MEDA: Metric-Conditioned Decision Adaptation for Multi-Target CT Recognition

Dharini Raghavan[1], Amritpal Singh[2], and Anant Madabhushi[1,2]

[1] Georgia Institute of Technology, Atlanta, GA, USA
[2] Emory University, Atlanta, GA, USA
draghavan7@gatech.edu

Abstract. While modern CT foundation models can learn strong shared representations, the performance gap comes from how those features are translated into decisions for downstream tasks. Different CT abnormalities have different statistical properties, class imbalance, and separability. A single classifier and threshold cannot optimally serve all tasks. Additionally, good ranking performance (e.g. AUROC) does not guarantee good decisions, as inappropriate thresholds can result in poor binary predictions. Target-specific operating-point selection can bridge part of this gap and improves balanced accuracy for most targets. Since it is not always possible to fine-tune the foundation model, foundation models should be adapted at the decision layer, not necessarily the feature layer. MEDA is a lightweight, target-aware framework, with decision layers optimised for multiple clinical endpoints, shifting the focus from representation learning to decision adaptation. MEDA acts as an adaptation layer that sits on top of any pretrained CT foundation model. A shared encoder extracts representations once, and for each target, MEDA automatically chooses both the best classifier and the best operating threshold, producing a decision rule tailored to that clinical task. On a 17-target CT validation benchmark spanning abdominal and chest findings, MEDA achieves 0.738 mean balanced accuracy and 0.746 mean AUROC. Target-specific operating-point calibration increases mean balanced accuracy from 0.635 to 0.738 and improves 15 of 17 targets. These results support a decision-centric view of CT representation adaptation: for heterogeneous multi-target recognition, the downstream rule should be selected with respect to the evaluation metric rather than fixed independently of the target. Our code is available at https://github.com/rdharini2001/MEDA

Keywords: CT representation learning · Multi-target classification · Decision calibration · Threshold selection · Metric-conditioned adaptation

1 Introduction

Recent advances in CT foundation models have demonstrated that a single pretrained encoder can learn reusable volumetric representations applicable to a wide range of downstream clinical tasks. By leveraging large-scale self-supervised,

vision-language, or multimodal pretraining, these models have substantially reduced the need to train task-specific feature extractors, making representation learning increasingly general-purpose. Consequently, the central challenge is no longer how to learn better CT features, but how to most effectively adapt these representations to diverse downstream endpoints.

Existing adaptation strategies typically treat the downstream prediction layer as a fixed component. A shared representation is evaluated using a standard linear probe or lightweight classifier, often followed by a default decision threshold. This implicitly assumes that all clinical targets share a common decision interface and differ only in classifier parameters. However, multi-target CT recognition violates this assumption. Different abnormalities exhibit distinct class imbalance, score distributions, separability regimes, and operating requirements. A decision rule that is well suited for detecting diffuse organ-level abnormalities may be poorly matched to small focal lesions or rare pathological findings, even when both are represented within the same feature space.

This observation highlights an important distinction between representation quality and decision quality. Metrics such as AUROC measure how well a model ranks positive and negative cases, but clinical deployment ultimately requires binary decisions at a specific operating point. A representation may produce highly informative rankings while still yielding poor classification performance if the downstream classifier or decision threshold is mismatched to the target. Improving feature representations alone therefore cannot fully address performance variability across heterogeneous CT tasks.

These limitations are becoming increasingly relevant as CT foundation models grow larger and more general-purpose. In many practical settings, retraining or fine-tuning the encoder is computationally expensive, unavailable, or undesirable due to limited labeled data. Instead, adaptation should occur at the decision layer, where lightweight task-specific decision rules can be learned while preserving the shared representation. This perspective shifts the focus from representation adaptation toward decision adaptation, treating the downstream interface as the primary object of optimization.

In this work, we introduce MEDA, (**ME**tric-Conditioned **D**ecision **A**daptation), a lightweight framework for adapting pretrained CT representations to multiple clinical endpoints. Rather than assuming a single downstream classifier, MEDA constructs a compact family of candidate decision functions and jointly selects both the scoring model and the operating threshold that maximize the target evaluation metric. The resulting target-specific decision rule is optimized for the empirical score geometry of each abnormality while requiring no modification of the underlying encoder. In this way, MEDA, acts as a plug-and-play adaptation layer that can be applied on top of any pretrained CT foundation model.

