Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Transforming Pixels to Phenomics-based Representations for Generalizable Medical AI: Findings from the International AID-MRI Study

Fereshteh Hasanzadeh
University of Calgary
Rylan Marianchuk
University of Calgary
Cory Jensen
University of Calgary
Zahra Ghods
University of Calgary
Jose Aponte
University of Calgary
Talia Beckie
University of Calgary
Jacqueline Flewitt
University of Calgary
Sandra Rivest
University of Calgary
Kate Hanneman
University of Toronto
Ming-Yen Ng
The University of Hong Kong
Juan Gaztanaga
NYU Langone Hospital–Long Island
Edward Nabet
NYU Langone Hospital–Long Island
Ntobeko A. B. Ntusi
University of Cape Town (UCT) and Groote Schuur Hospital
D. Ian Paterson
University of Ottawa Heart Institute
Idan Roifman
University of Toronto
Mahesh Vidula

University of Pennsylvania

Colin B Josephson

University of Calgary

Robert Miller

University of Calgary

Dina Labib

University of Calgary

James A White

jawhit@ucalgary.ca

University of Calgary

Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11663703/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Transforming Pixels to Phenomics-based Representations for Generalizable Medical AI: Findings from the International AID-MRI Study

Fereshteh Hasanzadeh[1], Rylan Marianchuk[1], Cory Jensen[1], Zahra Ghods[1], Jose Aponte[1], Talia Beckie[1], Jacqueline Flewitt[1], Sandra Rivest[1], Kate Hanneman[2,3], Ming-Yen Ng[2,4], Juan Gaztanaga[2,5], Edward Nabet[2,5], Ntobeko A. B. Ntusi[2,6], D. Ian Paterson[2,7], Idan Roifman[2,8], Mahesh Vidula[2,9], Colin B Josephson[1], Robert Miller[1], Dina Labib[1,10], and James A White[1,2]

[1]Libin Cardiovascular Institute, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada.

[2] Investigators of the International AID-MRI Study (AI-based Diagnosis of cardiomyopathy using cardiac MRI)

[3] Department of Medical Imaging, Temerty Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada

[4] Department of Diagnostic Radiology, School of Clinical Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China

[5] Department of Cardiology, NYU Langone Hospital–Long Island, Mineola, New York, United States

[6] Department of Medicine, University of Cape Town (UCT) and Groote Schuur Hospital, Cape Town, South Africa; and South African Medical Research Council, Cape Town, South Africa.

[7] Division of Cardiology, University of Ottawa Heart Institute, Ottawa, Ontario, Canada

[8] Department of Medicine and Division of Cardiology, Sunnybrook Health Sciences Centre and University of Toronto, Toronto, Ontario, Canada

[9] Department of Cardiovascular Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States.

[10] Department of Cardiovascular Medicine, Cairo University School of Medicine, Cairo, Egypt.

Corresponding Author:

James White, MD, FRCP(C)

Director - Nelson PULSE Centre for Innovation

Director of Precision Medicine - Libin Cardiovascular Institute

Professor of Medicine

Departments of Cardiac Sciences and Diagnostic Imaging

Cumming School of Medicine

University of Calgary

Email: jawhit@ucalgary.ca

Website: www.nelsonpulsecentre.ca

ABSTRACT

Medical images contain hardware-, acquisition-, and population-specific features beyond those relevant to pathology. Artificial intelligence (AI) models can exploit these noncausal signals, leading to shortcut learning, susceptibility to domain shift, and reduced real-world generalizability. We evaluated whether four-dimensional (4D) phenomics representations of cardiac shape and deformation from cine cardiac magnetic resonance (CMR) mitigate these vulnerabilities using 1,091 participants of the International AID-MRI study (644 with hypertrophic cardiomyopathy [HCM] and 447 without cardiomyopathy). We trained HCM classifiers using both original CMR pixel data and 4D phenomics representations then systematically tested their ability to generalize across scanner vendor, scanner model, and the sex and race of participants. Pixel-based models achieved high in-distribution performance (AUC, 0.91–0.98), but performance declined in out-of-distribution tests. By comparison, 4D phenomics models maintained robust performance (AUC, 0.97–0.99) across all stress tests. These findings support phenomics-based representations as a robust foundation for generalizable AI in medical imaging.

INTRODUCTION

Artificial intelligence (AI) is catalyzing rapid transformations in the diagnosis and management of disease.[1,2] However, contemporary AI models derive predictions from observable patterns in signals embedded within training data rather than from fundamental understandings of how disease alters biological traits. This dependence establishes a core vulnerability toward model bias, whereby non-causal patterns unrelated to pathology but associated with disease presence may be embedded into model predictions.[3] When these biases are not anticipated and eliminated, model deployment in new clinical environments or underrepresented populations can erode diagnostic accuracy, compromise clinical decision-making, and diminish patient trust.[4,5]

These challenges are amplified in medical imaging as non-causal signals are intrinsically embedded within pixel data. Although imaging aims to characterize anatomy and pathology, image-based classification models can capture latent information related to image acquisition or patient characteristics.[6] For example, scanner hardware, reconstruction algorithms, and post-processing introduce systematic, non-biological variation,[7] while demographic attributes such as sex and race may also be implicitly encoded.[8] Together, these non-causal factors contribute to two major sources of model failure: shortcut learning, where models rely on spurious correlates rather than pathology,[9,10] and domain shift, where performance deteriorates when the source data or population characteristics differ from those observed during model training[11] (Figure 1). Commonly engaged mitigation strategies, such as data augmentation or class balancing, incompletely address these vulnerabilities[12], and motivate a need for fundamentally different approaches to how imaging data is represented in both model training and deployment.

The systematic transformation of medical imaging data into standardized, anatomy- and physiology-centered representations of disease, termed here *phenomics*, aims to maximize the expression of biologically meaningful variation while minimizing sensitivity to acquisition-specific and non-causal features. In the field of cardiovascular imaging, recent studies have demonstrated the capacity of cine cardiac magnetic resonance (CMR) data to be transformed into dynamic three-dimensional (i.e. 3D + *t*, 4D) digital representations of cardiac structure and function.[13–16] These 4D phenomics-based representations generate multi-dimensional quantitative descriptors of cardiovascular phenotypes, inclusive of ventricular shape, wall thickness, and myocardial strain, while excluding non-pathologic image pixel features inherently susceptible to shortcut learning and domain shift.

Leveraging a prospectively recruited multi-centre cohort of the International Artificial Intelligence-based Diagnosis of cardiomyopathy using cardiac MRI (AID-MRI) study, we systematically evaluated vulnerabilities to shortcut learning and domain shift in pixel-based disease classification and tested whether 4D phenomics can support identical disease classification tasks while reducing the susceptibility to non-causal signal learning. Models were systematically stress-tested across controlled variations in scanner vendor, site, sex and self-reported race to quantify susceptibility to shortcut learning and domain shift under each condition.

