AbdCTBench: Learning Clinical Biomarker Representations from Abdominal Surface Geometry

Anonymous authors
Paper under double-blind review

Abstract

Body composition analysis through CT and MRI imaging provides critical insights for cardiometabolic health assessment but remains limited by accessibility barriers including radiation exposure, high costs, and infrastructure requirements. We present AbdCTBench, a large-scale dataset containing 23,506 CT-derived abdominal surface meshes from 18,719 patients, paired with 87 comorbidity labels, 31 specific diagnosis codes, and 16 CT-derived biomarkers. Our key insight is that external surface geometry is predictive of internal tissue composition, enabling accessible health screening through consumer devices. We establish comprehensive benchmarks across seven computer vision architectures (ResNet-18/34/50, DenseNet-121, EfficientNet-B0, ViT-Small, Swin Transformer-Base), demonstrating that models can learn robust surface-to-biomarker representations directly from 2D mesh projections. Our best-performing models achieve clinically relevant accuracy: age prediction with MAE 6.22 years (R^2=0.757), mortality prediction with AUROC 0.839, and diabetes (with chronic complications) detection with AUROC 0.801. Notably, smaller architectures consistently matched or surpassed larger models, while medical-domain pretraining (RadImageNet) and self-supervised pretraining (DINOv2) showed competitive but not superior performance. AbdCTBench represents the largest publicly available dataset bridging external body geometry with internal clinical measurements, enabling future research in accessible medical AI. We plan to release the dataset, evaluation protocols, and baseline models to accelerate research in representation learning for medical applications, immediately following the review period.

1 Introduction

Body composition analysis has emerged as a critical avenue for advancing preventive and diagnostic medicine, offering valuable insights into cardiometabolic health (Amato et al., 2013; Rosenquist et al., 2013). While traditional metrics such as body mass index (BMI) and waist circumference are widely used, they fail to differentiate between metabolically active visceral adipose tissue (VAT), intramuscular fat infiltration, and organ-specific pathologiesall critical for cardiometabolic risk stratification (Lee et al., 2018; Sweatt et al., 2024; Therkelsen et al., 2013). This limitation has led to the adoption of advanced imaging biomarkers derived from computed tomography (CT) and magnetic resonance imaging (MRI), which provide quantitative assessments of tissue composition with unprecedented precision (Thomas et al., 2025).

However, the clinical utility of these modalities remains limited by accessibility barriers. CT exposes patients to ionizing radiation, precluding repeated use, while MRI is costly and has limited availability. Both modalities require specialized infrastructure and trained radiologists, creating bottlenecks in resource-constrained settings and perpetuating health disparities. These limitations underscore the need for alternative approaches that provide access to clinically useful biomarkers while overcoming accessibility barriers.

To address this challenge, we present AbdCTBench, the first-of-its-kind dataset of 2D surface meshes derived from formerly conducted abdominal CT scans. AbdCTBench is carefully curated to enable development of computer vision techniques for learning representations from surface ge-

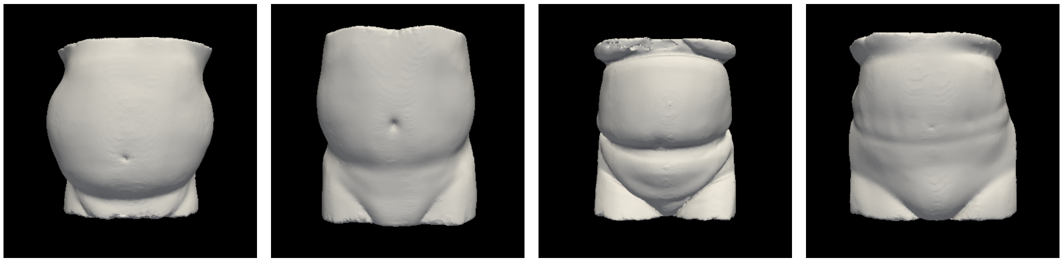


Figure 1: Sample 2D abdominal surface meshes from AbdCTBench dataset. These CT-derived surface geometries demonstrate the range of external anatomical features used to predict internal body composition biomarkers without radiation exposure. Complete biomarker details for these images are provided in the Appendix A.1

ometry to predict a variety of clinically useful biomarkers. We fine-tune an array of state-of-the-art models and benchmark their performance on the task of predicting the associated biomarkers. Our models are trained on detailed internal body composition data from CT scans and reflect the feasibility of effectively capturing the predictive signal by using only external body measurements and surface geometry at inference. Model predictions can be obtained simply by providing 2D surface mesh images as input, aimed at bridging the gap between high-precision clinical imaging and widely accessible consumer technology.

By validating against the gold standard of CT-derived biomarkers, AbdCTBench can facilitate the development of reliable models that ultimately eliminate the need for CT scans in routine screening and monitoring through surface mesh imagery. Recent breakthroughs in consumer-grade depth sensing, such as LiDAR-enabled devices like the iPhone, offer viable alternatives for body composition assessment (Oberhofer et al., 2024; Zamotsin et al., 2022; Boczar et al., 2024; Vasic et al., 2024) and provide accessible methods for generating the surface meshes required by these models. While early implementations struggle with complex torso geometries (Galaaoui et al., 2025), iterative scanning protocols and machine learning-based mesh reconstruction algorithms are rapidly closing the fidelity gap with clinical CT. Meanwhile, imaging foundation models trained on extensive radiographic data demonstrate the feasibility of extracting biomarkers from various imaging modalities. When applied to 2D surface scans, such models could learn associations between external surface geometry and internal composition patterns.

Our research represents a transitional step in this evolution with two main contributions:

1. We curate AbdCTBench, a dataset of 23,506 CT-derived abdominal surface mesh images from 18,719 unique patients. To our knowledge, AbdCTBench is the first and largest publicly available dataset of its kind. Our Dataset Release Statement is provided in Appendix A.2.
2. We benchmark a variety of computer vision architectures on predicting CT-derived ground truth biomarkers available in AbdCTBench, requiring models to learn the relationship between external abdominal geometry and internal CT-derived biomarkers.

Once validated, these models could operate solely on surface-derived 2D meshes without requiring CT imaging. This approach could transform biomarker accessibility by providing individuals, clinicians, and researchers with actionable physiological insights through non-invasive, radiation-free surface scans, potentially enabling broad, low-cost, and scalable health screening tools for early disease risk detection.

2 Related Work

2.1 Medical Image Analysis Benchmarks

Medical imaging benchmarks have played an important role in advancing computer vision methods for healthcare. Large-scale datasets such as ImageNet inspired challenges (e.g. CheXpert (Irvin et al., 2019), MIMIC-CXR (Johnson et al., 2019)) have provided valuable testbeds for algorithm development and reproducibility. These resources are typically derived from high-fidelity imaging modalities and clinician-verified ground truth labels such as radiologist reports or disease codes. While these benchmarks have accelerated progress in medicine, they remain tied to modalities like

CT, MRI, or X-ray that require specialized infrastructure and are often inaccessible for population-scale screening.

In contrast, external body shape analysis has primarily been explored in non-clinical domains such as human pose estimation (Cao et al., 2019) or anthropometric studies. To date, no benchmark has systematically linked external abdominal surface geometry with internal, CT-derived biomarkers at scale. AbdCTBench fills this gap by bridging the rigor of medical imaging benchmarks with the accessibility of surface-based imaging, enabling a new class of methods for non-invasive biomarker assessment.

2.2 Architecture Evaluation in Medical Imaging

Benchmarking diverse neural architectures has been central to medical image analysis research. Early convolutional neural networks (CNNs) demonstrated success on tasks such as tumor segmentation (Ronneberger et al., 2015), while later work has shown the advantages of transformers (Dosovitskiy et al., 2021a; Chen et al., 2021a) and hybrid CNN-transformer architectures for capturing global context in medical images. More recently, foundation models pretrained on massive radiology corpora (e.g., RadImageNet (Mei et al., 2022a)) have highlighted the benefits of transfer learning for downstream tasks.

Existing architectural studies, however, primarily benchmark models on imaging modalities that capture internal anatomy directly (CT, MRI, X-ray). By contrast, AbdCTBench evaluates the ability of architectures to infer internal body composition from external abdominal surface meshes. This task differs fundamentally from conventional medical imaging because the predictive signal is indirect, requiring models to learn associations between geometry and physiology. As such, AbdCTBench provides a new arena to assess whether architectural advances – spanning CNNs, transformers, and emerging vision models – generalize to this novel, indirect inference problem.

3 AbdCTBench Dataset

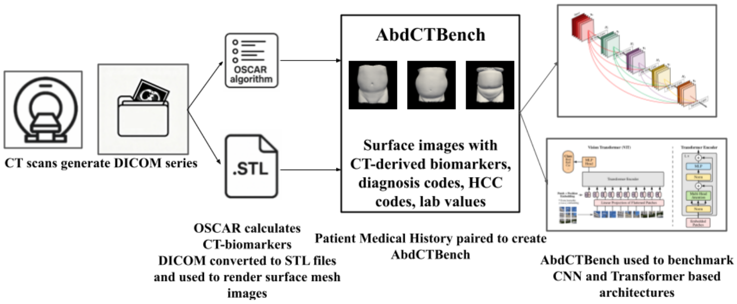

CT scans generate DICOM series
OSCAR algorithm
.STL
OSCAR calculates CT-biomarkers DICOM converted to STL files and used to render surface mesh images
AbdCTBench
Surface images with CT-derived biomarkers, diagnosis codes, HCC codes, lab values
Patient Medical History paired to create AbdCTBench
AbdCTBench used to benchmark CNN and Transformer based architectures

Figure 2: AbdCTBench dataset overview showing the pipeline from CT scans to surface mesh extraction and biomarker prediction.

3.1 Dataset Collection and Curation

AbdCTBench is a comprehensive dataset derived from 23,506 abdominal CTs of 18,719 patients ($\approx$ 1.26 scans per patient), representing one of the largest CT-derived biomarker datasets for abdominal composition analysis. The data was collected from all available CT scans conducted at facilities of a privately-owned healthcare company. This included CT scans conducted from August 11, 2003, to September 9, 2021, under IRB approval from the University of Wisconsin (Protocol: "Opportunistic CT Screening"). The utilization of all available CT scans allowed the creation of the largest possible dataset available, but may introduce implicit biases in the data which we discuss further in section 7. Following collection, processing proceeded in two parallel phases: surface mesh rendering and CT-derived biomarker calculation.

For surface mesh rendering, DICOM image series were converted to stereo-lithography (STL) files, then to 2D PNG images of size 384 x 384 (Figure 1) via PyVista (Sullivan & Kaszynski, 2019). The conversion pipeline consists of three sequential stages: volume processing applies optional shrinking and anisotropic smoothing to prepare the data; surface extraction generates 3D triangular meshes using VTK's contour filter, which are refined through mesh cleaning and smoothing operations before being exported as binary STL files; and finally, the STL to 2D image conversion renders each mesh into standardized PNG images with fixed camera positioning and mesh smoothing. Further technical details about the DICOM to STL and STL to PNG conversions are provided in the Appendix A.3. For biomarker calculation, DICOM series were processed by OSCAR (Pickhardt et al., 2020), which creates segmentation masks to calculate metrics at vertebral levels (L1-L5, T10-T12) and organ-specific regions (liver, spleen, kidneys, aorta). Key measurements include bone mineral density, adipose tissue distribution, muscle composition, organ volumes, and calcium scoring metrics, yielding 16 biomarkers measuring body composition at various anatomical levels.