The paper makes three contributions.

- First, our work frames multi-target CT adaptation as metric-conditioned decision-layer selection, where the learner family and operating threshold are selected for each target over a shared representation.

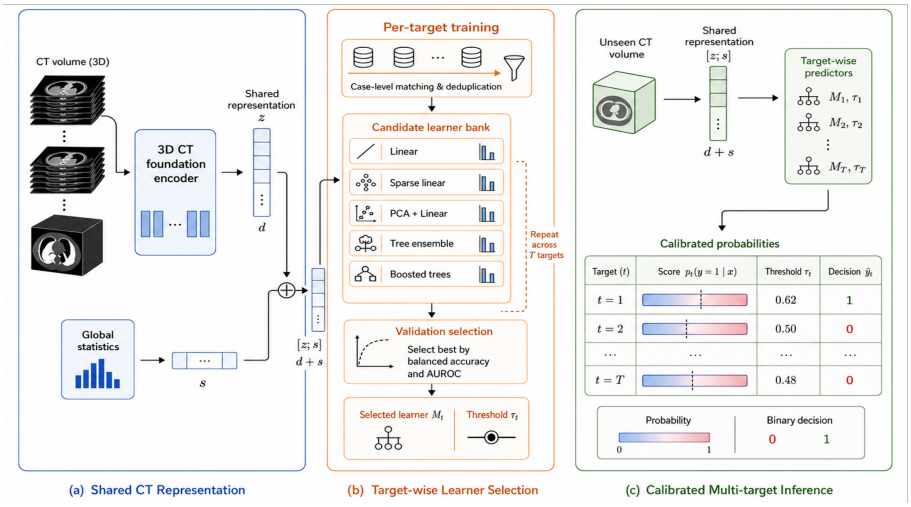

Fig. 1. Overview of MEDA. A 3D CT volume is mapped to a shared volumetric representation. For each target, MEDA constructs a case-level representation matrix, evaluates a compact family of candidate scoring functions, and selects both the learner family and operating threshold under the validation metric. The resulting target-wise rules produce calibrated scores and binary decisions over the same representation space.

- Second, it empirically characterizes target-dependent decision geometry across 17 CT recognition tasks, showing that the selected learner family varies across linear, low-rank, ensemble, and boosted decision rules.
- Third, it quantifies the role of operating-point calibration under balanced-accuracy evaluation: target-specific thresholds increase mean balanced accuracy from 0.635 to 0.738 and improve 15 of 17 targets, while the final method retains a mean AUROC of 0.746.

Across a 17-target benchmark spanning abdominal and chest CT findings, MEDA achieves a mean balanced accuracy of 0.738 and a mean AUROC of 0.746. Target-specific operating-point optimization improves mean balanced accuracy from 0.635 to 0.738, benefiting 15 of 17 clinical endpoints. These results support a decision-centric view of foundation model adaptation, demonstrating that for heterogeneous multi-target recognition, optimizing the downstream decision rule can be as important as learning the underlying representation.

2 Method

2.1 Problem Setup

We consider a multi-target binary classification setting over 3D CT volumes. Each volume x_i is associated with T binary labels $y_i^{(t)} \in \{0, 1\}$, one per clinical

target $t \in \{1, \ldots, T\}$. A shared volumetric encoder f_θ, pretrained (and frozen), maps each volume to a fixed-dimensional representation:

$$z_i = f_\theta(x_i), \qquad z_i \in \mathbb{R}^d. \tag{1}$$

Rather than modifying f_θ, we adapt the decision layer. For each target t, the downstream prediction is defined by two objects: a scoring function $g_t : \mathbb{R}^d \to [0, 1]$ and an operating threshold $\tau_t \in [0, 1]$. The binary decision rule is:

$$h_t(x_i) = \mathbb{1}[g_t(f_\theta(x_i)) \geq \tau_t]\,. \tag{2}$$

The full adapted model is a collection of target-wise decision rules:

$$\mathcal{H} = \{(g_t, \tau_t)\}_{t=1}^{T}\,. \tag{3}$$

This formulation separates representation extraction from decision formation. The encoder defines a common feature space; each target receives an independent decision rule whose inductive bias and operating point are selected from validation data. No encoder updates are required.