RESULTS

Disease classification models using both pixel-based and 4D phenomics-based data representations were successfully trained across a range of stress-testing scenarios to assess susceptibility to vendor-related shortcut learning and domain shifts associated with scanner vendor, scanner model, sex, and self-reported race. Both pixel- and 4D phenomics-based modeling

achieved comparable in-distribution (ID) performance for the exemplar classification task of discriminating hypertrophic cardiomyopathy (HCM) from those without cardiomyopathy. However, out-of-distribution (OOD) performance exposed vulnerabilities to shortcut learning and a loss of generalizability under domain shift conditions for pixel-based models, as summarized in Table 1. 4D phenomics-based models remained stable and demonstrated greater generalizability to previously unseen clinical environments.

Shortcut learning experiments showed that, under permissive training conditions, pixel-based models relied heavily on vendor-specific image features and exhibited marked declines in discrimination on out-of-distribution (OOD) data. The pixel-based model declined from an AUC of 0.91 (95% CI, 0.86–0.95) under ID conditions to 0.52 (95% CI, 0.37–0.67) during OOD testing (ΔAUC −0.39). As shown by the confusion matrices (Supplementary Figure S9), Healthy cases were consistently misclassified as HCM when exposed to the Siemens vendor signal, demonstrating that this vendor's identity had been adopted as a diagnostic shortcut. In contrast, the 4D phenomics-based model maintained identical diagnostic performance, with an AUC of 0.99 (95% CI, 0.97–1.00) in both ID and OOD testing (ΔAUC 0.00).

Domain shift experiments showed similar patterns in performance across different stress testing scenarios (Table 1). During sex-based domain shift, a pixel-based model trained on male subjects declined from an AUC of 0.97 (95% CI, 0.94–0.99) to 0.88 (95% CI, 0.81–0.95) when evaluated in female subjects, with women with HCM frequently misclassified as Healthy. In contrast, the 4D phenomics-based model maintained stable performance, with an AUC of 0.98 during training and 0.99 during OOD evaluation. Domain shift testing related to scanner model and self-reported race produced modest reductions across data representations, resulting in minimal differences

between approaches. Receiver-operating characteristic curves for all stress tests are shown in Figure 3, with corresponding confusion matrices in Supplementary Figures S9 and S10.

Pixel-based Model Explainability:

Occlusion-based class-contrast saliency maps were generated to reveal the source of pixel-embedded signals maximally contributing to disease classification. Across all pixel-based models, inconsistent and diffuse activation patterns were commonly observed, typically lacking anatomic localization (Supplementary Figure S11), suggesting a reliance on distributed pixel-level signals rather than discrete pathological markers of disease.

4D Phenomics-based Model Explainability:

SHAP-based evaluations of the 4D phenomics models demonstrated strong stability in feature importance rankings across ID and OOD cohorts. Patient-specific SHAP waterfall plots provided transparent, interpretable visualization of region-level features contributing to predicted probabilities. Figure 4 presents four representative patient cases selected from the sex domain shift OOD stress test. Waterfall plots illustrate the sequential contribution of region-level features used to derive the final prediction probability, demonstrated across a spectrum of phenotypes from Healthy individuals to patients with overt HCM, accompanied by their respective 4D phenomics models.

DISCUSSION

This study systematically evaluated susceptibility to shortcut learning and domain shift in AI models trained from medical imaging pixel data. Using a prospectively enrolled multi-national study cohort, we demonstrated that pixel-based models can readily learn non-causal vendor-

specific signals (i.e., vendor shortcut learning), compromising generalizability across new clinical environments. Despite balanced class exposures and standard data augmentation, performance degraded across all domain shift experiments. In contrast, standardized phenomics-based representations of the same source data preserved diagnostic performance while preventing model reliance on non-causal image features, resulting in robust generalizability across a range of external vendor, site and population scenarios. Collectively, these findings provide multicenter evidence that 4D phenomics-based representations of cardiac shape and motion provide a robust foundation for the exploration of generalizable AI-assisted diagnosis from routine cine CMR.

Sensitivities of AI models to learn acquisition-related image signatures are well recognized, reflecting systematic differences in intensity/contrast characteristics and noise texture profiles across vendors and hardware models. The vulnerabilities identified in this study are consistent with multimodal benchmarks reported by Guo, et al., demonstrating greater domain shift degradation in MRI-based classification versus CT or radiography derived models.[17] This has been attributed to greater variability in MRI acquisition parameters (e.g., repetition time, echo time, and flip angles), which introduce confounding signals that facilitate shortcut learning. Ly, et al., further showed that AI models frequently exploit latent acquisition biases, including non-semantic intensity distributions or noise profiles, as effective surrogates for diagnosis.[18] The pervasive nature of these manufacturer-specific signals was shown by Biondetti, et al., demonstrating how CNNs can distinguish scanner vendor from pixel data with near-perfect accuracy (>99%) even when imperceptible to expert readers.[19]

Importantly, we identified performance degradation even when scanner vendor was held constant and only scanner model differed (Stress Test 3). This indicates that acquisition-related signatures are not confined to specific vendors, but extend to specific hardware models or site-specific protocols. Unsupervised clustering of MRI data by field strength (1.5T versus 3T) and model (Prisma versus Skyra) has been previously reported by Kushol et al.[20] Previous studies have attempted to suppress these signals at the representation level by applying adversarial 'unlearning' while preserving disease-discriminative features. However, a key limitation is that bias sources must be a-priori identified for this to be engaged, limiting application to unknown or unmeasured biases.[21] In contrast, phenomics-driven transformations remove acquisition-dependent information through a standardized anatomical representation, reducing susceptibility to both measured and unrecognized biases without requiring their explicit identification during model development.

We observed pixel-based classification models to experience performance degradation under sex and race domain shifts, indicating sensitivity to demographic-associated image characteristics (e.g., body habitus, chest wall configuration, fat distribution, etc.). This observed vulnerability to sex domain shifts is consistent with the findings of Larrazabal, et al., who demonstrated that sex imbalance in training data produces biased classifiers.[22] In their large-scale chest X-ray study, models trained on single-sex cohorts experienced substantial reductions in classification performance when tested in the underrepresented sex. Our findings demonstrate that similar vulnerabilities extend to race-based domain shifts, consistent with the landmark study by Gichoya, et al., that revealed deep learning models can predict self-reported race from medical images with near-perfect performance (AUC 0.91–0.99).[23] Importantly, they demonstrated that these latent

signatures persisted across anatomic regions and image frequency spectrums, remaining detectable despite substantial image perturbation.