The CT-derived biomarkers were then paired with 31 diagnosis (ICD-10) codes, 87 hierarchical condition category (HCC) comorbidity labels (groupings of ICD-10 codes (Amerigroup, 2019)) and 2 longitudinal lab values (HbA1c and C-reactive protein) from patient medical history. The dataset was then processed for HIPAA Safe Harbor (U.S. Department of Health and Human Services, Office for Civil Rights, 2012) compliance, removing PII for safe public release, and we performed a 70/20/10 split at the patient ID level for train, validation, and test sets to prevent data leakage. All hyperparameter tuning and model selection procedures used only the train and validation sets, and the test set was held out only to perform the final evaluation for the reported results. The resulting dataset integrates quantitative CT biomarkers with clinical outcomes, comorbidity codes, and demographic information for comprehensive cardiometabolic health analysis and architecture benchmarking.

3.2 Dataset Statistics and Analysis

AbdCTBench represents a diverse patient population with mean age 55.3 years (SD: 16.51). For HIPAA compliance, ages 90 were categorized as (90+), with 298 such cases excluded from age statistics. The dataset shows balanced sex distribution (56.8% female, 43.2% male) and significant clinical diversity, with high prevalence of: essential hypertension (53.7%), Type 2 Diabetes (44.6%), impaired glucose tolerance (38.0%), tobacco use (26.8%), and MI (23.1%).

HCC comorbidity burden analysis reveals substantial heterogeneity, with patients carrying an average of 1.8 HCC conditions (max: 24 per patient). Most prevalent HCC codes include: HCC 108 (Vascular Disease; 22.6%), HCC 19 (Diabetes without complications; 13.0%), and HCC 12 (Breast, Prostate, and other Cancers; 10.9%). Overall mortality rate is 11.4%.

Primary CT-derived biomarkers include: Calcium Scoring Abdominal Agatston score (mean: 1200.9 ± 3126.5), kidney median Hounsfield units (mean: 90.0 ± 58.8HU), spleen median Hounsfield units (mean: 82.1 ± 37.2HU), spleen volume (mean: 223.9 ± 127.2cm^3), and comprehensive adipose tissue analysis at vertebral levels (L1-L5, T10-T12). Comprehensive dataset statistics are provided in the Appendix A.4.

4 AbdCTBench Benchmark

From AbdCTBench, we curate 10 biomarker prediction tasks from 2D surface mesh images. We design a single-target learning framework to benchmark selected architectures on biomarker prediction. The goal is twofold: (i) design a standardized evaluation framework for comparing computer vision architectures across diverse biomarker prediction tasks, and (ii) benchmark foundation models for CT-derived biomarker prediction using surface geometry representations from 2D images. The biomarker prediction tasks are as follows:

- Mortality prediction (11.4%): binary classification for patient death during follow-up.
- HCC-108 (Vascular Disease; 22.6%): binary classification for HCC 108 code at scan time.
- HCC-12 (Breast, Prostate, and other Cancers; 10.9%): binary classification for HCC 12 code at scan time.
- HCC-96 (Cardiac Arrhythmias; 9.0%): binary classification for HCC 96 code at scan time.
- HCC-18 (Diabetes with Chronic Complications; 8.3%): binary classification for HCC 18 code at scan time.
- HCC-111 (COPD; 7.1%): binary classification for HCC 111 code at scan time.

- Calcium Scoring Abdominal Agatston Score: binary classification for score > 1000 (Janjua et al., 2021) (21.2%).
- Myocardial Infarction (MI; 23.1%): binary classification for previous myocardial infarction.
- Type 2 Diabetes (44.6%): binary classification for diabetes at scan time.
- Age (mean 55.3): regression for patient age at scan time.

4.1 Architecture Selection

We selected 6 representative architectures spanning different families to ensure comprehensive coverage of modern computer vision approaches for medical image representation learning. **CNN Architectures:** Four CNN-based models representing different design philosophies:

- **ResNet-18/34** (He et al., 2016): Residual networks with skip connections, for strong baseline performance with efficient parameter usage (4-6GB and 6-8GB GPU memory)
- **DenseNet-121** (Huang et al., 2017): Densely connected networks maximizing feature reuse through concatenation-based connections (8-10GB GPU memory)
- **EfficientNet-B0** (Tan & Le, 2019): Compound scaling approach balancing depth, width, and resolution for optimal efficiency (6-8GB GPU memory)

Vision Transformers: **ViT-Small (DINOv2)** (Oquab et al., 2024) (8-10GB GPU memory) evaluates self-supervised pre-training effectiveness with vision transformers on medical imaging tasks. DINOv2's self-supervised pre-training has shown superior performance compared to supervised ImageNet pre-training on various downstream tasks. On the other hand, **Swin Transformer-Base** (Liu et al., 2021) (12-16GB GPU memory) represents hierarchical vision transformers with shifted window attention, providing an alternative and modern transformer architecture for comparison.

Medical-Specific Architecture: ResNet-50 (RadImageNet) (Mei et al., 2022b) (6-8GB GPU memory) represents domain-specific pre-training, utilizing RadImageNet weights trained on medical images to assess medical domain knowledge benefits.

4.2 Standardized Training Protocol

For fair and reproducible comparison across architectures, we establish a standardized training protocol for all models, designed based on best practices from the medical imaging literature.

Optimization Configuration: All models use the AdamW optimizer (Loshchilov & Hutter, 2019) with weight decay 1×10^{-4} and cosine annealing learning rate scheduling. AdamW is the optimizer of choice for various medical imaging tasks (Chang, 2024), (Mortazi et al., 2023). We evaluate three learning rates: 1×10^{-5}, 1×10^{-4}, and 1×10^{-3} to capture different optimization regimes and ensure robust performance across architectures.

Training Hyperparameters: Batch size of 16 balances memory efficiency with gradient stability. All models train for 100 epochs with early stopping based on validation performance (patience: 10 epochs). Dropout of 0.2 prevents overfitting, as established in medical imaging literature (Wang & Huang (2024), Maruyama et al. (2025), Adebayo & et al. (2025)).

Loss Function: We standardized across all models to use binary cross-entropy loss with logits for binary classification tasks, and mean squared error loss for regression tasks.

Fine-tuning Strategy: All models use full fine-tuning to maximize performance. Pre-trained weights are loaded from ImageNet (CNN models), DINOv2 (ViT), or RadImageNet (medical-specific ResNet-50), with final classification layers replaced with task-specific heads.

Data Augmentation: A standardized augmentation pipeline is applied to all models using pre-training weights with slight modifications for medical imaging tasks. For ImageNet/DINOv2 pre-trained models: random horizontal flips (p=0.3), geometric augmentations including random rotations (±7.5) and less aggressive crops (0.9-1.0 original size) with aspect ratio 0.8-1.2 (p=0.6). For grayscale images, reduced intensity color augmentations (halved brightness/contrast shift, p=0.4) are applied. Images are then converted to 3-channel by repeating tensors along channel dimension. For normalization, ImageNet models use ImageNet mean (0.485, 0.456, 0.406) and std (0.229, 0.224, 0.225); DINOv2 models use CT-derived means (0.55001191, 0.55001191, 0.55001191) and stds (0.18854326, 0.18854326, 0.18854326) (Pyrros et al., 2023). For RadImageNet (medical imaging specific), augmentations are more conservative: random horizontal flips (p=0.2, reduced from 0.3), random rotations (±5, reduced from ±7.5), less aggressive crops (0.95-1.0 original size, reduced from 0.9-1.0), narrower aspect ratio (0.9-1.1, reduced from 0.8-1.2), applied with p=0.4 (instead of 0.6). Color augmentations: color jitter for brightness/contrast with reduced intensity 0.3 (instead of 0.5), applied with p=0.3 (instead of 0.4). Images are then similarly converted to 3-channel and normalized using ImageNet mean/std as recommended by Mei et al. (2022b).

4.3 Single-Target Learning Framework and Class Imbalance Handling

Our benchmark focuses on single-target learning to establish clear performance baselines for biomarker prediction tasks, allowing direct comparison of architectures without multi-task learning dynamics.

Architecture Adaptation: CNN models use direct classification heads; Vision Transformers employ CLS token classification. For each, the final layer is replaced with a task-specific head outputting the appropriate number of classes or continuous values.

Evaluation Strategy: Each architecture is evaluated independently on each biomarker task, allowing comprehensive analysis of architectural strengths across different prediction types. This approach provides clear insights into which architectures excel at specific biomarker prediction types.

Further, as described in section 3, the dataset exhibits severe class imbalance, significantly impacting model performance and evaluation. We implement all the strategies described below to address this challenge as a standard training protocol across all models and architectures:

Inverse Frequency Weighting: Applied to the loss function, where class weight is calculated as the inverse of class frequency in the dataset.

Balanced Batch Sampling: During training, balanced batch sampling ensures each batch contains approximately equal representation from all classes, preventing model domination by majority classes.

Threshold Optimization: For binary classification tasks, classification thresholds are optimized using F1-score on the validation set, searching the range [0.1, 0.9] with 9 discrete steps. This ensures optimal performance for imbalanced datasets where default 0.5 thresholding may be suboptimal.

5 Experimental Setup

All experiments were implemented in PyTorch with CUDA 12.4 support. We utilized the timm library for Vision Transformer implementations and torchvision for CNN architectures. Experiments are conducted on NVIDIA GeForce RTX 2080 Ti GPUs (11GB memory each) with 10 GPUs available for parallel execution. Each model is allocated sufficient memory based on expected requirements (4-10GB range), ensuring no memory-related performance degradation.

5.1 Evaluation Metrics

Our evaluation framework employs task-specific metrics for comprehensive assessment of model performance across different biomarker types:

Binary Classification Metrics: AUROC (area under the receiver operating characteristic curve, providing threshold-independent performance assessment), F1-Score, Precision/Recall, and Specificity. Given clinical relevance, the best model is identified based on highest AUROC. Since F1-Score, Precision, Recall, and Specificity are threshold dependent, the F1-score optimizing threshold is selected from the validation set, then used to calculate performance metrics on the test set.

Regression Metrics: Mean Absolute Error (MAE) and Mean Squared Error (MSE). Since age prediction is our only regression task, the best model is selected based on lowest MAE, allowing interpretation of prediction quality in terms of absolute difference in years from true patient age.

Statistical Significance: Results are reported with bootstrapped 95% confidence intervals to assess reliability of performance differences. The bootstrapping was performed with 1000 samples each of size equal to the test set, using simple random sampling with replacement. Samples may be excluded in case of low label diversity (e.g. both classes not present for the binary classification tasks). The 95% confidence intervals were calculated using the percentile method from the distribution of bootstrap statistics for each metric of interest.