2.2 Case-level representation construction

Different clinical targets have different labeled cohorts, class priors, and validation splits. Although all targets share the same frozen encoder, each defines a distinct downstream learning problem. We make this explicit by constructing a target-specific dataset for each endpoint.

For target t, we define:

$$\mathcal{D}_t = \left\{(z_i, y_i^{(t)})\right\}_{i=1}^{N_t}, \tag{4}$$

where $z_i \in \mathbb{R}^d$ is the frozen representation of the i-th CT volume and $y_i^{(t)} \in \{0, 1\}$ is its binary label for target t. Representations are organized into a case-level feature matrix:

$$X_t = \begin{bmatrix} z_1^\top \\ \vdots \\ z_{N_t}^\top \end{bmatrix} \in \mathbb{R}^{N_t \times d}, \qquad y_t = \left[y_1^{(t)}, \ldots, y_{N_t}^{(t)}\right]^\top. \tag{5}$$

Case identifiers are canonicalized before label matching, and duplicate entries are removed so that each CT study contributes exactly one representation. This preserves the case as the unit of statistical analysis and prevents duplicated representations from biasing model selection or validation estimates. Constructing target-indexed matrices in this way allows MEDA to adapt each decision rule to the empirical score geometry of its endpoint, while leaving the shared encoder unchanged.

2.3 Candidate scoring families

Because different targets induce different separability regimes over the same representation space, no single inductive bias is universally appropriate. We therefore define a compact candidate family of scoring functions,

$$\mathcal{G}_t = \{g_{t,1}, \ldots, g_{t,K}\}\,. \tag{6}$$

where each candidate maps a representation to a positive-class score:

$$p_i^{(t,k)} = g_{t,k}(z_i). \tag{7}$$

The family is designed to cover complementary low-capacity inductive biases over the same representation space: class-balanced logistic regression tests direct linear separability; sparse logistic models identify a restricted subset of discriminative axes; low-rank probes via PCA followed by logistic regression suppress nuisance variation before fitting; random forests and ExtraTrees capture nonlinear coordinate interactions through ensembles; and histogram gradient boosting and XGBoost [6] model boosted nonlinear decision rules.

This family does not form a test-time ensemble. It defines a small hypothesis set for target-wise selection. Each target ultimately uses a single scoring function drawn from $\mathcal{G}_t$, chosen to best match the empirical decision geometry of that endpoint.

2.4 Metric-conditioned selection

Let $\mathcal{D}_t^{\mathrm{tr}}$ and $\mathcal{D}_t^{\mathrm{val}}$ denote the training and validation subsets for target t. The evaluation metric for the binary decision is balanced accuracy,

$$\mathrm{BA}(y, \hat{y}) = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\right), \tag{8}$$

which weights sensitivity and specificity equally and is therefore appropriate when class priors differ across targets [7]. We retain AUROC as a complementary threshold-independent ranking measure [8].

A key observation motivates our selection procedure: balanced accuracy is realized only after converting a continuous score into a binary decision. Selecting a scoring function without jointly selecting its operating threshold therefore decouples the optimization from the evaluation metric. We avoid this mismatch by optimizing both objects together. For each candidate $g_{t,k} \in \mathcal{G}_t$, we fit the scorer on $\mathcal{D}_t^{\mathrm{tr}}$ and compute validation scores on $\mathcal{D}_t^{\mathrm{val}}$. The selected decision rule is:

$$(g_t^*, \tau_t^*) \in \arg\max_{g_{t,k}\in\mathcal{G}_t,\ \tau\in\mathcal{T}_{t,k}} \mathrm{BA}\left(y_{\mathrm{val}}^{(t)}, \mathbb{1}[g_{t,k}(z_{\mathrm{val}}) \geq \tau]\right). \tag{9}$$

where $\mathcal{T}_{t,k}$ is the threshold set associated with the validation scores of candidate $g_{t,k}$. When candidates achieve comparable validation balanced accuracy, AUROC and cross-validated balanced accuracy serve as secondary diagnostics. This joint selection ensures that the downstream decision rule is conditioned on the evaluation metric, not fixed independently of it.