The bias mitigation strategy evaluated in this study sought to maximize phenotypic representation while minimizing reliance on non-causal image features. Across all stress tests, this approach substantially improved OOD generalizability and demonstrated resilience to shifts in scanner vendor, scanner model, sex, and self-reported race. Phenomics-based models showed relative immunity to permissive vendor shortcut learning and markedly reduced susceptibility to vendor domain shift, although a small residual performance reduction remained for the latter. One possible explanation is the upstream segmentation network that was developed predominantly using Siemens datasets, from which subtle vendor-related differences in segmentation accuracy may have propagated into the downstream 4D phenomics model. This interpretation is consistent with findings from the M&Ms Challenge, in which Campello *et al.* demonstrated persistent vendor-specific differences in cardiac MRI segmentation despite extensive data augmentation and domain adaptation strategies.[24]

Sex- and race-based stress testing similarly demonstrated a mitigation of OOD performance degradation using 4D phenomics-based representation, indicating the preferential learning of pathology-relevant features. Interestingly, female OOD testing showed a non-significant improvement in performance rather than degradation. One postulated explanation is greater phenotypic expression of HCM among women at time of clinical referral. Butters *et al.* reported that women with HCM are typically referred later than men and present with greater disease expression.[25] As 4D phenomics focuses on anatomic markers of disease expression, discrimination performance may have been robustly maintained (or augmented) due to greater expression of the

learned phenotype. Similarly, OOD race testing, while limited to reduced sample sizes, showed optimal OOD performance, demonstrating successful blinding to race-related confounders.

Our approach builds on the interpretable deep learning framework proposed by Biffi, *et al.*, who sought to isolate anatomical feature learning by replacing source pixel data with LV segmentation masks as input to a three-dimensional convolutional variational autoencoder (VAE).[26] This anatomy-centered approach improved external performance on the ACDC MICCAI dataset while enabling visualization of learned latent features through reconstruction of structural phenotypes. Extending this concept, we transformed deep learning–derived multiaxial, multiphase segmentation masks into a standardized finite element–based 4D phenomic model of cardiac structure and function. Rather than learning an implicit latent embedding, our framework converts source images into an explicit spatiotemporal phenomic representation that can be interrogated at the element, regional, or segmental level. The resulting tabular feature set supports conventional machine learning with transparent feature attribution (e.g., SHAP values) while enabling future graph neural network approaches operating directly on time-resolved 3D meshes.

Our study has several limitations. First, stress-testing required strict stratification across scanner vendor, site, field strength, disease status, sex, and race to isolate individual bias sources. Although derived from a large multicenter cohort, these constraints reduced sample sizes within individual partitions, increasing uncertainty around some performance estimates. Consequently, race could not be fully balanced in the hardware-focused stress tests (Stress Tests 1–3), leaving the possibility of residual demographic imbalance. Second, evaluation was limited to binary disease classification and should be extended to multiclass tasks. Third, although key demographic variables were considered, age, body mass index (BMI), and comorbidities were not explicitly controlled. Finally,

the 4D phenomic framework depends on an upstream segmentation network trained predominantly on Siemens datasets, which may contribute to residual vendor sensitivity. Expanding segmentation training with diverse multi-vendor datasets is an important direction for future work.

CONCLUSION

Pixel-based disease classification is vulnerable to learning non-causal image features related to scanner vendor, scanner model, sex, and race. Reliance on these signals can produce clinically meaningful performance degradation when models are deployed across new clinical environments and distribution shifts. In contrast, standardized 4D phenomic representations preserve disease classification performance while minimizing reliance on non-causal image features, substantially improving model generalizability. These findings support phenomic representations as a robust, interpretable foundation for bias-aware AI in medical imaging.

METHODS

Dataset

This study utilized a sub-cohort of the multicenter International AID-MRI study (NCT05793840); prospectively enrolling patients referred to clinical CMR services for known or suspected cardiomyopathy states. Study inclusion criteria and imaging protocols are described in the Supplementary Figure S1. At the time of data access, 6,960 of the planned 11,000 participants had been enrolled across eight international sites. This analysis included patients with a physician-adjudicated diagnosis of hypertrophic cardiomyopathy (HCM; n=644) and those determined to have a normal phenotype with no detectable cardiomyopathy or cardiac pathology, referred to in the analysis as Healthy (n=447). These respective cohorts were selected to support a well-defined disease classification task (HCM versus Healthy) through which the effect of data representation

on model generalizability could be assessed. To reduce class imbalance, the healthy cohort was augmented with 100 participants from the Cardiovascular Imaging Registry of Calgary (CIROC; NCT04367220). Baseline characteristics of the final cohort (n=1,091) are summarized in Supplementary Table S7. Protocol-defined standardized criteria used for disease adjudication across participating sites are provided in Supplementary Methods Figure S3.

All patient studies were de-identified and uploaded to a cloud data management platform supporting protocol compliance assessment, series labeling, data pre-processing and model training. Model inputs consisted of sequential short-axis (SAX) and multiplanar long-axis cine CMR images, or the corresponding 4D phenomic representations, plus disease labels derived from physician-adjudicated case report forms. All data pre-processing, inclusive of 4D phenomics, was performed with no user interaction to ensure a fair and transparent assessment of automated data management for AI-enabled diagnostics. The International AID-MRI study was approved by the University of Calgary Conjoint Health Research Ethics Board (CHREB; REB22-0319). Informed consent was obtained from participants in accordance with the approved protocol.

Study Design

To quantify the influence of data representation on susceptibility to shortcut learning and domain shift, sequential SAX and multiplanar LAX cine CMR datasets were presented to the same disease classification task in two forms: source pixel data and corresponding 4D finite element–encoded representations of the left ventricle (LV). As described in the Supplementary Methods, the latter encoded cardiac structure and function as spatiotemporal descriptors on an anatomically registered mesh of 480 finite elements.

Models trained from each representation were then subjected to identical shortcut learning and domain shift stress tests using bespoke partitions of the multicenter dataset to compare generalizability under out-of-distribution (OOD) conditions (Figure 2). Five stress tests were performed, each introducing a single shift between in-distribution (ID) and OOD data while balancing all other characteristics as closely as possible. Collectively, these experiments evaluated susceptibility to vendor-related shortcut learning and domain shifts associated with scanner vendor, scanner model, sex, and self-reported race.