6 Results and Analysis

We evaluated single-biomarker predictors across the six architectures, using the best checkpoints selected on validation and report test set performance. For the regression task (age), EfficientNet-B0 achieved the best error (MAE 6.22), with ResNet-50 (RadImageNet, MAE 6.34) and ViT-Small (MAE 6.47) close behind. All models substantially outperformed a naive baseline ($R^2 > 0.719$, reported in Appendix A.5), indicating strong representation learning from abdominal surface geometry for age estimation. For non-HCC binary targets, AUROC varied by biomarker and architecture:

- Calcium Scoring Abdominal Agatston: ResNet-34 achieved the best AUROC of 0.848, with DenseNet-121 at 0.847
- Myocardial infarction (MI): Swin Transformer-Base achieved the best AUROC of 0.742, with EfficientNet-B0 at 0.732 (reported in table 1) trailing closely
- Mortality Prediction: ResNet-18 with an AUROC of 0.839 was the best performing. EfficientNet-B0 at 0.830, Swin transformer-Base at 0.828 close behind
- Type 2 Diabetes (T2D): ResNet-34 achieved the best AUROC of 0.742, with EfficientNet-B0 and Swin Transformer-Base at 0.740

Table 1: Results for non-HCC biomarkers by architecture on the test set. AUROC is reported for the binary classification tasks and MAE is reported for Age prediction (regression task). Bootstrapped 95% CIs are reported in parentheses.

Architecture	Age (MAE)	Calcium Score	MI	Mortality	T2D
Naive Baseline	13.16 (12.79–13.57)	0.500 —	0.500 —	0.500 —	0.500 —
DenseNet-121	6.769 (6.551–6.994)	0.847 (0.829–0.863)	0.730 (0.703–0.752)	0.823 (0.800–0.845)	0.728 (0.709–0.750)
EfficientNet-B0	**6.223** (6.016–6.422)	0.847 (0.829–0.864)	0.732 (0.708–0.754)	0.830 (0.805–0.852)	0.740 (0.720–0.761)
ResNet-18	6.472 (6.264–6.678)	0.843 (0.825–0.859)	0.729 (0.705–0.752)	**0.839** (0.816–0.861)	0.735 (0.714–0.756)
ResNet-34	6.486 (6.284–6.692)	**0.848** (0.831–0.864)	0.731 (0.706–0.753)	0.825 (0.799–0.848)	**0.742** (0.722–0.762)
ResNet-50 (RadImageNet)	6.341 (6.154–6.532)	0.833 (0.815–0.849)	0.716 (0.693–0.739)	0.810 (0.784–0.834)	0.733 (0.714–0.753)
ViT-Small (DINOv2)	6.465 (6.260–6.684)	0.829 (0.809–0.846)	0.732 (0.707–0.754)	0.811 (0.785–0.836)	0.735 (0.714–0.755)
Swin Transformer-Base	6.540 (6.338–6.758)	0.845 (0.828–0.862)	**0.742** (0.718–0.763)	0.828 (0.803–0.851)	0.740 (0.720–0.759)

- HCC-108 (Vascular Disease): best AUROC of 0.768 as achieved by Swin Transformer-Base, with ResNet-18 at 0.763, and EfficientNet-B0 at 0.753
- HCC-111 (Chronic Obstructive Pulmonary Disease): best AUROC of 0.769 again achieved by ResNet-18, with Swin Transformer-Base trailing at 0.765
- HCC-12 (Breast, Prostate, and other Cancers): ResNet-34 achieved the best AUROC of 0.591. Most performance numbers clustered near chance-level (0.571–0.591) across architectures. We believe this reflects both limited biological plausibility and label design: HCC-12 aggregates multiple, heterogeneous cancer types with differing and often weak relationships to obesity and abdominal body composition. Further, the timing of the HCC code relative to the CT scan can span pre-diagnosis, active treatment, and long-term survivorship. Taken together, these factors likely attenuate any subtle signal from the surface geometry, consistent with our interpretation that external abdominal surface geometry is primarily predictive of cardio-metabolic and musculoskeletal biomarkers rather than oncologic comorbidities
- HCC-18 (Diabetes with Chronic Complications): best AUROC of 0.801 was achieved by Swin Transformer-Base. It is noteworthy that this AUROC is higher than the highest AUROC of 0.742

Table 2: Results for HCC code biomarkers by architecture on the test set. All biomarkers report AUROC. Bootstrapped 95% CIs are reported in parentheses.

Architecture	HCC-108	HCC-111	HCC-12	HCC-18	HCC-96
Naive Baseline	0.500 —	0.500 —	0.500 —	0.500 —	0.500 —
DenseNet-121	0.740 (0.720–0.760)	0.716 (0.681–0.749)	0.587 (0.551–0.622)	0.766 (0.735–0.797)	0.757 (0.723–0.787)
EfficientNet-B0	0.753 (0.732–0.775)	0.747 (0.708–0.782)	0.586 (0.550–0.623)	0.789 (0.760–0.816)	0.763 (0.732–0.790)
ResNet-18	0.763 (0.742–0.784)	**0.769** (0.734–0.802)	0.580 (0.544–0.614)	0.799 (0.770–0.828)	0.760 (0.727–0.791)
ResNet-34	0.749 (0.729–0.769)	0.766 (0.730–0.799)	**0.591** (0.557–0.624)	0.775 (0.746–0.805)	0.728 (0.694–0.761)
ResNet-50 (RadImageNet)	0.718 (0.698–0.739)	0.739 (0.704–0.774)	0.571 (0.536–0.607)	0.782 (0.754–0.812)	0.738 (0.705–0.769)
ViT-Small (DINOv2)	0.743 (0.723–0.764)	0.760 (0.727–0.793)	0.576 (0.542–0.610)	0.785 (0.755–0.817)	0.757 (0.721–0.788)
Swin Transformer-Base	**0.768** (0.749–0.788)	0.765 (0.732–0.796)	0.580 (0.545–0.616)	**0.801** (0.776–0.828)	**0.770** (0.739–0.798)

for Type 2 Diabetes prediction as achieved by ResNet-34. This difference may be reflective of the greater predictive capacity of representations from surface geometry for diabetes with chronic complications more so than simply the presence of Type 2 Diabetes

- HCC-96 (Cardiac Arrhythmias): best AUROC of 0.770 was achieved by Swin Transformer-Base with EfficientNet-B0 at 0.763, ResNet-18 at 0.760, and ViT-Small at 0.757

Across biomarkers, smaller-to-midsized CNNs (ResNet-18/34, EfficientNet-B0) consistently matched or surpassed larger ResNet-50 models. ViT-Small with DINOv2 pretraining showed competitive performance, often ranking in the top 2-3 architectures but not achieving the best results on any biomarkers. On the other hand, Swin Transformer-Base achieved the best performance on several biomarkers (MI, HCC-108, HCC-18, HCC-96).

Classification accuracies tracked AUROC and generally fell in the 0.59–0.83 range depending on task difficulty, while F1-scores were modest due to class imbalance and conservative thresholds used during testing. Detailed results with all metrics are provided in Appendix A.5.

Taken together, Tables 1 and 2 highlight that diverse architectures, including the newly added Swin Transformer-Base, can learn discriminative representations from abdominal surface geometry to predict clinically meaningful biomarkers, and we expect that scaling to larger or more specialized backbones will further improve performance.

6.1 Architectural Family Analysis

Within architectural families, we observed that ResNet-18 often led on biomarkers (HCC-108, HCC-111, HCC-18, and Mortality), whereas ResNet-34 performed the best on Calcium Score and Type 2 Diabetes. ResNet-50 (RadImageNet) underperformed ResNet-18/34 on most biomarkers despite greater capacity and domain specific pre-training. EfficientNet-B0 was best on Age (regression) and multiple other binary tasks (HCC-96, MI), showing strong accuracy/efficiency trade-offs. DenseNet-121 was consistently strong but rarely the best, as the model ranked near the top on Calcium Score and Mortality. ViT-Small with DINOv2 pre-training showed competitive performance across all tasks, but never achieving the top result. This transformer architecture demonstrated robustness, but its self-attention mechanism could not surpass CNN baselines on this medical imaging benchmark, which may potentially be caused by the specific pre-training. On the other hand, Swin Transformer-Base achieved the best AUROC on several biomarkers (MI, HCC-108, HCC-18, HCC-96)

We believe that the task of predicting internal biomarkers from external abdominal surface geometry fundamentally differs from conventional medical imaging. As noted in Section 2.2, AbdCTBench requires learning ”associations between geometry and physiology” from an indirect predictive signal. CNNs’ inductive bias for local feature extraction through convolutional operations appears particularly well-suited for this task, as surface geometry patterns that correlate with internal biomarkers likely manifest as local spatial features (e.g., subtle curvature variations, adipose tissue distribution patterns) rather than global dependencies.

This aligns with established findings that CNNs excel at capturing local spatial patterns through their translation-equivariant convolutional operations (Dosovitskiy et al., 2021b), whereas vision transformers, while powerful for global context, may require more data or architectural modifications to effectively capture fine-grained local features. Since Swin Transformer-Base (which incorporates hierarchical local attention through shifted windows) achieved best performance on several supports this interpretation, as Swin's architecture explicitly balances local and global feature extraction.

6.2 Domain Specialization Impact

ResNet-50 models initialized from RadImageNet did not outperform lighter architectures trained with standard initializations. For example, on Mortality prediction, the RadImageNet pre-trained ResNet-50 reached an AUROC of 0.810, lagging ResNet-18 (0.839) and EfficientNet-B0 (0.830). Similarly, on HCC tasks, ResNet-50 results (e.g., 0.718 on HCC-108, 0.738 on HCC-96) trailed the best smaller models. While RadImageNet pre-training is domain-specific, the difference between AbdCTBench and typical CT-scans, MRIs, or X-rays may be crucial: AbdCTBench captures the abdominal surface geometry derived from CT-scans rather than the raw CT imagery. Consequently, models with standard initializations consistently outperform the RadImageNet initialized ResNet-50.

ViT-Small with DINOv2 pretraining, while competitive, also did not outperform the best CNN architectures. This indicates that both medical-domain pretraining in larger backbones and general vision transformer pretraining are not sufficient by themselves to overcome optimization and generalization benefits offered by smaller, more regularized networks in this specific setting.

6.3 Threshold-dependent Metrics

We report threshold-derived metrics at fixed operating points used during testing (often 0.8–0.9) in Appendix A.5. Under these settings, several biomarkers exhibit high recall but low precision (e.g., HCC-18 with ResNet-18: recall 0.93 vs. precision 0.15), or conversely high specificity with moderate recall (e.g., Calcium Score with ResNet-34: specificity 0.758, recall 0.773). This underscores the importance of task-specific threshold selection: the same classifier can trade precision and recall substantially without changing AUROC. We reiterate that for each task and architecture, we select the F1-optimal threshold from the validation set, and use the same threshold to report metrics from the test set. Complete threshold-dependent metrics with 95% CIs are reported in Appendix A.5.

6.4 Subgroup Analysis

To understand how model performance varies across patient demographics, we conducted subgroup analyses stratified by gender. For each biomarker, we performed the subgroup analysis by using the best-performing model (prior to the addition of Swin Transformer-Base experiments) on the test set. These analyses reveal important heterogeneity in predictive performance that may inform clinical deployment strategies. For age prediction, models achieved substantially better performance in male patients (MAE 5.76, R^2 = 0.81) compared to female patients (MAE 6.63, R^2 = 0.70), suggesting that abdominal surface geometry may be more predictive of age in males. This difference may reflect biological variations in body composition changes with age between genders. For binary classification tasks, gender differences varied by biomarker. Males showed superior performance on several cardiovascular and respiratory conditions: Calcium Score (AUROC 0.858 vs. 0.838), MI (AUROC 0.724 vs. 0.699), and HCC-111 (AUROC 0.802 vs. 0.740). In contrast, females demonstrated better performance on HCC-18 (AUROC 0.824 vs. 0.773) and Mortality prediction (AUROC 0.844 vs. 0.831).