2.5 Operating-point calibration

The threshold τ_t^* is estimated directly from validation scores rather than fixed at a conventional value. For the selected scoring function g_t^*, the candidate threshold set combines a uniform probability grid with empirical score quantiles:

$$\mathcal{T}_{t,k} = \mathcal{U} \cup \mathcal{Q}\Big(\Big\{g_{t,k}(z_i) : (z_i, y_i^{(t)}) \in \mathcal{D}_t^{\text{val}}\Big\}\Big)\,, \tag{10}$$

where $\mathcal{U}$ is a uniform grid over $[0,1]$ and $\mathcal{Q}(\cdot)$ denotes empirical quantiles of the validation scores. The uniform grid provides stable coverage of the probability interval; the quantile candidates adapt the search to the score distribution induced by each target and learner family.

The final binary decision rule for target t is:

$$h_t^*(x) = \mathbb{1}[g_t^*(f_\theta(x)) \geq \tau_t^*]\,. \tag{11}$$

After model selection, the selected learner family is refit using the available labeled cases for that target and paired with the calibrated threshold for final inference. The result is a compact target-wise decision layer $\{(g_t^*, \tau_t^*)\}_{t=1}^T$ over a shared CT representation, requiring no modification to the underlying encoder

3 Experiments

3.1 Evaluation setting

We evaluate MEDA in a multi-target CT classification setting with $T = 17$ binary endpoints spanning abdominal and chest CT. The benchmark includes targets with different visual support and score geometry, including organ-scale findings, localized high-attenuation findings, distributed abnormalities, cystic findings, lesion-level decisions, and malignancy-risk estimation. This diversity allows us to test whether a shared volumetric representation can be converted into reliable target-specific decisions using a metric-conditioned adaptation layer.

The target set consists of adrenal hyperplasia, ascites, atherosclerosis, cholecystitis, colorectal cancer, COVID-19, fatty liver, gallstone, hydronephrosis, kidney stone, liver calcifications, liver cyst, liver lesion, lung nodule malignancy, lymphadenopathy, renal cyst, and splenomegaly. Across the target-wise validation splits, we evaluate 1,898 validation instances, with 949 positive and 949 negative examples after target-wise validation construction. The abdominal subset contains 15 targets and the chest subset contains 2 targets.

3.2 Implementation details

The encoder f_θ is instantiated by a 3D vision-language CT foundation model pretrained on paired CT volumes, radiology reports, and structured clinical supervision [1]. Representations are extracted once per case and shared across all targets.

Table 1. Aggregate validation performance grouped by region. Cross-validation metrics are computed on the training partition of each target and averaged across targets.

Region	Targets	N_{val}	Pos.	BalAcc	AUROC	CV-BalAcc	CV-AUROC
Abdomen	15	1634	817	0.742	0.750	0.676	0.713
Chest	2	264	132	0.708	0.715	0.645	0.656
Overall	17	1898	949	0.738	0.746	0.672	0.706

Candidate scoring functions are trained independently for each target. Linear and sparse logistic models use feature standardization; low-rank probes apply PCA after standardization. Tree-based models operate directly on the representation vector. Logistic models use class-balanced weighting where supported. For each target, candidates are fit on the training split and evaluated on the validation split. The selected scorer and threshold jointly maximize validation balanced accuracy, with AUROC and cross-validated balanced accuracy used as tiebreakers.

We report balanced accuracy and AUROC as complementary measures. Balanced accuracy evaluates the thresholded binary decision and is therefore the primary metric for operating-point selection. AUROC evaluates ranking quality independent of threshold. We additionally report cross-validated balanced accuracy and cross-validated AUROC for the selected learners, the distribution of selected learner families, and the effect of target-wise operating-point calibration relative to the default threshold of 0.5.