Pixel-based disease classification models:

Pixel-based classification models were trained using two-dimensional convolutional neural networks (2D CNNs) to discriminate hypertrophic cardiomyopathy (HCM) from a Healthy phenotype using end-diastolic (ED) and end-systolic (ES) short-axis (SAX) and long-axis (LAX) cine CMR images. As detailed in the Supplementary Methods, standard preprocessing, image cropping, and data augmentation were performed before training. Models were trained using five-fold stratified cross-validation on the ID datasets, with patient-level predictions obtained by averaging four view-level outputs (one SAX stack and three LAX views). Performance was evaluated on the corresponding OOD datasets. To support model interpretability, class-contrast occlusion saliency maps[27] were generated to identify image regions contributing to model predictions.

4D phenomics-based classification models:

4D phenomics-based classification models were trained using tabular representations of phase-specific engineered features derived from the finite element mesh. Features included regional wall

thickness, myocardial volume, endocardial curvature, and peak deformation (minimum and maximum principal strain), each available at varying model resolutions. Region-level features (2 × 2 elements per region; 120 regions total) were selected to balance spatial resolution with feature dimensionality. Descriptions of engineered features are provided in the Supplementary Methods and Supplementary Figures S4 and S5, and Supplementary Table S1.

To match pixel-based models, only features engineered from the end-diastolic and end-systolic phases were included. Disease classification models were trained using a LightGBM classifier[28] with five-fold cross-validation on the in-distribution datasets and evaluated on the corresponding OOD datasets. Model interpretability was assessed using SHAP (SHapley Additive exPlanations) values, computed with the TreeExplainer algorithm[29] applied to the cross-validation fold achieving the median area under the receiver-operating characteristic curve (AUC). For each patient, SHAP values quantified the contribution of each region-level feature to the predicted probability of HCM relative to the model's average output across the training cohort.

Stress Test 1 - Vendor Shortcut Learning:

To assess susceptibility to vendor shortcut learning, 188 unique patients were available following a balancing of vendor, field strength, diagnosis, sex, and site. Models were trained using 128 study subjects (64 HCM scanned on Siemens 3T / 64 Healthy scanned on GE 3T), followed by OOD testing on 60 subjects acquired exclusively using Siemens 3T scanners (30 HCM and 30 Healthy). This test aimed to identify learned vendor features as a shortcut to disease classification.

Stress Test 2 - Vendor Domain Shift:

To assess susceptibility to vendor domain shift, 652 unique patients were available following a balancing of diagnosis, sex, site, and scanner model. Training was performed using 620 Siemens scanned subjects (310 HCM and 310 Healthy), followed by OOD testing using 32 GE-scanned subjects (16 HCM and 16 healthy).

Stress Test 3 - Scanner-Model Domain Shift:

To assess susceptibility to scanner model domain shift, 700 unique patients were available following a balancing of diagnosis, sex, and restriction to Siemens vendor. Training was performed using 620 study subjects (310 HCM and 310 Healthy scanned using Prisma or Skyra models), followed by OOD testing performed using 80 disease-balanced subjects scanned using Siemens Skyra Fit or Vida models.

Stress Test 4 - Sex Domain Shift:

To assess susceptibility to sex domain shift, 244 unique patients were available following a balancing of diagnosis, vendor, site, scanner model, and race. Training was performed using 160 male-only study subjects (80 HCM and 80 normal), followed by OOD testing in a similarly balanced cohort of 84 female-only subjects.

Stress Test 5 - Race Domain Shift:

To assess susceptibility to race domain shift, 200 unique patients were available following a balancing of diagnosis, vendor, site, scanner model, and sex. Training was performed using 168 White study subjects (84 HCM and 84 Healthy), followed by OOD testing in a similarly balanced cohort of 32 Asian subjects.

Detailed split definitions and sample stratification for each stress test are provided in the Supplementary Methods, Figures S7 and S8, and Tables S2–S6.

Statistical Analysis: Model performance was assessed using the Area Under the Receiver Operating Characteristic Curve (AUC), accuracy, F1-score, and Matthews Correlation Coefficient (MCC). All performance metrics are reported as the mean across 5-fold cross-validation with 95% Confidence Intervals (95% CIs) estimated across 100 bootstraps to quantify uncertainty. Additionally, confusion matrices were generated to visualize class-specific errors. Performance metrics were compared between pixel-based and 4D phenomic based classification models across all stress tests to evaluate their respective robustness, defined as performance degradation from ID to OOD testing.

FIGURES

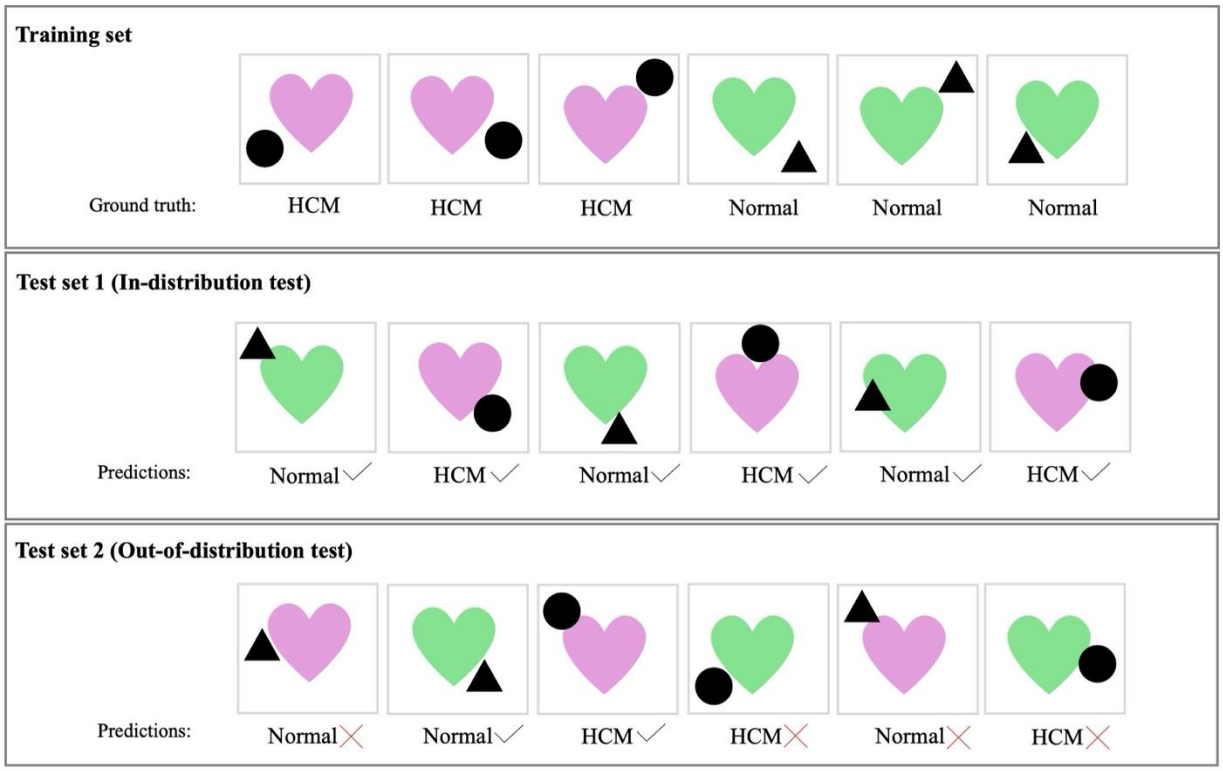

Figure 1. Shortcut learning and domain shift in image-based disease classification.