These gender-stratified differences likely reflect genuine biological variation in how body composition relates to health outcomes. Sex differences in fat distribution (gynoid vs android adiposity) and age-related changes (menopause effects) are well-documented (Lee et al., 2013; Lovejoy et al., 2008). The improved age prediction in males may reflect more consistent age-related changes in male body composition, while the better diabetes complication detection in females aligns with known sex differences in diabetes presentation and complications (Wells, 2007; Regitz-Zagrosek, 2012). Importantly, performance remains clinically meaningful across both groups, with no subgroup showing near-chance performance. These findings suggest that while sex-specific model training may yield marginal improvements, a unified model provides robust predictions across demographics.

Table 3: Gender-stratified performance metrics on the test set. Bootstrapped 95% CIs are shown in parentheses.

Biomarker	Best Model	Overall	Male (47.8 %)	Female (52.2%)	Difference (Female - Male)
Age (MAE)	EfficientNet-B0	6.223	5.76 (5.51–6.05)	6.63 (6.34–6.93)	+0.87
Calcium Score	ResNet-34	0.848	0.858 (0.835–0.879)	0.838 (0.812–0.861)	-0.020
MI	EfficientNet-B0	0.732	0.724 (0.691–0.756)	0.699 (0.661–0.736)	-0.025
Mortality	ResNet-18	0.839	0.831 (0.801–0.860)	0.844 (0.809–0.877)	+0.013
T2D	ResNet-34	0.742	0.740 (0.710–0.769)	0.743 (0.715–0.771)	+0.003
HCC-108	ResNet-18	0.763	0.763 (0.730–0.794)	0.766 (0.735–0.795)	+0.003
HCC-111	ResNet-18	0.769	0.802 (0.749–0.847)	0.740 (0.691–0.788)	-0.062
HCC-12	ResNet-34	0.591	0.623 (0.575–0.670)	0.552 (0.503–0.603)	-0.071
HCC-18	ResNet-18	0.799	0.773 (0.728–0.814)	0.824 (0.789–0.859)	+0.051
HCC-96	EfficientNet-B0	0.763	0.756 (0.714–0.794)	0.774 (0.731–0.815)	+0.018

Because chronological age is one of our main prediction targets in AbdCTBench, stratifying performance by age would conflate subgroup effects with task-defined signal. We therefore avoid age-based subgroup reporting and instead use gender, which is independent of the benchmark's prediction targets and has sufficient sample size for reliable analysis.

We performed additional follow-up experiments with Multi-Task Learning as detailed in Appendix A.6. Further, to demonstrate the effectiveness of learning representations from abdominal surface geometry, we present additional analysis by using Gradient-weighted Class Activation Mappings (Grad-CAM) (Selvaraju et al., 2017) in Appendix A.7.

7 Limitations and Future Directions

Our benchmarking focuses on single-biomarker predictors to isolate signal detectability and architectural effects. Extending to multi-target learning (shared encoders, task-specific heads) is a promising direction to leverage inter-target correlations (we present preliminary experiment results in Appendix A.6); given the strength of smaller CNNs here, we hypothesize shared lightweight backbones with calibrated thresholds could improve macro-level performance without sacrificing efficiency. Further, due to computational constraints, we limited our benchmarking to include smaller convolutional neural networks and transformer architectures. While ViT-Small showed competitive performance and Swin Transformer-Base performed the best on a few biomarkers, future work should explore larger vision transformers, different pretraining strategies (e.g., medical-specific self-supervised learning), specialized architectures such as U-Net variants (Zhou et al., 2018; Chen et al., 2021b; Lu et al., 2022; Vasa et al., 2024), and architectural modifications tailored to medical imaging. Additionally, incorporating calibration methods (temperature scaling, focal loss tuning) and uncertainty estimation may yield better decision thresholds.

From the perspective of real-world deployment of these models for low-cost, non-invasive cardiometablic risk assessment, cross-site validation is a key next step, as AbdCTBench was collected from a single site. While we have standardized the dataset curation and benchmarking procedure, protocols to conduct CT-scans may vary slightly across sites, and thus may meaningfully change the surface geometry visible in the corresponding abdominal surface images. With the goal of capturing the most expansive dataset possible from the collection site, the absence of specific inclusion/exclusion criteria (e.g. related to age, sex, race, pre-existing conditions, etc.) may have introduced implicit demographic biases in the dataset. Thus, multi-site evaluation is crucial for generalizability assessment.

Finally, validation of the models using surface geometry captured from consumer-grade devices would take us closer to the goal of widely accessible cardiometablic risk assessment.

REFERENCES

Fatima Adebayo and et al. Lightweight brain tumor segmentation on low-resource systems. *protocols.io*, 2025. URL https://www.protocols.io/view/lightweight-brain-tumor-segmentation-on-low-resour-gz4sbx8wf.

M. C. Amato, V. Guarnotta, and C. Giordano. Body composition assessment for the definition of cardiometabolic risk. *Journal of Endocrinological Investigation*, 36(7):537–543, 2013. ISSN 1720-8386. doi: 10.3275/8943.

Amerigroup. Icd-10-cm to cms-hcc crosswalk - amerigroup providers. https://provider.amerigroup.com/dam/publicdocuments/ALL_CARE_CMSHCCRAModel_mrdcoding_tips.pdf, 2019. [Accessed September 23, 2025].

Dawid Boczar, Magdalena Kitala, Klaudia Nowak, Bartlomiej Nowak, and Rafal Slojewski. 3d breast scanning in plastic surgery utilizing free iphone lidar applications and standard consumer devices: A comparative analysis. *Aesthetic Surgery Journal*, 2024. doi: 10.1093/asj/sjae251. URL https://academic.oup.com/asj/advance-article-abstract/doi/10.1093/asj/sjae251/7941969.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE TPAMI*, 2019.

Chia-Yu et al. Chang. Overcoming data scarcity in biomedical imaging with a foundational multi-task model. *Nature Computational Science*, 4:582–594, 2024. doi: 10.1038/s43588-024-00662-z.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. In *MICCAI*, pp. 485–495. Springer, 2021a.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiaoyang Wang, Jin Ye, Yuxin Zhou, Shuyang Li, Chengrui Wang, Tianyu Zha, Yuyin Zhou, Jing Zhang, Lequan Yu, and Ling Shao. Transunet: Transformers make strong encoders for medical image segmentation. *Medical Image Analysis*, 2021b. doi: 10.1016/j.media.2024.102965. URL https://www.sciencedirect.com/science/article/pii/S1361841524002056.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 794–803. PMLR, 2018. URL https://proceedings.mlr.press/v80/chen18a.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021a.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021b. URL http://arxiv.org/abs/2010.11929. arXiv:2010.11929 [cs].

Salma Galaaoui, Eduardo Valle, David Picard, and Nermin Samet. 3d human pose and shape estimation from lidar point clouds: A review. *arXiv preprint*, arXiv:2509.12197v1, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, June 2016. IEEE. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.90. URL http://ieeexplore.ieee.org/document/7780459/.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, Honolulu, HI, July 2017. IEEE. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.243. URL `https://ieeexplore.ieee.org/document/8099726/`.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jacob Seekins, Daniel Mong, Safwan Halabi, John Sandberg, Ross Jones, David Larson, Curtis Langlotz, Bhavik Patel, Matthew Lungren, and Andrew Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *AAAI*, 2019.

Sumbal A. Janjua, Joseph M. Massaro, Michael L. Chuang, Agostino Ralph B. D, Udo Hoffmann, and Donnell Christopher J. O. Thresholds for Abdominal Aortic Calcium That Predict Cardiovascular Disease Events in the Framingham Heart Study. *JACC: Cardiovascular Imaging*, 14(3): 695–697, March 2021. doi: 10.1016/j.jcmg.2020.09.019. URL `https://www.jacc.org/doi/10.1016/j.jcmg.2020.09.019`. Publisher: American College of Cardiology Foundation.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathan R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Yifan Peng, Zhenglong Lu, Roger G. Mark, and Steven Horng. Mimic-cxr: A large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schüttl, Sven Dähne, Dumitru Erhan, and Been Kim. The (un)reliability of saliency methods. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller (eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*, pp. 267–280. Springer, Cham, 2019. doi: 10.1007/978-3-030-28954-6_14.

Jane J. Lee, Alison Pedley, Udo Hoffmann, Joseph M. Massaro, Daniel Levy, and Michelle T. Long. Visceral and intrahepatic fat are associated with cardiometabolic risk factors above other ectopic fat depots: The Framingham Heart Study. *The American journal of medicine*, 131 (6):684–692.e12, June 2018. ISSN 0002-9343. doi: 10.1016/j.amjmed.2018.02.002. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5964004/`.

Mi-Jeong Lee, Yuanyuan Wu, and Susan K. Fried. Adipose tissue heterogeneity: Implication of depot differences in adipose tissue for obesity complications. *Molecular Aspects of Medicine*, 34 (1):1–11, 2013. doi: 10.1016/j.mam.2012.10.001.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL `https://arxiv.org/abs/1711.05101`.

J. C. Lovejoy, C. M. Champagne, L. de Jonge, H. Xie, and S. R. Smith. Increased visceral fat and decreased energy expenditure during the menopausal transition. *International Journal of Obesity*, 32(6):949–958, June 2008. doi: 10.1038/ijo.2008.25.

Haoran Lu, Yifei She, Jun Tie, and Shengzhou Xu. Half-unet: A simplified u-net architecture for medical image segmentation. *Frontiers in Neuroinformatics*, 16:911679, 2022. doi: 10.3389/fninf.2022.911679. URL `https://www.frontiersin.org/articles/10.3389/fninf.2022.911679/full`.

Tomoko Maruyama, Norio Hayashi, Yusuke Sato, Toshihiro Ogura, Masumi Uehara, Haruyuki Watanabe, and Yoshihiro Kitoh. Artifact estimation network for mr images: effectiveness of batch normalization and dropout layers. *BMC Medical Imaging*, 25:144, 2025. doi: 10.1186/s12880-025-01663-8.

Xuefeng Mei, Zhongyi Liu, Paul M. Robson, Benjamin Marinelli, Michael Huang, Ankit Doshi, Adam Jacobi, Chang Cao, Daniel Link, Tao Yang, Zahi A. Fayad, and Yiyu Yang. Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 4(5):e210315, 2022a.

Xueyan Mei, Zelong Liu, Philip M. Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E. Link, Thomas Yang, Ying Wang, Hayit Greenspan, Timothy Deyer, Zahi A. Fayad, and Yang Yang. Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 0(ja):e210315, 2022b. doi: 10.1148/ryai.210315. URL https://doi.org/10.1148/ryai.210315.

Aliasghar Mortazi, Vedat Cicek, Elif Keles, and Ulas Bagci. Selecting the best optimizers for deep learningbased medical image segmentation. *Frontiers in Radiology*, 3:1175473, 2023. doi: 10.3389/fradi.2023.1175473.

