removing the hand-crafted feature-engineering step required by radiomics pipelines.

Table 10 Representative prior work on adipose tissue analysis for CAD/cardiovascular risk prediction.

Study	Method	Tissue	n	AUC (%)	Validation
Antonopoulos et al. [2]	Fat attenuation index	PCAT	273	72–76	Single-centre
Al’Aref et al. [1]	ML + calcium score	–	13,054	88.1	Internal
Militello et al. [6]	Radiomics + ML	PCAT	118	82.0	Nested CV
This study	Multimodal MIL	EAT+PAT	117	83.3	5-fold OOF

The error analysis provides an additional layer of interpretability that goes beyond aggregate discrimination metrics. The concentration of false positives within a single cross-validation fold suggests that the markedly lower specificity observed in that fold reflects a subset of borderline cases rather than a systematic weakness of the model. The association between false positives and smoking status is biologically plausible, potentially reflecting smoking-related inflammation of adipose tissue, but remains exploratory given the small subgroup size and the absence of correction for multiple comparisons. Nevertheless, this finding motivates targeted evaluation in a larger, stratified cohort. The case in which the imaging component overrode a clinically plausible risk signal highlights a specific failure mode: active disagreement between imaging and clinical information, rather than simple non-contribution. Such cases would not be detected by aggregate metrics alone and would need to be flagged for human review in any clinically deployed system.

Workflow integration considerations.

A practical integration of this pipeline into CCTA reporting would require the inference step to run on segmentation masks already produced during routine EAT/PAT

quantification, adding a probability score and its associated calibration uncertainty to the existing report rather than requiring a separate acquisition or manual annotation step. The error-analysis findings suggest concrete points at which such a system would need human oversight rather than autonomous decision-making: patients with borderline model probabilities, or cases in which the imaging-derived prediction conflicts with the broader clinical context, such as a strong smoking history, where the model’s output should be interpreted alongside, rather than instead of, clinical judgement. Given the exploratory calibration and the single-centre cohort, any prospective workflow evaluation would need to treat model output as a triage or second-reader aid rather than a stand-alone diagnostic decision, consistent with current practice for computer-aided detection tools in cardiac CT.

Limitations.

All experiments derive from a single-centre cohort of 117 patients; multicentre external validation is the natural next step. Segmentation masks were expert-provided rather than automatic, and their generalisability to noisier, automatically generated masks is untested. CAD-positive patients were on average ten years younger than controls, a characteristic of the recruitment protocol; although the clinical-only model’s low AUC and the preserved discrimination within age strata argue against age alone driving the imaging result, residual confounding cannot be excluded given the observational design. Attention weights have not been validated against expert anatomical annotations and should be interpreted as relative contributions to the learned representation rather than validated localisation. The volumetric baseline used only a small set of first-order attenuation statistics and total depot volume; it does not rule out that more elaborate hand-crafted radiomic features (e.g. texture or shape descriptors) could partially explain the deep models’ performance, a comparison left for future work. Calibration parameters were fitted and evaluated on the same out-of-fold predictions, so calibration metrics (unlike discrimination metrics, unaffected by this monotonic transformation) may be mildly optimistic. The CAD label itself relies on expert visual assessment of plaque biomarkers rather than an independent angiographic or invasive reference standard, so the present results should be read as detection of a radiologist-defined phenotype rather than of obstructive CAD confirmed by invasive coronary angiography or fractional flow reserve; future studies would benefit from cross-referencing against such a reference standard where available. The 22.6M-parameter dual-backbone architecture, while shown to be reducible to an 11.4M-parameter shared-backbone variant without measurable performance loss, remains substantial relative to the 117-patient cohort, and the negligible bootstrap optimism reported here should not be over-interpreted as a guarantee against over-fitting in cohorts with different case-mix or class balance. Decision thresholds and early-stopping checkpoints for the full model were selected using each held-out validation fold’s own predictions rather than a separate inner validation split; this may introduce mild optimism in threshold-dependent metrics (F1, precision, recall, specificity). The primary discrimination metric, AUC-ROC, is unaffected by this issue, as it

does not depend on a decision threshold. Finally, the exploratory error-analysis associations should be regarded as hypothesis-generating pending replication in a larger cohort.

4 Conclusion

In this pilot cohort, combining paired EAT and PAT imaging with structured clinical variables improved CAD discrimination relative to the clinical-only baseline within a weakly supervised MIL framework. The multimodal configuration was the only comparison showing a statistically significant performance improvement, whereas the architectural variations examined, including aggregation strategy and backbone sharing, did not produce significant differences. A volumetric/attenuation-only baseline performed at chance level, indicating that the deep imaging models capture spatial or textural information beyond gross tissue quantity or density. These findings suggest that, within this dataset, multimodal information contributed more to the observed performance differences than the specific architectural refinements evaluated. On the basis of these results, we recommend the parameter-efficient shared-backbone configuration as the preferred design for future work on cardiac imaging cohorts of similar size, since it matches the discrimination of the larger dual-backbone model with roughly half the parameters.