Schematic illustrating the relationship between shortcut learning and domain shift in supervised disease classification. During training (top), disease labels are spuriously associated with non-causal image features (black circle or triangle), representing latent confounders such as scanner vendor, acquisition protocol, sex, or race. The resulting model achieves high performance on in-distribution (ID) test data (middle), where the same associations are preserved. During out-of-distribution (OOD) testing (bottom), these associations are disrupted, exposing reliance on non-causal features and resulting in degraded classification performance despite unchanged underlying pathology. The schematic illustrates how strong ID performance may conceal shortcut learning that becomes apparent only under domain shift.

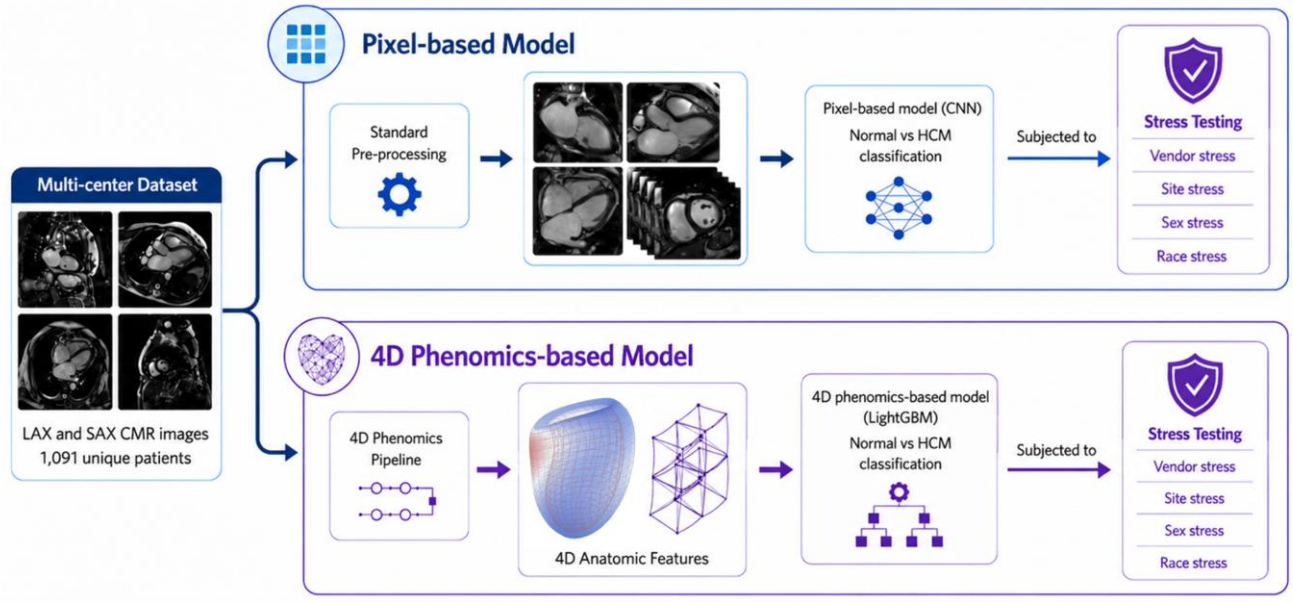

Figure 2. Study design and comparative evaluation of pixel-based and phenomic classification models. A total of 1,091 cine cardiac magnetic resonance (CMR) studies were processed by two parallel classification pipelines. **Top:** A pixel-based model received preprocessed long- and short-axis cine images as input to a convolutional neural network (CNN) for classification of hypertrophic cardiomyopathy (HCM) versus Healthy. **Bottom:** A 4D Phenomics-based model receives the same cine image data and transforms these into a standardized bi-layer four-dimensional (4D) finite element model, followed by the generation of engineered architectural and biomechanical features. These are used to train an explainable machine learning model for the same classification task. Both pipelines were evaluated by identical stress-testing protocols to compare their susceptibility to vendor shortcut learning and domain shifts related to scanner vendor, scanner model, sex, and self-reported race.

Table 1. Comparative performance of pixel-based and 4D phenomic-based disease classification models. Performance is reported for in-distribution (ID) and out-of-distribution (OOD) cohorts (metrics shown as point estimate with 95% CI).

Stress Tests	Metrics	Pixel-based Classifier		4D Phenomics-based Classifier	
		ID	OOD	ID	OOD
Vendor Shortcut learning	AUC	0.91 (0.86–0.95)	0.52 (0.37–0.67)	0.99 (0.97–1.00)	0.99 (0.97–1.00)
	ACC	0.80 (0.73–0.87)	0.48 (0.43–0.53)	0.94 (0.90–0.98)	0.95 (0.88–1.00)
	F1- macro	0.80 (0.73–0.87)	0.35 (0.30–0.43)	0.94 (0.89–0.98)	0.95 (0.88–1.00)
	MCC	0.59 (0.45–0.73)	−0.08 (−0.27–0.19)	0.88 (0.80–0.95)	0.90 (0.79–1.00)
Vendor domain shift	AUC	0.98 (0.97–0.99)	0.88 (0.75–0.98)	0.98 (0.97–0.99)	0.96 (0.87–1.00)
	ACC	0.94 (0.92–0.96)	0.81 (0.69–0.94)	0.94 (0.92–0.96)	0.86 (0.71–0.96)
	F1- macro	0.94 (0.92–0.96)	0.81 (0.68–0.94)	0.94 (0.92–0.96)	0.86 (0.71–0.96)
	MCC	0.88 (0.85–0.92)	0.63 (0.38–0.88)	0.88 (0.84–0.91)	0.72 (0.45–0.93)
Scanner-Model domain shift	AUC	0.98 (0.98–0.99)	0.93 (0.87–0.98)	0.98 (0.97–0.99)	0.96 (0.90–0.99)
	ACC	0.94 (0.92–0.95)	0.82 (0.74–0.91)	0.94 (0.92–0.96)	0.89 (0.81–0.95)
	F1- macro	0.94 (0.92–0.95)	0.82 (0.74–0.91)	0.94 (0.92–0.96)	0.89 (0.81–0.95)
	MCC	0.87 (0.84–0.91)	0.65 (0.48–0.83)	0.88 (0.84–0.91)	0.78 (0.63–0.90)
Sex domain shift	AUC	0.97 (0.94–0.99)	0.88 (0.81–0.95)	0.98 (0.96–0.99)	0.99 (0.97–1.00)
	ACC	0.94 (0.91–0.98)	0.80 (0.71–0.88)	0.92 (0.88–0.96)	0.94 (0.89–0.99)
	F1- macro	0.94 (0.91–0.98)	0.80 (0.71–0.88)	0.92 (0.88–0.96)	0.94 (0.89–0.99)
	MCC	0.89 (0.82–0.95)	0.60 (0.45–0.77)	0.85 (0.78–0.93)	0.89 (0.80–0.98)
Race domain shift	AUC	0.95 (0.92–0.98)	0.93 (0.84–0.99)	0.97 (0.95–0.99)	1.00 (1.00–1.00)
	ACC	0.89 (0.85–0.93)	0.81 (0.69–0.94)	0.92 (0.88–0.96)	0.97 (0.91–1.00)
	F1-macro	0.89 (0.85–0.93)	0.81 (0.69–0.94)	0.92 (0.88–0.96)	0.97 (0.91–1.00)
	MCC	0.79 (0.69–0.87)	0.62 (0.38–0.88)	0.85 (0.76–0.93)	0.94 (0.83–1.00)