Katja Oberhofer, Cline Knopfli, Basil Achermann, and Silvio R Lorenzetti. Feasibility of using laser imaging detection and ranging technology for contactless 3d body scanning and anthropometric assessment of athletes. *Sports*, 12(4):92, 2024. doi: 10.3390/sports12040092. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11054930/.

Maxime Oquab, Timothe Darcet, Tho Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herv Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, February 2024. URL http://arxiv.org/abs/2304.07193. arXiv:2304.07193 [cs].

Perry J. Pickhardt, Ronald M. Summers, John W. Garrett, and et al. Automated ct-based body composition analysis: Opportunistic prediction of future major osteoporotic fractures in asymptomatic adults. *Radiology*, 297(1):64–74, 2020. doi: 10.1148/radiol.2020191465.

Ayis Pyrros, Stephen M. Borstelmann, Ramana Mantravadi, Zachary Zaiman, Kaesha Thomas, Brandon Price, Eugene Greenstein, Nasir Siddiqui, Melinda Willis, Ihar Shulhan, John Hines-Shah, Jeanne M. Horowitz, Paul Nikolaidis, Matthew P. Lungren, Jorge Mario Rodrguez-Fernndez, Judy Wawira Gichoya, Sanmi Koyejo, Adam E. Flanders, Nishith Khandwala, Amit Gupta, John W. Garrett, Joseph Paul Cohen, Brian T. Layden, Perry J. Pickhardt, and William Galanter. Opportunistic detection of type 2 diabetes using deep learning from frontal chest radiographs. *Nature Communications*, 14(1):4039, July 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-39631-x. URL https://www.nature.com/articles/s41467-023-39631-x.

Vera Regitz-Zagrosek. *Sex and Gender Differences in Pharmacology*, volume 214. Springer, Berlin, Heidelberg, 2012. ISBN 978-3-642-30726-3. doi: 10.1007/978-3-642-30726-3.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pp. 234–241. Springer, 2015.

Klara J. Rosenquist, Alison Pedley, Joseph M. Massaro, Kate E. Therkelsen, Joanne M. Murabito, Udo Hoffmann, and Caroline S. Fox. Visceral and Subcutaneous Fat Quality and Cardiometabolic Risk. *JACC: Cardiovascular Imaging*, 6(7):762–771, July 2013. doi: 10.1016/j.jcmg.2012.11.021. URL https://www.jacc.org/doi/10.1016/j.jcmg.2012.11.021. Publisher: American College of Cardiology Foundation.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Bane Sullivan and Alexander Kaszynski. PyVista: 3D plotting and mesh analysis through a streamlined interface for the Visualization Toolkit (VTK). *Journal of Open Source Software*, 4(37):1450, May 2019. doi: 10.21105/joss.01450. URL https://doi.org/10.21105/joss.01450.

Katherine Sweatt, W. Timothy Garvey, and Catia Martins. Strengths and Limitations of BMI in the Diagnosis of Obesity: What is the Path Forward? *Current Obesity Reports*, 13(3):584–595, 2024. ISSN 2162-4968. doi: 10.1007/s13679-024-00580-1. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11306271/.

Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/tan19a.html.

Kate E. Therkelsen, Alison Pedley, Elizabeth K. Speliotes, Joseph M. Massaro, Joanne Murabito, Udo Hoffmann, and Caroline S. Fox. Intramuscular Fat and Associations With Metabolic Risk Factors in the Framingham Heart Study. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 33(4):863–870, April 2013. doi: 10.1161/ATVBAHA.112.301009. URL https://www.ahajournals.org/doi/10.1161/atvbaha.112.301009. Publisher: American Heart Association.

Diana M. Thomas, Ira Crofford, John Scudder, Brittany Oletti, Ashok Deb, and Steven B. Heymsfield. Updates on methods for body composition analysis: Implications for clinical practice. *Current Obesity Reports*, 14(1):8, 2025. doi: 10.1007/s13679-024-00593-w.

U.S. Department of Health and Human Services, Office for Civil Rights. Methods for de-identification of protected health information under the hipaa privacy rule. https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html, 2012. [Accessed September 23, 2025].

Vamsi Krishna Vasa, Wenhui Zhu, Xiwen Chen, Peijie Qiu, Xuanzhao Dong, and Yalin Wang. Sta-unet: Rethink the semantic redundant for medical imaging segmentation. *arXiv preprint arXiv:2410.11578*, 2024. URL https://arxiv.org/abs/2410.11578.

Zoran Vasic et al. Lidar-based scaling of opensim musculoskeletal human models is accurate and efficienta validation study using 3d body scanning. *Journal of Biomechanics*, 2024. doi: 10.1016/j.jbiomech.2024.111891. URL https://www.sciencedirect.com/science/article/pii/S0021929024005189.

Fan Wang and Jian Huang. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2402.02491*, 2024.

Jonathan C. K. Wells. Sexual dimorphism of body composition. *Best Practice & Research Clinical Endocrinology & Metabolism*, 21(3):415–430, 2007. doi: 10.1016/j.beem.2007.04.007.

Mikalai Zamotsin, Andrey Dyagilev, Hawas S. Hawaz, Ttiana olovenko, and Oleksandr Shovkomud. Human body measurement with the iphone 12 pro lidar scanner. In *American Institute of Physics Conference Series*, volume 2430, pp. 090009, 2022. doi: 10.1063/5.0078310. URL http://ui.adsabs.harvard.edu/abs/2022AIPC.2430i0009M/abstract.

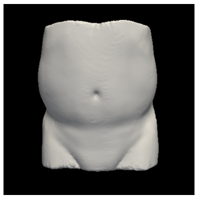
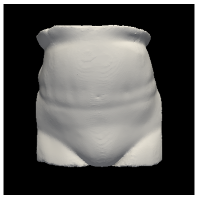
Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (MICCAI)*, pp. 3–11, 2018. doi: 10.1007/978-3-030-00889-5_1. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC7329239/.

A APPENDIX

A.1 BIOMARKER DETAILS FOR FIGURE 1

This section provides complete biomarker details for the sample images shown in Figure 1.

Table 4: Complete biomarker details for the sample images shown in Figure 1.

Image				
Gender	Male	Male	Female	Female
Age (years)	66	75	72	73
Calcium Score	Present	Absent	Present	Absent
MI	Absent	Absent	Present	Absent
Mortality	Absent	Absent	Present	Absent
T2D	Present	Absent	Present	Absent
HCC-108	Absent	Absent	Present	Present
HCC-111	Absent	Absent	Absent	Absent
HCC-12	Absent	Absent	Absent	Absent
HCC-18	Absent	Absent	Absent	Absent
HCC-96	Absent	Absent	Absent	Absent

A.2 Dataset Release Statement

The AbdCTBench dataset will be released at `https://abdctbenchrepo.github.io/AbdCTBench/` (an anonymized url compliant with the double-blind submission policy) under Creative Commons BY 4.0 license immediately upon paper acceptance. The release will include:

- 23,506 2D depth map projections (PNG format, ≈50KB each)
- 23,506 3D STL surface meshes (≈1-2MB each)
- Complete DICOM-to-STL-to-PNG processing pipeline (Python code)
- OSCAR biomarker extraction pipeline
- Pre-trained model checkpoints for all 8 architectures (including Swin and multi-task models)
- Train/val/test splits and evaluation protocols
- HIPAA-compliant de-identified labels: 87 comorbidities, 31 diagnoses, 16 biomarkers

While multi-site CT datasets exist (e.g., Stanford Merlin), they lack the HCC and ICD-10 diagnosis codes required for our benchmark tasks. Integrating imaging with structured clinical outcomes at scale remains challenging. With the release of AbdCTBench and all associated elements, any institution with CT DICOM series and corresponding HCC/ICD-10 codes can reproduce the benchmark and validate results locally. Given current data availability, this is the most practical and scalable path forward for multi-site validation.

Further, no publicly available datasets currently contain paired consumer depth sensor captures and CT scans of abdominal regions, in order to quantify the fidelity gap. With the release of AbdCTBench, researchers with consumer depth devices can apply our pipeline and quantify geometric fidelity gaps. Modern smartphone LiDAR (≈5mm depth resolution) should capture coarse body contour features that our models learn. We view AbdCTBench as establishing proof-of-concept that surface geometry contains predictive biomarker information, with the open-source pipeline enabling community-driven real-world validation.

A.3 Technical Details of DICOM to STL to PNG conversion

Once the DICOM image series are provided as input and loaded, the volume processing step applies optional shrinking to reduce volumes to a maximum of 256^3 voxels using shrink factors calculated per dimension (optionally applied when the shrink factor exceeds 3). Anisotropic smoothing uses sitk.CurvatureAnisotropicDiffusion with a time step of 0.03 to convert the data to sitkFloat32 format for further processing. Then, the surface extraction step generates 3D triangular meshes using VTK's contour filter, and the subsequent mesh processing applies refinement such as mesh cleaning, small object removal, and mesh smoothing. The processed mesh is then exported as a binary STL file. The STL to 2D image conversion step uses a square 384 x 384 pixel window, automatically centers the mesh at the origin, positions the camera using a distance factor of 3.0, applies an 80% zoom factor, and includes a fixed 10° rotation around the Z-axis. Following the application of 5000 mesh smoothing iterations, the output is a PNG image of size 384 x 384 pixels.

A.4 Complete Dataset Statistics

Below we present a variety of dataset statistics, feature distributions, variable correlations, and patient clustering visualizations.

Table 5: Complete CT-derived biomarker statistics for AbdCTBench dataset.

Biomarker	Count	Mean	Std Dev	Min	Max
BMD L1 High Sensitivity (HU)	22,083	148.4	44.3	-46.2	603.2
BMD L1 Standard (HU)	22,119	179.1	54.4	-41.1	1,180.1
Calcium Scoring Abdominal Agatston	23,506	1,200.9	3,126.5	0.0	37,152.0
Kidney Median HU	23,279	90.0	58.8	5.0	298.9
Kidney Volume (cm^3)	23,401	349.0	94.8	50.0	750.0
L3 SAT Area (cm^2)	23,467	218.4	111.2	5.9	838.2
L3 TAT Area (cm^2)	23,467	365.2	177.0	8.5	1,136.5
L3 VAT Area (cm^2)	23,467	146.9	104.6	0.0	647.7
L3 VAT Median (HU)	23,418	-89.0	10.5	-118.5	-30.0
L3 VAT/SAT Ratio	23,466	0.73	0.56	0.0	4.46
Liver Median HU	23,501	82.3	30.3	-24.7	233.2
Liver Volume (cm^3)	23,498	1,578.4	432.8	269.0	4,973.0
L3 Muscle Area (cm^2)	23,469	147.5	37.4	29.2	315.3
L3 Muscle Mean HU	23,468	36.2	16.4	-47.5	89.8
Spleen Median HU	23,473	82.1	37.2	10.3	488.1
Spleen Volume (cm^3)	23,242	223.9	127.1	50.0	4,323.0

Diagnosis	Prevalence (%)
Essential Hypertension	53.7
Type 2 Diabetes	44.6
Impaired Glucose Tolerance	38.0
Tobacco Use	26.8
Myocardial Infarction	23.1
Osteoporosis	14.7
Heart Failure	11.4
CVD	15.5
Hypertensive CKD	10.0
Chronic Liver Disease	7.8

HCC Code	Condition	Prevalence (%)
HCC 108	Vascular Disease	22.6
HCC 19	Diabetes without Complications	13.0
HCC 12	Breast, Prostate, Other Cancers	10.9
HCC 85	Congestive Heart Failure	9.9
HCC 48	Coagulation defects	9.8
HCC 18	Diabetes with Chronic Complications	8.3
HCC 11	COPD	5.6
HCC 40	Rheumatologic arthritis	6.1
HCC 23	Other Significant Endocrine Disorders	5.0
HCC 22	Morbid obesity	4.7

Table 6: Clinical diagnosis prevalence (left) and HCC comorbidity codes (right) in AbdCTBench.