3.3 Aggregate performance

Table 1 reports validation performance averaged within each anatomical subset and across all targets. MEDA achieves 0.738 mean balanced accuracy and 0.746 mean AUROC over the full 17-target benchmark. The abdominal subset reaches 0.742 balanced accuracy and 0.750 AUROC, while the chest subset reaches 0.708 balanced accuracy and 0.715 AUROC. The chest aggregate contains only two targets, so it is interpreted as a domain-specific diagnostic rather than as a stable estimate of chest-wide performance.

Table 2 compares MEDA with the CT-NEXUS reference baseline. MEDA improves the overall mean balanced accuracy by 3.9 points and mean AUROC by 3.1 points. The improvement is observed in both the abdominal and chest subsets, supporting the use of target-specific decision adaptation over a shared representation.

3.4 Target-wise decision rules

Table 3 reports the selected learner and validation metrics for each target. The strongest balanced accuracy is observed for splenomegaly, fatty liver, atherosclerosis, lymphadenopathy, and COVID-19. The lowest scores occur for liver calcifications and lung nodule malignancy, two targets where the relevant evidence

Table 2. Comparison with the CT-NEXUS reference baseline. Values are mean validation scores over targets within each region. CT-NEXUS EAO was used as baseline model.

Region	Targets	Method	BalAcc	AUROC
Abdomen	15	baseline	0.702	0.721
		MEDA	**0.742**	**0.750**
Chest	2	baseline	0.670	0.670
		MEDA	**0.708**	**0.715**
Overall	17	baseline	0.699	0.715
		MEDA	**0.738**	**0.746**

Table 3. Target-wise validation performance for MEDA. The table includes validation metrics and cross-validation diagnostics for the selected learner. Positive counts are reported as train/validation.

Region	Target	N_{tr}	N_{val}	Pos. tr/val	Head	BalAcc	AUROC	CV-BalAcc	CV-AUROC
Abdomen	Adrenal hyperplasia	935	70	85/35	PCA-LR (8)	0.743	0.744	0.650	0.674
	Ascites	1107	68	101/34	PCA-LR (8)	0.779	0.824	0.842	0.901
	Atherosclerosis	1107	68	81/34	ExtraTrees	0.838	0.876	0.719	0.773
	Cholecystitis	792	102	132/51	XGB	0.686	0.659	0.645	0.665
	Colorectal cancer	1107	52	82/26	XGB	0.769	0.735	0.613	0.629
	Fatty liver	774	78	129/39	XGB	0.885	0.949	0.781	0.854
	Gallstone	640	116	160/58	LR	0.681	0.673	0.601	0.616
	Hydronephrosis	576	64	144/32	PCA-LR (64)	0.781	0.812	0.757	0.801
	Kidney stone	897	158	299/79	PCA-LR (64)	0.684	0.651	0.664	0.703
	Liver calcifications	822	72	137/36	PCA-LR (128)	0.583	0.525	0.555	0.557
	Liver cyst	632	244	316/122	PCA-LR (32)	0.672	0.690	0.598	0.615
	Liver lesion	627	134	209/67	PCA-LR (64)	0.664	0.669	0.667	0.710
	Lymphadenopathy	1107	72	111/36	XGB	0.806	0.797	0.639	0.680
	Renal cyst	796	266	398/133	RF	0.665	0.704	0.660	0.690
	Splenomegaly	1107	70	71/35	PCA-LR (64)	0.886	0.936	0.742	0.825
Chest	COVID-19	308	64	60/32	LR	0.797	0.811	0.709	0.726
	Lung nodule mali.	1000	200	100/100	PCA-LR (256)	0.620	0.620	0.582	0.586

may occupy a small fraction of the full volume or require fine-grained lesion localization.

The selected learner family varies across targets. PCA-logistic models are selected for 9 of 17 targets, XGBoost for 4 targets, balanced logistic regression for 2 targets, and random forest or ExtraTrees for one target each. This variation is the main empirical signature of target-dependent decision geometry: different endpoints favor different low-capacity scoring functions even when the underlying representation is shared.

3.5 Candidate-family analysis

Table 4 summarizes performance across all candidate learners evaluated within each family. ExtraTrees and random forests obtain the strongest average family-level balanced accuracy, while the final selected heads are dominated by PCA-logistic and XGBoost models. This distinction is expected: average robustness across all targets and optimality for a particular target are different criteria. A family can perform consistently across many targets, while another can provide the best score-to-decision mapping for a smaller subset of endpoints.