Abbreviations: AUC, area under the receiver operating characteristic curve; ACC, accuracy; F1-macro, macro-averaged F1-score; MCC, Matthews correlation coefficient; ID, in-distribution; OOD, out-of-distribution; CI, confidence interval; HCM, hypertrophic cardiomyopathy.

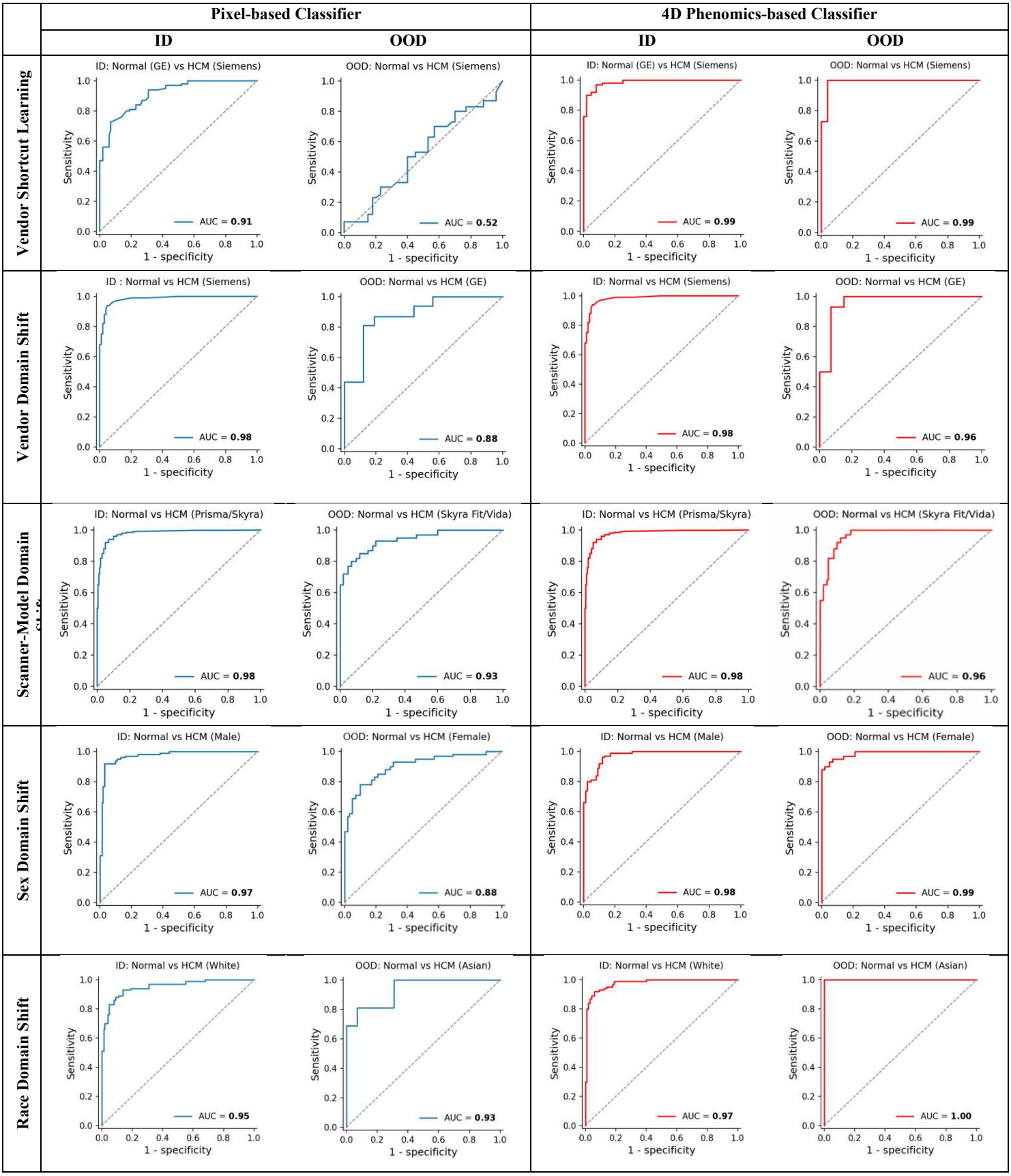

Figure 3. Comparative performance of pixel-based and 4D phenomics-based disease classification models across five stress tests. AUC is reported for both ID and OOD.