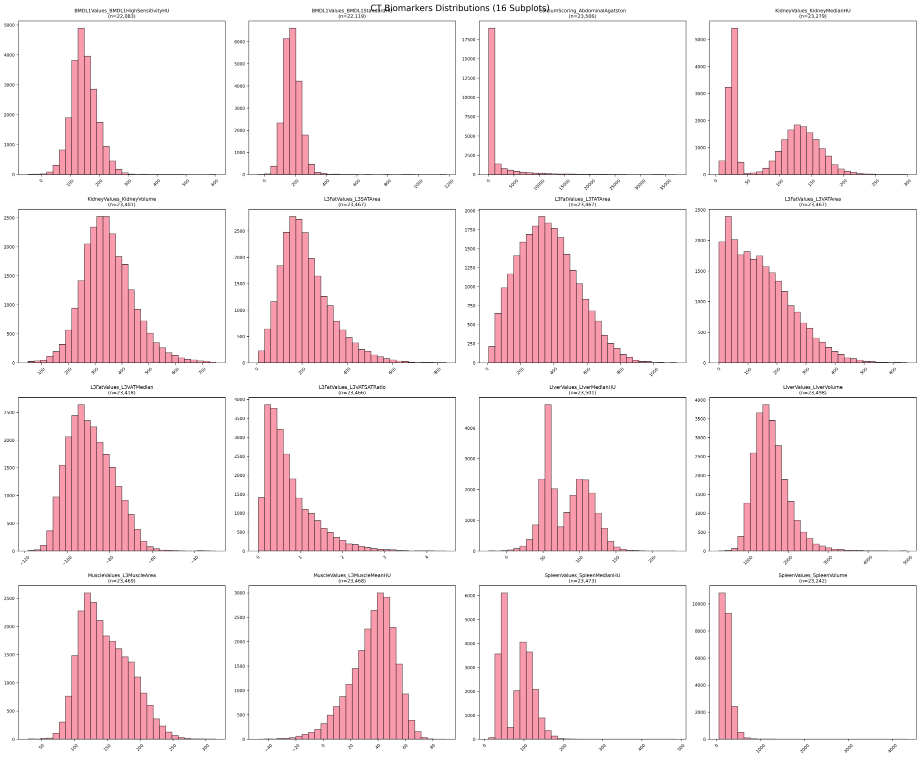


Figure 3: CT-derivedbiomarker distributions across the AbdCTBench dataset.

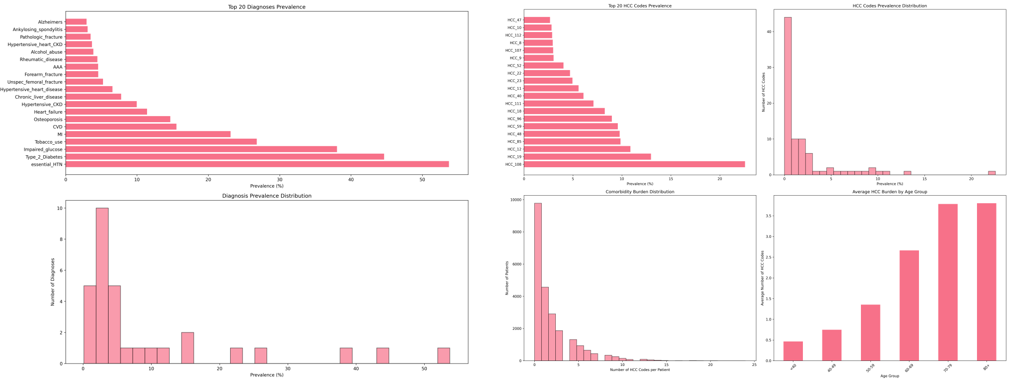


Figure 4: Prevalence of (a) diagnosis (ICD-10) codes, and (b) Hierarchical Condition Categories (HCC) codes.

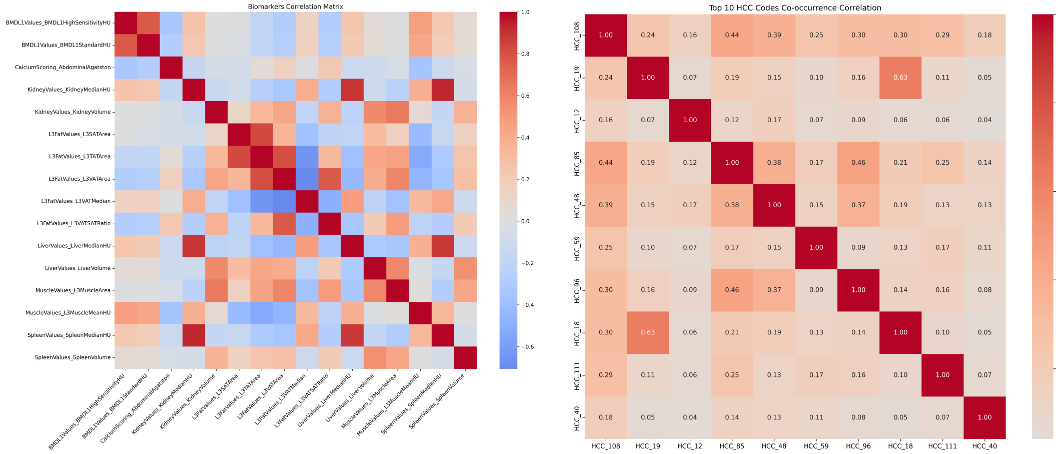


Figure 5: Correlations between (a) CT-derived biomarkers, and (b) HCC codes.

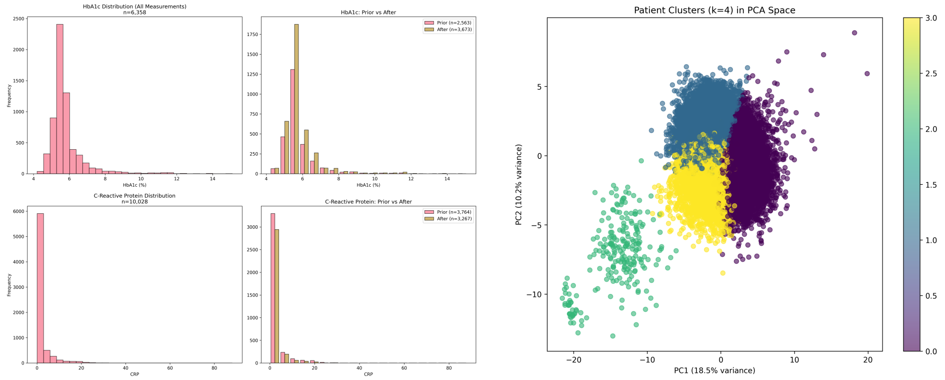


Figure 6: (a) Longitudinal laboratory value distributions, and (b) patient clustering patterns. The clusters were obtained via k-means clustering (with k = 4) by using the first 22 Principal Components (which explained 80% of the variance in the data from all the numeric features).

A.5 Detailed Results by Biomarker

This section provides comprehensive results for each biomarker across all architectures, including all metrics with 95% confidence intervals computed via patient-level bootstrapping.

Table 7: Complete results for Age prediction across all architectures.

Architecture	MAE	MSE	R^2
Naive Baseline	13.16 (12.79, 13.57)	265.11 (252.99, 278.70)	-0.0001 (-0.0030, 0.0000)
DenseNet-121	6.769 (6.551, 6.994)	74.391 (69.636, 79.591)	0.719 (0.696, 0.741)
EfficientNet-B0	6.223 (6.016, 6.422)	64.447 (60.141, 69.166)	0.757 (0.735, 0.776)
ResNet-18	6.472 (6.264, 6.678)	67.985 (63.650, 73.052)	0.744 (0.722, 0.763)
ResNet-34	6.486 (6.284, 6.692)	68.853 (64.397, 73.698)	0.740 (0.718, 0.761)
ResNet-50	6.341 (6.154, 6.532)	65.604 (61.517, 70.203)	0.753 (0.733, 0.770)
ViT-Small (DINOv2)	6.465 (6.260, 6.684)	70.542 (65.913, 75.829)	0.734 (0.710, 0.755)
Swin Transformer-Base	6.540 (6.338, 6.758)	70.80 (66.43, 75.85)	0.733 (0.710, 0.754)

Table 8: Complete results for Calcium Score prediction across all architectures.

Architecture	AUROC	Precision	Recall	Specificity	F1	Accuracy	Threshold
Naive Baseline	0.5 —	0.0 —	0.0 —	1.0 —	0.0 —	0.786 (0.770, 0.803)	—
DenseNet-121	0.847 (0.829, 0.863)	0.497 (0.460, 0.531)	0.746 (0.707, 0.782)	0.795 (0.776, 0.813)	0.597 (0.563, 0.626)	0.784 (0.767, 0.800)	0.8
EfficientNet-B0	0.847 (0.829, 0.864)	0.481 (0.446, 0.515)	0.765 (0.728, 0.803)	0.775 (0.757, 0.794)	0.591 (0.557, 0.621)	0.773 (0.757, 0.790)	0.8
ResNet-18	0.843 (0.825, 0.859)	0.442 (0.407, 0.472)	0.817 (0.782, 0.849)	0.719 (0.698, 0.738)	0.573 (0.540, 0.601)	0.740 (0.721, 0.756)	0.7
ResNet-34	0.848 (0.831, 0.864)	0.465 (0.433, 0.498)	0.773 (0.737, 0.807)	0.758 (0.738, 0.778)	0.581 (0.549, 0.609)	0.761 (0.744, 0.779)	0.8
ResNet-50	0.833 (0.815, 0.849)	0.411 (0.379, 0.439)	0.837 (0.801, 0.868)	0.673 (0.651, 0.695)	0.551 (0.518, 0.579)	0.708 (0.688, 0.727)	0.5
ViT-Small (DINOv2)	0.829 (0.809, 0.846)	0.426 (0.394, 0.455)	0.819 (0.784, 0.852)	0.700 (0.678, 0.719)	0.561 (0.529, 0.589)	0.725 (0.706, 0.742)	0.8
Swin Transformer-Base	0.845 (0.828, 0.862)	0.474 (0.438, 0.505)	0.769 (0.729, 0.805)	0.767 (0.748, 0.786)	0.586 (0.554, 0.615)	0.768 (0.750, 0.784)	0.9

Table 9: Complete results for HCC-108 prediction across all architectures.