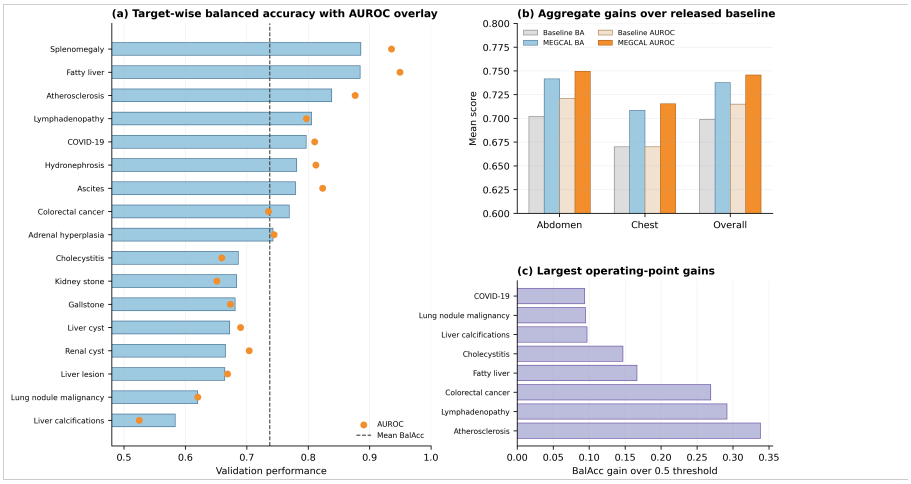

Fig. 2. Validation behavior across targets. (a) Target-wise balanced accuracy with AUROC overlay. (b) Aggregate comparison against the reference baseline. (c) Largest gains from target-specific operating-point calibration relative to the default threshold of 0.5.

3.6 Effect of operating-point calibration

Table 5 isolates the contribution of target-specific operating-point calibration. Applying the default threshold gives 0.635 mean balanced accuracy. Selecting the threshold on validation scores increases mean balanced accuracy to 0.738, for a gain of 0.103. Calibration improves 15 of 17 targets. The selected thresholds range from 0.023 to 0.790, showing that the score scale varies substantially across targets and learner families.

The largest gains from operating-point calibration occur for atherosclerosis (+0.338), lymphadenopathy (+0.292), colorectal cancer (+0.269), fatty liver (+0.167), and cholecystitis (+0.147). These targets share the same qualitative pattern: the scoring function contains useful ranking information, but the default threshold is poorly aligned with the validation metric. The result supports the central empirical claim of MEDA: in multi-target CT recognition, representation quality must be evaluated together with the metric-conditioned rule that converts scores into binary decisions.

4 Discussion

The experiments highlight the importance of treating multi-target CT recognition as a decision-adaptation problem over a shared representation. The first observation is that target-wise learner selection captures meaningful variation in downstream geometry. Different endpoints select different low-capacity scoring families: COVID-19 is selected by a balanced linear model, atherosclerosis

Table 4. Candidate-family analysis across validation reports. Scores are averaged over all candidates evaluated within each family.

Family	Candidates	BalAcc	AUROC	CV-BalAcc
ExtraTrees	17	0.700	0.718	0.680
Random forest	17	0.699	0.717	0.677
XGBoost	68	0.696	0.711	0.671
Sparse logistic	68	0.692	0.708	0.668
HistGradientBoosting	17	0.691	0.708	0.663
PCA-logistic	101	0.689	0.703	0.660
Balanced logistic	102	0.682	0.692	0.654

Table 5. Effect of target-wise operating-point calibration for the selected learners.

Quantity	Value Interpretation
Mean BalAcc at default threshold	0.635 fixed operating point
Mean BalAcc after calibration	0.738 target-wise operating point
Mean absolute BalAcc gain	0.103 calibrated minus default
Targets improved by calibration	15/17 positive gain in BalAcc
Mean selected threshold	0.318 average target-wise threshold
Threshold range	0.023–0.790 score-scale variation across targets

by ExtraTrees, renal cyst by a random forest, several abdominal targets by low-rank logistic probes, and fatty liver, colorectal cancer, cholecystitis, and lymphadenopathy by boosted trees. This variation indicates that a shared volumetric representation does not induce a single separability regime across all targets. Some endpoints are well described by nearly linear directions, others benefit from low-rank stabilization, and others require nonlinear coordinate interactions. The selected learner family therefore serves as an empirical probe of target-dependent decision geometry.