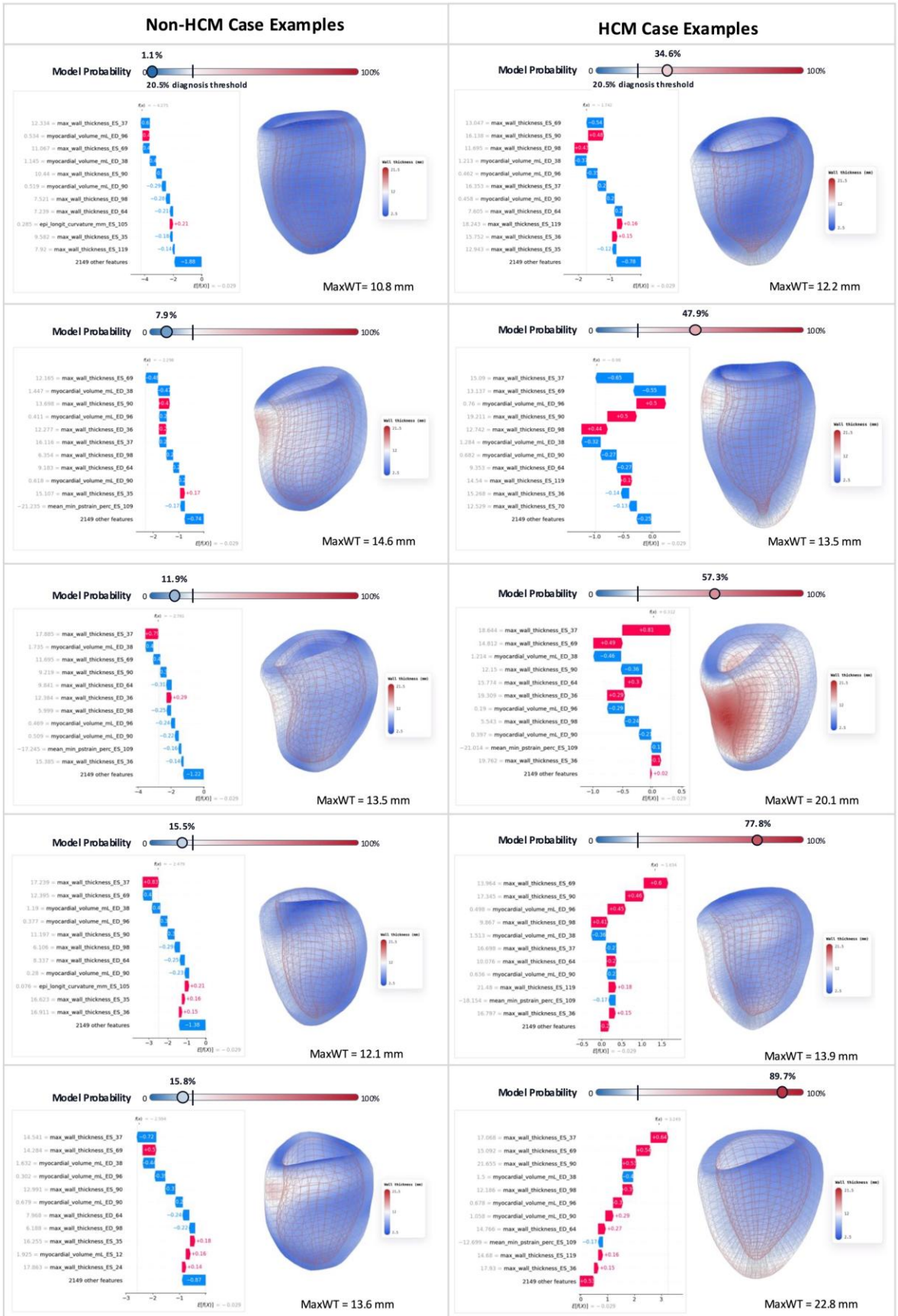

Figure 4. Representative 4Deep ventricular phenotypes and SHAP explanations across the prediction spectrum.

Out-of-distribution cases arranged by model-predicted probability relative to the prespecified 20.5% diagnostic threshold. Examples extracted from sex domain shift model ranging from low to high HCM prediction confidence for patients with Non-HCM and HCM ground truth labels. Each patient-level SHAP waterfall plot is shown with its corresponding 4Deep representation, shown at end-diastole and colour-coded by regional wall thickness. SHAP features shifting the prediction toward Non-HCM shown in blue, and toward HCM in red. Global maximal end-diastolic wall thickness (MaxWT) is provided as a reference, illustrating poor congruence between model confidence and historic diagnostic MaxWT thresholds for HCM.

Data Availability

The data used in this study are not publicly available as they contain sensitive health information and are subject to participant privacy, institutional ethics, and data-use restrictions. Access may be considered upon reasonable request to the corresponding author, subject to research ethics approval, institutional authorization, and execution of applicable data-sharing agreements. The related source code for 4D phenomics modelling and disease classification is not currently available as it remains under active development and validation as part of the ongoing multi-centre study. Future availability will be determined following completion of the study, in accordance with institutional policies and applicable intellectual property arrangements.

Author Contributions

FH led the research and was responsible for the study design, methodology, data curation, cardiac MRI image analysis, machine learning and deep learning analyses, statistical analysis, interpretation of results, and manuscript preparation. JW led the study and supervised the research, providing guidance on the conceptual framework, methodology, research direction, and manuscript development. RM developed the 4D-phenomics and feature extraction pipeline and provided guidance on the research direction. CJ contributed to software development, segmentation, and research direction. ZG provided technical and methodological support during machine learning model development. DL contributed to cohort curation and technical aspects of image review and contour adjudication. JA, TB, JF, and SR provided technical and administrative support. CBJ and RM provided guidance on the research direction. KH, MYN, JG, EN, NABN, DIP, IR, and MV contributed as principal investigators to the international AID-MRI study, including its clinical and multicentre research activities. All authors contributed to the interpretation of the findings, critically reviewed the manuscript, and approved the final version.

Funding: This study was funded by the Canadian Institutes of Health Research (CIHR)

Competing Interests: All authors declare no competing interests related to this work.

Acknowledgements

This study was supported by the Nelson PULSE Centre for Innovation at the Cumming School of Medicine. The authors acknowledge the Stephenson Cardiac Imaging Centre and the Personalized Diagnostics Program, which is supported by the Libin Precision Medicine Initiative (LPMI) of the Libin Cardiovascular Institute. We also acknowledge the contributions of the International AID-MRI study teams and participating centres in facilitating the multicentre research infrastructure and data collection.

REFERENCES

1. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022 Jan;28(1):31–8. doi:10.1038/s41591-021-01614-0

2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019 Jan;25(1):44–56. doi:10.1038/s41591-018-0300-7

3. Castro DC, Walker I, Glocker B. Causality matters in medical imaging. Nat Commun. 2020 Jul 22;11(1):3673. doi:10.1038/s41467-020-17478-w

4. Hasanzadeh F, Josephson CB, Waters G, Adedinsewo D, Azizi Z, White JA. Bias recognition and mitigation strategies in artificial intelligence healthcare applications. npj Digit Med. 2025 Mar 11;8(1):154. doi:10.1038/s41746-025-01503-7

5. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med. 2021 Dec;27(12):2176–82. doi:10.1038/s41591-021-01595-0

6. Lotter W. Acquisition parameters influence AI recognition of race in chest x-rays and mitigating these factors reduces underdiagnosis bias. Nat Commun. 2024 Aug 29;15(1):7465. doi:10.1038/s41467-024-52003-3

7. Roschewitz M, Khara G, Yearsley J, Sharma N, James JJ, Ambrózay É, et al. Automatic correction of performance drift under acquisition shift in medical image classification. Nat Commun. 2023 Oct 19;14(1):6608. doi:10.1038/s41467-023-42396-y

8. Glocker B, Jones C, Bernhardt M, Winzeck S. Algorithmic encoding of protected characteristics in chest X-ray disease detection models. eBioMedicine. 2023 Mar 1;89. doi:10.1016/j.ebiom.2023.104467 PubMed PMID: 36791660.