Architecture	AUROC	Precision	Recall	Specificity	F1	Accuracy	Threshold
Naive Baseline	0.5 —	0.0 —	0.0 —	1.0 —	0.0 —	0.763 (0.745, 0.780)	—
DenseNet-121	0.740 (0.720, 0.760)	0.384 (0.354, 0.414)	0.709 (0.673, 0.747)	0.647 (0.626, 0.669)	0.498 (0.468, 0.526)	0.662 (0.643, 0.681)	0.8
EfficientNet-B0	0.753 (0.732, 0.775)	0.376 (0.349, 0.405)	0.775 (0.741, 0.807)	0.601 (0.579, 0.624)	0.507 (0.477, 0.536)	0.643 (0.623, 0.663)	0.8
ResNet-18	0.763 (0.742, 0.784)	0.375 (0.347, 0.404)	0.791 (0.759, 0.822)	0.590 (0.567, 0.612)	0.508 (0.480, 0.539)	0.638 (0.620, 0.657)	0.7
ResNet-34	0.749 (0.729, 0.769)	0.362 (0.336, 0.390)	0.804 (0.770, 0.834)	0.561 (0.539, 0.584)	0.499 (0.470, 0.528)	0.619 (0.599, 0.638)	0.8
ResNet-50	0.718 (0.698, 0.739)	0.368 (0.340, 0.396)	0.737 (0.700, 0.772)	0.607 (0.584, 0.629)	0.491 (0.462, 0.519)	0.638 (0.620, 0.655)	0.2
ViT-Small (DINOv2)	0.743 (0.723, 0.764)	0.375 (0.348, 0.403)	0.730 (0.696, 0.766)	0.623 (0.603, 0.645)	0.496 (0.467, 0.524)	0.649 (0.631, 0.667)	0.8
Swin Transformer-Base	0.768 (0.749, 0.788)	0.378 (0.352, 0.407)	0.800 (0.766, 0.830)	0.592 (0.569, 0.614)	0.514 (0.486, 0.543)	0.641 (0.623, 0.661)	0.8

Table 10: Complete results for HCC-111 prediction across all architectures.

Architecture	AUROC	Precision	Recall	Specificity	F1	Accuracy	Threshold
Naive Baseline	0.5 —	0.0 —	0.0 —	1.0 —	0.0 —	0.929 (0.919, 0.939)	—
DenseNet-121	0.716 (0.681, 0.749)	0.122 (0.103, 0.142)	0.790 (0.731, 0.844)	0.570 (0.550, 0.591)	0.212 (0.181, 0.242)	0.585 (0.567, 0.605)	0.9
EfficientNet-B0	0.747 (0.708, 0.782)	0.131 (0.111, 0.152)	0.717 (0.643, 0.784)	0.638 (0.617, 0.658)	0.221 (0.190, 0.254)	0.643 (0.621, 0.662)	0.9
ResNet-18	0.769 (0.734, 0.802)	0.121 (0.103, 0.140)	0.843 (0.788, 0.895)	0.533 (0.513, 0.554)	0.211 (0.182, 0.240)	0.555 (0.535, 0.575)	0.9
ResNet-34	0.766 (0.730, 0.799)	0.114 (0.097, 0.133)	0.886 (0.829, 0.933)	0.479 (0.459, 0.500)	0.203 (0.175, 0.231)	0.508 (0.488, 0.528)	0.9
ResNet-50	0.739 (0.704, 0.774)	0.154 (0.123, 0.187)	0.416 (0.343, 0.497)	0.826 (0.811, 0.841)	0.224 (0.182, 0.269)	0.797 (0.780, 0.813)	0.2
ViT-Small (DINOv2)	0.760 (0.727, 0.793)	0.117 (0.100, 0.134)	0.898 (0.853, 0.939)	0.483 (0.463, 0.504)	0.206 (0.179, 0.233)	0.512 (0.492, 0.532)	0.9
Swin Transformer-Base	0.765 (0.732, 0.796)	0.150 (0.128, 0.175)	0.753 (0.687, 0.817)	0.676 (0.656, 0.696)	0.251 (0.217, 0.284)	0.682 (0.663, 0.700)	0.9

Table 11: Complete results for HCC-12 prediction across all architectures.

Architecture	AUROC	Precision	Recall	Specificity	F1	Accuracy	Threshold
Naive Baseline	0.5 —	0.0 —	0.0 —	1.0 —	0.0 —	0.884 (0.871, 0.897)	—
DenseNet-121	0.587 (0.551, 0.622)	0.136 (0.117, 0.154)	0.585 (0.526, 0.642)	0.513 (0.493, 0.534)	0.220 (0.192, 0.247)	0.521 (0.502, 0.540)	0.7
EfficientNet-B0	0.586 (0.550, 0.623)	0.162 (0.136, 0.189)	0.456 (0.394, 0.512)	0.691 (0.670, 0.711)	0.239 (0.205, 0.273)	0.663 (0.643, 0.683)	0.3
ResNet-18	0.580 (0.544, 0.614)	0.134 (0.118, 0.153)	0.713 (0.659, 0.766)	0.397 (0.377, 0.418)	0.226 (0.201, 0.253)	0.434 (0.415, 0.455)	0.9
ResNet-34	0.591 (0.557, 0.624)	0.148 (0.129, 0.170)	0.621 (0.566, 0.674)	0.533 (0.511, 0.554)	0.240 (0.211, 0.269)	0.543 (0.524, 0.563)	0.9
ResNet-50	0.571 (0.536, 0.607)	0.150 (0.128, 0.176)	0.434 (0.380, 0.495)	0.678 (0.658, 0.697)	0.223 (0.192, 0.257)	0.650 (0.630, 0.669)	0.9
ViT-Small (DINOv2)	0.576 (0.542, 0.610)	0.140 (0.119, 0.163)	0.449 (0.390, 0.508)	0.640 (0.620, 0.661)	0.214 (0.184, 0.245)	0.618 (0.599, 0.636)	0.9
Swin Transformer-Base	0.580 (0.545, 0.616)	0.133 (0.116, 0.150)	0.728 (0.674, 0.777)	0.378 (0.358, 0.399)	0.225 (0.200, 0.250)	0.418 (0.398, 0.440)	0.8

Table 12: Complete results for HCC-18 prediction across all architectures.

Architecture	AUROC	Precision	Recall	Specificity	F1	Accuracy	Threshold
Naive Baseline	0.5 —	0.0 —	0.0 —	1.0 —	0.0 —	0.908 (0.896, 0.920)	—
DenseNet-121	0.766 (0.735, 0.797)	0.236 (0.197, 0.281)	0.454 (0.388, 0.528)	0.852 (0.836, 0.866)	0.311 (0.266, 0.362)	0.815 (0.800, 0.831)	0.8
EfficientNet-B0	0.789 (0.760, 0.816)	0.203 (0.177, 0.232)	0.764 (0.706, 0.819)	0.697 (0.677, 0.716)	0.321 (0.285, 0.359)	0.703 (0.685, 0.721)	0.9
ResNet-18	0.799 (0.770, 0.828)	0.145 (0.127, 0.164)	0.926 (0.886, 0.960)	0.446 (0.426, 0.467)	0.250 (0.222, 0.279)	0.490 (0.472, 0.510)	0.9
ResNet-34	0.775 (0.746, 0.805)	0.167 (0.146, 0.190)	0.857 (0.808, 0.904)	0.569 (0.547, 0.589)	0.280 (0.249, 0.311)	0.595 (0.575, 0.613)	0.9
ResNet-50	0.782 (0.754, 0.812)	0.218 (0.188, 0.250)	0.648 (0.584, 0.715)	0.765 (0.748, 0.782)	0.326 (0.286, 0.369)	0.754 (0.737, 0.771)	0.9
ViT-Small (DINOv2)	0.785 (0.755, 0.817)	0.181 (0.157, 0.208)	0.792 (0.738, 0.849)	0.637 (0.617, 0.658)	0.295 (0.261, 0.331)	0.652 (0.633, 0.671)	0.9
Swin Transformer-Base	0.801 (0.776, 0.828)	0.184 (0.161, 0.210)	0.829 (0.779, 0.878)	0.627 (0.608, 0.646)	0.301 (0.268, 0.335)	0.646 (0.627, 0.663)	0.9

Table 13: Complete results for HCC-96 prediction across all architectures.

Architecture	AUROC	Precision	Recall	Specificity	F1	Accuracy	Threshold
Naive Baseline	0.5 —	0.0 —	0.0 —	1.0 —	0.0 —	0.914 (0.901, 0.925)	—
DenseNet-121	0.757 (0.723, 0.787)	0.156 (0.134, 0.178)	0.818 (0.763, 0.866)	0.581 (0.560, 0.602)	0.262 (0.229, 0.294)	0.601 (0.582, 0.621)	0.9
EfficientNet-B0	0.763 (0.732, 0.790)	0.152 (0.131, 0.174)	0.818 (0.764, 0.866)	0.569 (0.548, 0.590)	0.257 (0.226, 0.288)	0.591 (0.571, 0.611)	0.9
ResNet-18	0.760 (0.727, 0.791)	0.162 (0.138, 0.186)	0.773 (0.715, 0.829)	0.621 (0.600, 0.641)	0.268 (0.233, 0.301)	0.634 (0.615, 0.653)	0.9
ResNet-34	0.728 (0.694, 0.761)	0.177 (0.144, 0.211)	0.453 (0.386, 0.519)	0.801 (0.785, 0.817)	0.255 (0.212, 0.297)	0.771 (0.755, 0.787)	0.8
ResNet-50	0.738 (0.705, 0.769)	0.198 (0.165, 0.232)	0.522 (0.448, 0.589)	0.800 (0.784, 0.817)	0.287 (0.244, 0.329)	0.776 (0.760, 0.792)	0.8
ViT-Small (DINOv2)	0.757 (0.721, 0.788)	0.162 (0.139, 0.186)	0.773 (0.711, 0.828)	0.621 (0.601, 0.643)	0.268 (0.234, 0.302)	0.635 (0.615, 0.654)	0.9
Swin Transformer-Base	0.770 (0.739, 0.798)	0.178 (0.152, 0.206)	0.699 (0.632, 0.761)	0.695 (0.676, 0.714)	0.284 (0.246, 0.321)	0.695 (0.677, 0.713)	0.9

Table 14: Complete results for Myocardial Infarction (MI) prediction across all architectures.