The results should be considered a feasibility benchmark rather than evidence of clinical readiness. The single-centre design, limited sample size, expert-provided segmentations, and age imbalance between groups constrain generalisability. Future work should focus on independent multicentre validation, evaluation with automated or semi-automated segmentation, and prospective assessment within routine CCTA workflows. A second-reader or triage role would provide a more appropriate initial clinical application than stand-alone diagnosis.

Statements and Declarations

Funding. This research has been funded by the Scientific Research Deanship at University of Ha'il, Saudi Arabia, through project number IRP-24 014.

Competing interests. The authors declare that they have no relevant financial or non-financial interests to disclose.

Ethics approval. This study is a secondary analysis of the controlled-access MultiD4CAD dataset. The original data collection was approved by the Ethics Committee of the University Hospital Policlinico "P. Giaccone" of Palermo, Italy (minutes N. 8/2024), and written informed consent was waived owing to the retrospective nature of the original study. As the present work used only previously collected, de-identified data, no additional ethics approval was required.

Consent to participate. Informed consent was waived for the original retrospective data collection, as documented in the MultiD4CAD data descriptor.

Consent to publish. Not applicable. The present study used de-identified data and contains no identifiable individual information.

Data availability. The MultiD4CAD dataset analysed in this study is third-party, controlled-access data distributed via Zenodo (<https://doi.org/10.5281/zenodo>).

15063615) and described in its data descriptor. Access is granted by the data providers under their terms, which require registration and approval. To support reproducibility, the training and evaluation code, preprocessing scripts, model configuration files, fold assignments, and a synthetic toy example that reproduces the pipeline without restricted patient data are openly available at <https://github.com/haneneguesmi/multimodal-mil-cad>.

Code availability. The code and supporting configuration files are available at <https://github.com/haneneguesmi/multimodal-mil-cad>.

Declaration of generative AI use. Claude (Anthropic) was used to assist with language editing and readability in selected sections. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the scientific content, including experimental design, data analysis, and conclusions.

Author contributions. Hanene Guesmi led the conceptualization, methodology, software, formal analysis, original draft preparation, and funding acquisition. Haifa Harrouch contributed to methodology, software, formal analysis, validation, resources, data curation, and original draft preparation. Karim Gasmi contributed to validation and to review and editing. Muna O. Alamoudi contributed to resources, data curation, and review and editing. All authors contributed to visualization, read the manuscript, and approved the submitted version.

References

- [1] Al'Aref SJ, Maliakal G, Singh G, et al (2020) Machine learning of clinical variables and coronary artery calcium scoring for the prediction of obstructive coronary artery disease on coronary computed tomography angiography: analysis from the CONFIRM registry. *European Heart Journal* 41(3):359–367
- [2] Antonopoulos AS, Sanna F, Sabharwal N, et al (2017) Detecting human coronary inflammation by imaging perivascular fat. *Science Translational Medicine* 9(398):eaal2658
- [3] Chen M, Hao G, Xu J, et al (2024) Radiomics analysis of lesion-specific pericoronary adipose tissue to predict major adverse cardiovascular events in coronary artery disease. *BMC Medical Imaging* 24(1):150
- [4] Iacobellis G (2015) Local and systemic effects of the multifaceted epicardial adipose tissue depot. *Nature Reviews Endocrinology* 11(6):363–371
- [5] Ilse M, Tomczak J, Welling M (2018) Attention-based deep multiple instance learning. In: *Proceedings of the 35th International Conference on Machine Learning*, pp 2127–2136
- [6] Militello C, Prinzi F, Sollami G, et al (2023) CT radiomic features and clinical biomarkers for predicting coronary artery disease. *Cognitive Computation* 15:238–253

- [7] Oikonomou EK, Marwan M, Desai MY, et al (2018) Non-invasive detection of coronary inflammation using computed tomography and prediction of residual cardiovascular risk (the CRISP CT study). *The Lancet* 392(10151):929–939
- [8] Prinzi F, Militello C, Sollami G, et al (2025) MultiD4CAD: multimodal dataset composed of CT and clinical features for coronary artery disease analysis. *Scientific Data* 12(1):1570
- [9] Roth GA, Mensah GA, Johnson CO, et al (2020) Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. *Journal of the American College of Cardiology* 76(25):2982–3021
- [10] Vickers AJ, Elkin EB (2006) Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making* 26(6):565–574
- [11] Yang W, Ding X, Yu Y, et al (2024) Long-term prognostic value of CT-based high-risk coronary lesion attributes and radiomic features of pericoronary adipose tissue in diabetic patients. *Clinical Radiology* 79(12):931–940
- [12] Zhan W, Luo Y, Luo H, et al (2024) Predicting major adverse cardiovascular events in angina patients using radiomic features of pericoronary adipose tissue based on CCTA. *Frontiers in Cardiovascular Medicine* 11:1462451
- [13] Zou Q, Qiu T, Liang C, et al (2025) Multimodal prediction of major adverse cardiovascular events in hypertensive patients with coronary artery disease. *La Radiologia Medica* 130(6):767–781