9. Brown A, Tomasev N, Freyberg J, Liu Y, Karthikesalingam A, Schrouff J. Detecting shortcut learning for fair medical AI using shortcut testing. Nat Commun. 2023 Jul 18;14(1):4314. doi:10.1038/s41467-023-39902-7

10. DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. Nat Mach Intell. 2021 Jul;3(7):610–9. doi:10.1038/s42256-021-00338-7

11. Perkonigg M, Hofmanninger J, Herold CJ, Brink JA, Pianykh O, Prosch H, et al. Dynamic memory to alleviate catastrophic forgetting in continual learning with medical imaging. Nat Commun. 2021 Sep 28;12(1):5678. doi:10.1038/s41467-021-25858-z

12. Zhao Q, Adeli E, Pohl KM. Training confounder-free deep learning models for medical applications. Nat Commun. 2020 Nov 26;11(1):6010. doi:10.1038/s41467-020-19784-9

13. Qiao M, McGurk KA, Wang S, Matthews PM, O'Regan DP, Bai W. A personalized time-resolved 3D mesh generative model for unveiling normal heart dynamics. Nat Mach Intell.

2025;7(5):800–11. doi:10.1038/s42256-025-01035-5 PubMed PMID: 40416329; PubMed Central PMCID: PMC12101970.

14. Meng Q, Bai W, O'Regan DP, Rueckert D. DeepMesh: Mesh-Based Cardiac Motion Tracking Using Deep Learning. IEEE Trans Med Imaging. 2024 Apr;43(4):1489–500. doi:10.1109/TMI.2023.3340118 PubMed PMID: 38064325; PubMed Central PMCID: PMC7615801.

15. Ma Q, Meng Q, Wu Y, Wang S, Qiao M, Niederer S, et al. Learning a dynamic four-chamber shape model of the human heart for 95,695 UK Biobank participants [Internet]. arXiv; 2026 [cited 2026 Aug 9]. Available from: http://arxiv.org/abs/2603.28711 doi:10.48550/arXiv.2603.28711

16. Meng Q, Qin C, Bai W, Liu T, de Marvao A, O'Regan DP, et al. MulViMotion: Shape-Aware 3D Myocardial Motion Tracking From Multi-View Cardiac MRI. IEEE Trans Med Imaging. 2022 Aug;41(8):1961–74. doi:10.1109/TMI.2022.3154599 PubMed PMID: 35201985; PubMed Central PMCID: PMC7613225.

17. Guo B, Lu D, Szumel G, Gui R, Li P, Wang T, et al. The impact of scanner domain shift on deep learning performance in medical imaging: an experimental study. Int J Comput Assist Radiol Surg. 2026 Apr 30. doi:10.1007/s11548-026-03655-7 PubMed PMID: 42062639.

18. Ong Ly C, Unnikrishnan B, Tadic T, Patel T, Duhamel J, Kandel S, et al. Shortcut learning in medical AI hinders generalization: method for estimating AI model generalization without external data. NPJ Digit Med. 2024 May 14;7(1):124. doi:10.1038/s41746-024-01118-4 PubMed PMID: 38744921; PubMed Central PMCID: PMC11094145.

19. Biondetti GP, Gauriau R, Bridge CP, Lu C, Andriole KP. "Name that manufacturer". Relating image acquisition bias with task complexity when training deep learning models: experiments on head CT [Internet]. arXiv; 2020 [cited 2026 Jul 10]. Available from: http://arxiv.org/abs/2008.08525 doi:10.48550/arXiv.2008.08525

20. Kushol R, Parnianpour P, Wilman AH, Kalra S, Yang YH. Effects of MRI scanner manufacturers in classification tasks with deep learning models. Sci Rep. 2023 Oct 5;13(1):16791. doi:10.1038/s41598-023-43715-5

21. Souza R, Stanley EAM, Gulve V, Moore J, Kang C, Camicioli R, et al. HarmonyTM: multi-center data harmonization applied to distributed learning for Parkinson's disease classification. J Med Imaging (Bellingham). 2024 Sep;11(5):054502. doi:10.1117/1.JMI.11.5.054502 PubMed PMID: 39308760; PubMed Central PMCID: PMC11413651.

22. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc Natl Acad Sci U S A. 2020 Jun 9;117(23):12592–4. doi:10.1073/pnas.1919012117 PubMed PMID: 32457147; PubMed Central PMCID: PMC7293650.

23. Gichoya JW, Banerjee I, Bhimireddy AR, Burns JL, Celi LA, Chen LC, et al. AI recognition of patient race in medical imaging: a modelling study. The Lancet Digital Health. 2022 Jun 1;4(6):e406–14. doi:10.1016/S2589-7500(22)00063-2 PubMed PMID: 35568690.

24. Campello VM, Gkontra P, Izquierdo C, Martin-Isla C, Sojoudi A, Full PM, et al. Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M&Ms Challenge. IEEE Trans Med Imaging. 2021 Dec;40(12):3543–54. doi:10.1109/TMI.2021.3090082 PubMed PMID: 34138702.

25. Butters A, Lakdawala NK, Ingles J. Sex Differences in Hypertrophic Cardiomyopathy: Interaction With Genetics and Environment. Curr Heart Fail Rep. 2021 Oct;18(5):264–73. doi:10.1007/s11897-021-00526-x PubMed PMID: 34478112; PubMed Central PMCID: PMC8484093.

26. Biffi C, Oktay O, Tarroni G, Bai W, De Marvao A, Doumou G, et al. Learning Interpretable Anatomical Features Through Deep Generative Models: Application to Cardiac Remodeling. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G, editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Cham: Springer International Publishing; 2018. p. 464–71. doi:10.1007/978-3-030-00934-2_52

27. Zeiler MD, Fergus R. Visualizing and Understanding Convolutional Networks. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. Computer Vision – ECCV 2014. Cham: Springer International Publishing; 2014. p. 818–33. doi:10.1007/978-3-319-10590-1_53

28. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2017 [cited 2026 Jul 10]. Available from: https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

29. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020 Jan;2(1):56–67. doi:10.1038/s42256-019-0138-9

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.docx