Architecture	AUROC	Precision	Recall	Specificity	F1	Accuracy	Threshold
Naive Baseline	0.5 —	0.0 —	0.0 —	1.0 —	0.0 —	0.765 (0.748, 0.783)	—
DenseNet-121	0.730 (0.703, 0.752)	0.377 (0.346, 0.408)	0.674 (0.633, 0.713)	0.657 (0.636, 0.680)	0.483 (0.451, 0.514)	0.661 (0.642, 0.681)	0.8
EfficientNet-B0	0.732 (0.708, 0.754)	0.399 (0.367, 0.432)	0.687 (0.647, 0.722)	0.683 (0.663, 0.704)	0.505 (0.472, 0.536)	0.684 (0.665, 0.702)	0.9
ResNet-18	0.729 (0.705, 0.752)	0.374 (0.344, 0.405)	0.669 (0.631, 0.707)	0.656 (0.634, 0.678)	0.480 (0.449, 0.510)	0.659 (0.640, 0.678)	0.8
ResNet-34	0.731 (0.706, 0.753)	0.375 (0.346, 0.405)	0.681 (0.640, 0.719)	0.651 (0.630, 0.673)	0.483 (0.452, 0.513)	0.658 (0.639, 0.678)	0.8
ResNet-50	0.716 (0.693, 0.739)	0.346 (0.316, 0.373)	0.717 (0.679, 0.755)	0.583 (0.561, 0.606)	0.467 (0.435, 0.496)	0.615 (0.595, 0.634)	0.8
ViT-Small (DINOv2)	0.732 (0.707, 0.754)	0.347 (0.320, 0.374)	0.823 (0.790, 0.852)	0.525 (0.501, 0.547)	0.488 (0.457, 0.515)	0.595 (0.574, 0.615)	0.7
Swin Transformer-Base	0.742 (0.718, 0.763)	0.368 (0.338, 0.394)	0.748 (0.710, 0.784)	0.606 (0.583, 0.628)	0.493 (0.461, 0.521)	0.639 (0.620, 0.658)	0.8

Table 15: Complete results for Mortality prediction across all architectures.

Architecture	AUROC	Precision	Recall	Specificity	F1	Accuracy	Threshold
Naive Baseline	0.5 —	0.0 —	0.0 —	1.0 —	0.0 —	0.886 (0.874, 0.900)	—
DenseNet-121	0.823 (0.800, 0.845)	0.281 (0.247, 0.311)	0.727 (0.675, 0.779)	0.761 (0.744, 0.779)	0.405 (0.364, 0.442)	0.757 (0.741, 0.774)	0.9
EfficientNet-B0	0.830 (0.805, 0.852)	0.318 (0.277, 0.355)	0.633 (0.574, 0.687)	0.826 (0.810, 0.841)	0.423 (0.377, 0.463)	0.804 (0.787, 0.819)	0.9
ResNet-18	0.839 (0.816, 0.861)	0.289 (0.256, 0.321)	0.749 (0.694, 0.800)	0.764 (0.748, 0.782)	0.418 (0.379, 0.455)	0.763 (0.746, 0.779)	0.9
ResNet-34	0.825 (0.799, 0.848)	0.359 (0.310, 0.405)	0.581 (0.516, 0.637)	0.867 (0.853, 0.881)	0.444 (0.391, 0.489)	0.835 (0.820, 0.849)	0.8
ResNet-50	0.810 (0.784, 0.834)	0.220 (0.194, 0.242)	0.850 (0.809, 0.891)	0.613 (0.593, 0.632)	0.349 (0.315, 0.378)	0.640 (0.620, 0.658)	0.9
ViT-Small (DINOv2)	0.811 (0.785, 0.836)	0.255 (0.223, 0.285)	0.745 (0.690, 0.796)	0.720 (0.701, 0.739)	0.379 (0.338, 0.417)	0.723 (0.705, 0.742)	0.9
Swin Transformer-Base	0.828 (0.803, 0.851)	0.280 (0.245, 0.312)	0.749 (0.696, 0.799)	0.753 (0.735, 0.772)	0.407 (0.367, 0.445)	0.752 (0.736, 0.770)	0.9

Table 16: Complete results for Type-2 Diabetes prediction across all architectures.

Architecture	AUROC	Precision	Recall	Specificity	F1	Accuracy	Threshold
Naive Baseline	0.5 —	0.0 —	0.0 —	1.0 —	0.0 —	0.551 (0.531, 0.571)	—
DenseNet-121	0.728 (0.709, 0.750)	0.533 (0.509, 0.555)	0.909 (0.892, 0.926)	0.351 (0.323, 0.376)	0.672 (0.650, 0.690)	0.601 (0.580, 0.621)	0.4
EfficientNet-B0	0.740 (0.720, 0.761)	0.532 (0.508, 0.553)	0.915 (0.897, 0.931)	0.344 (0.318, 0.368)	0.673 (0.652, 0.691)	0.600 (0.580, 0.619)	0.3
ResNet-18	0.735 (0.714, 0.756)	0.551 (0.526, 0.573)	0.888 (0.869, 0.906)	0.409 (0.381, 0.434)	0.680 (0.659, 0.699)	0.624 (0.603, 0.643)	0.4
ResNet-34	0.742 (0.722, 0.762)	0.538 (0.514, 0.560)	0.913 (0.894, 0.929)	0.361 (0.335, 0.387)	0.677 (0.656, 0.695)	0.609 (0.589, 0.628)	0.4
ResNet-50	0.733 (0.714, 0.753)	0.550 (0.526, 0.574)	0.881 (0.862, 0.899)	0.412 (0.386, 0.440)	0.677 (0.656, 0.696)	0.623 (0.603, 0.641)	0.4
ViT-Small (DINOv2)	0.735 (0.714, 0.755)	0.540 (0.517, 0.562)	0.895 (0.875, 0.913)	0.379 (0.354, 0.405)	0.674 (0.652, 0.692)	0.611 (0.591, 0.630)	0.4
Swin Transformer-Base	0.740 (0.720, 0.759)	0.533 (0.510, 0.556)	0.917 (0.900, 0.933)	0.347 (0.321, 0.373)	0.674 (0.654, 0.692)	0.603 (0.583, 0.622)	0.2

A.6 Multi-task Learning

In a follow-up experiment, we implemented a multi-task learning framework covering all 10 benchmarked biomarkers, by training multi-task ResNet-18, ResNet-34, and ResNet-50 (RadImageNet) models. Each of these had a shared backbone, followed by task-specific heads. Further, we kept the standardized single-target training protocol (learning rates, optimizers, augmentations, schedulers) wherever applicable, so the comparison isolates the multi-task objective rather than tuning differences. Each mini-batch optimized the sum of per-task losses, cross-entropy for the classification biomarkers and MSE for age regression. We incorporated GradNorm (Chen et al., 2018) to balance gradient magnitudes so that easier tasks cannot dominate optimization. Model selection used the median AUROC across the binary biomarkers on the validation split, ensuring that gains arise from broad improvements instead of a single outlier task.

This multi-task extension shows that AbdCTBench's standardized setup supports joint training. However, the results we obtained indicate that multi-task learning did not substantially improve, rather degraded performance, relative to single-task modeling:

Table 17: Multi-task learning results for non-HCC biomarkers by architecture on the test set. AUROC is reported for the binary classification tasks. MAE is reported for Age prediction (regression task). Bootstrapped 95% CIs are shown in parentheses.

Architecture	Age (MAE)	Calcium Score	MI	Mortality	T2D
ResNet-18	14.529	0.625	0.592	0.615	0.584
	(14.082–15.013)	(0.597–0.652)	(0.568–0.619)	(0.578–0.649)	(0.561–0.608)
ResNet-34	21.834	0.522	0.546	0.474	0.592
	(21.275–22.446)	(0.492–0.552)	(0.519–0.574)	(0.435–0.514)	(0.569–0.614)
ResNet-50	72.645	0.612	0.546	0.632	0.524
(RadImageNet)	(21.880–133.051)	(0.583–0.640)	(0.517–0.574)	(0.596–0.665)	(0.501–0.547)

Table 18: Multi-task learning results for HCC code biomarkers by architecture on the test set. All biomarkers report AUROC. Bootstrapped 95% CIs are shown in parentheses.

Architecture	HCC108	HCC111	HCC12	HCC18	HCC96
ResNet-18	0.603	0.508	0.527	0.590	0.611
	(0.577–0.631)	(0.464–0.551)	(0.494–0.564)	(0.551–0.629)	(0.567–0.653)
ResNet-34	0.537	0.507	0.508	0.655	0.530
	(0.510–0.566)	(0.463–0.550)	(0.472–0.544)	(0.614–0.692)	(0.486–0.576)
ResNet-50	0.552	0.628	0.492	0.570	0.574
(RadImageNet)	(0.525–0.577)	(0.583–0.673)	(0.456–0.532)	(0.527–0.614)	(0.528–0.614)
	(0.525–0.577)	(0.583–0.673)	(0.456–0.532)	(0.527–0.614)	(0.528–0.614)

The results above indicate that there may be negative transfer between the full set of 10 biomarkers. Further, the standardized training protocol for single-target modeling may not be amenable to the multi-task training problem, and will need to be investigated further. Multi-task learning with a smaller subset of biomarkers and hyper-parameters tuned specifically for that set of biomarkers may yield substantially better results, and AbdCTBench is training-ready for extensive multi-task learning modeling to be undertaken as future work.

A.7 Effective Learning of Representations from Abdominal Surface Geometry

From the performance metrics reported above, we do not see drastic differences between the architectures considered. This demonstrates the viability of effective representation learning from abdominal surface geometry for clinically relevant biomarker prediction. The effectiveness is observed across all architectures and biomarkers, except HCC-12 (Breast, Prostate, and other Cancers), whereby external surface geometry may not be predictive of the comorbidity from a clinical perspective either. To demonstrate this further, we apply Gradient-Weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) to the input images to visualize the representations.

In specific, we load ResNet-18 (the best-performing model on HCC-18 - Diabetes with Chronic Complications), and apply Grad-CAM on the last convolution layer to visualize the features learned from the surface geometry images. We collect a small random sample of size 100 from the test set, and select compelling examples to demonstrate the effectiveness of the learned representations as hypothesized. The heatmaps identify high attention regions from the surface geometry for the model to make predictions. The F1-optimal threshold of 0.9 was applied for the binary classification.

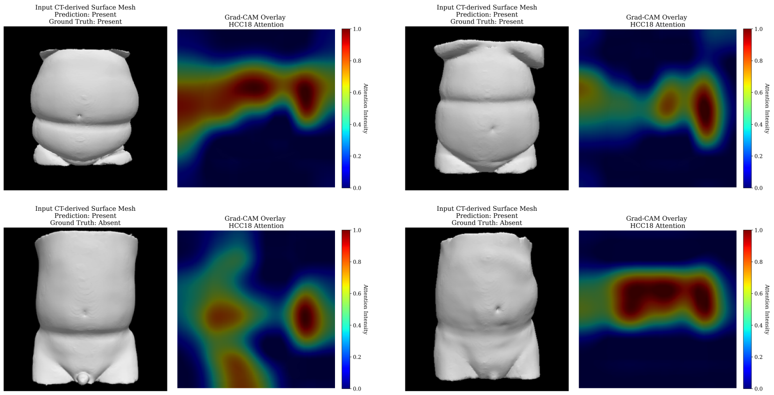


Figure 7: Grad-CAM visualizations showing learned representations from abdominal surface geometry. The heatmaps highlight regions of interest that the ResNet-18 model focuses on for HCC-18 (Diabetes with Chronic Complications) prediction.

While Grad-CAM is a popular interpretability method, it has been shown to be unreliable (Kindermans et al., 2019). We provide these visualizations as hypothesis-generating and these are not used to support any core claims. All of our main conclusions rely on quantitative performance metrics; the paper does not draw any causal or mechanistic inferences from Grad-CAM.

A.8 LLM Usage Declaration

We declare that we used LLMs for the following tasks:

- Literature review, finding relevant works, and understanding the state-of-the-art. All retrieved information was manually verified and validated.
- Improving the writing of the manuscript. All writing changes were manually verified and validated.
- Code generation and debugging. All code changes were manually reviewed, tested, and validated.