The second observation is that ranking quality and binary decision quality are distinct. AUROC measures the ordering induced by a scoring function, while balanced accuracy is realized only after selecting an operating point. In our experiments, several selected learners produce useful score rankings but perform poorly under the default threshold. Target-wise operating-point calibration corrects this mismatch by selecting the threshold under the same validation metric used for evaluation. The increase from 0.635 to 0.738 mean balanced accuracy, with improvement in 15 of 17 targets, shows that threshold estimation is a central component of metric-conditioned adaptation rather than a cosmetic post-processing step.

The third observation concerns the reliability of representation-based evaluation. In multi-target settings, each endpoint can have its own valid case subset, class prior, and validation split. Constructing target-indexed case-level matrices with deduplication preserves the statistical unit of analysis and prevents repeated feature entries from changing the empirical distribution seen during

learner selection. This step is methodologically simple but important: when representations are precomputed, validation estimates can be affected by the way cases are matched, deduplicated, and assigned to target-specific learning problems.

The main limitation is that MEDA operates on global scan-level representations. This design keeps the adaptation layer compact and makes the decision-rule analysis clean, but it can dilute evidence that occupies a small region of the volume. The weakest target-wise results occur for liver calcifications and lung nodule malignancy, both of which may require more localized evidence than a whole-scan representation can reliably expose. Future work should combine metric-conditioned decision adaptation with organ-conditioned representations, lesion-centered feature extraction, token-level pooling, and calibration procedures evaluated under center- or protocol-level distribution shift. These extensions would preserve the decision-centric formulation while giving the scoring functions access to more spatially focused information.

5 Conclusion

We introduced MEDA, a metric-conditioned adaptation framework for multi-target CT classification. The method converts a shared volumetric representation into target-specific decision rules by selecting both a scoring function and an operating threshold for each endpoint. Across 17 validation targets, MEDA achieved 0.738 mean balanced accuracy and 0.746 mean AUROC. Operating-point calibration increased mean balanced accuracy by 0.103 over the default threshold and improved 15 of 17 targets. These results support a decision-centric view of CT representation learning: reusable features should be evaluated together with the metric-conditioned rule that transforms scores into target-specific predictions.

Acknowledgements The authors thank the benchmark organizers and data contributors for releasing the evaluation infrastructure and baseline resources. Computational experiments were supported by Georgia Tech research computing resources.

Disclosure of Interests. The authors declare no competing interests relevant to this work.

References

1. Blankemeier, L., Cohen, J.P., Kumar, A., Van Veen, D., Gardezi, S.J.S., Paschali, M., Chen, Z., Delbrouck, J.-B., Reis, E., Truyts, C., et al.: Merlin: A vision language foundation model for 3D computed tomography. arXiv:2406.06512 (2024)
2. Hamamci, I.E., et al.: Developing generalist foundation models from a multimodal dataset for 3D computed tomography. arXiv:2403.17834 (2024)

3. Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., et al.: AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In: Advances in Neural Information Processing Systems, Dataset and Benchmark Track (2022)
4. Yang, X., He, X., Zhao, J., Zhang, Y., Zhang, S., Xie, P.: COVID-CT-Dataset: A CT scan dataset about COVID-19. arXiv:2003.13865 (2020)
5. Peeters, D., et al.: LUNA25: lung nodule malignancy risk estimation on screening CT. Zenodo (2025)
6. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
7. Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M.: The balanced accuracy and its posterior distribution. In: International Conference on Pattern Recognition (2010)
8. Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic curve. Radiology 143(1), 29–36 (1982)
9. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning, pp. 1321–1330. PMLR (2017)
10. Hernández-Orallo, J., Flach, P., Ferri, C.: A unified view of performance metrics: Translating threshold choice into expected classification loss. Journal of Machine Learning Research 13, 2813–2869 (2012)