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.

GS-ECG: a generalist–specialist approach to structural heart disease prediction from electrocardiograms

Jiacong Mi

j.mi.1@research.gla.ac.uk

University of Glasgow

Clarisse F. de Vries

University of Glasgow

Derek T. Connelly

University of Glasgow

Honghan Wu

University of Glasgow

Article

Keywords:

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11566820/v1

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

Additional Declarations: No competing interests reported.

GS-ECG: a generalist–specialist approach to structural heart disease prediction from electrocardiograms

Jiacong Mi[1*], Clarisse F. de Vries[1], Derek T. Connelly[1], Honghan Wu[1*]

[1]School of Health and Wellbeing, University of Glasgow, Glasgow, United Kingdom.

*Corresponding author(s). E-mail(s): j.mi.1@research.gla.ac.uk; Honghan.Wu@glasgow.ac.uk;

Abstract

Structural heart disease (SHD) often goes undiagnosed, and echocardiography, the standard test, cannot be offered to every patient promptly. AI analysis of routine electrocardiograms (ECGs) could help prioritize who is scanned first. ECG foundation models promise broad applicability across ECG tasks, but SHD abnormalities are novel targets, defined by echocardiography and outside the labels on which these models were pretrained, so how best to use them is unclear. Rather than adapting the foundation model to these targets, we propose GS-ECG, a synergistic generalist–specialist approach that keeps it frozen and pairs it with a compact specialist, DGRNet, trained on novel SHD targets. In this setup, the generalist summarizes each recording in long-context features, while DGRNet focuses on SHD-specific waveform morphology at beat-level resolution, with a receptive field of about two cardiac cycles per feature. We evaluated GS-ECG in two health systems: MIMIC-IV-Ext-SHD (49,229 patients), a new public benchmark, and the EchoNext-Mini Dataset (36,286 patients). Full fine-tuning of the foundation model improved mean AUPRC by only 1.3–3.6% over freezing it, whereas DGRNet alone exceeded fine-tuning by 6.1–6.4%. Combining the frozen generalist with DGRNet in GS-ECG performed best, exceeding fine-tuning by 8.1% in both datasets while training 4.9% as many parameters; its composite SHD AUROC was 0.825 and 0.839. Against the published EchoNext-Mini Model on its own dataset, GS-ECG found 17–25% more patients with aortic stenosis, mitral regurgitation or tricuspid regurgitation among the top 5% ranked

for echocardiography. Pairing frozen foundation models with purpose-built specialists offers an efficient route to clinical targets that lie beyond pretraining labels.

1 Introduction

Structural heart disease (SHD), including impaired ventricular function, thickened ventricular walls, and diseased heart valves, can progress to heart failure and other adverse outcomes, yet many affected patients remain undiagnosed [1, 2]. Transthoracic echocardiography (TTE) is the standard test for these abnormalities and guides their management [3–6]. However, TTE requires dedicated equipment, trained sonographers, and specialist interpretation, so not every patient can be scanned promptly, and deciding who should be scanned first becomes a practical clinical problem. The electrocardiogram (ECG), by contrast, is inexpensive, quick, and recorded routinely, and its signals reflect cardiac structure and function [7, 8]. A model that estimates TTE abnormalities from a routine ECG could therefore direct limited imaging capacity to the patients who need it most and identify those in whom a scan would rarely find disease.

Deep learning has shown that ECGs carry more information than conventional reading captures: AI-enabled ECGs can identify left ventricular dysfunction [9] and valvular disease [10, 11], and ECG-based screening has been evaluated prospectively in clinical settings [12–14]. Recent studies have moved from single conditions to SHD as a whole: rECHOmmend estimated the risk of undiagnosed SHD [1], PRESENT-SHD screened for SHD using ECG images [15], and EchoNext detected a composite of TTE-derived conditions with an area under the receiver operating characteristic curve (AUROC) of 0.852 on its NYP (NewYork-Presbyterian) multicentre internal test set [2]. A later study released the EchoNext-Mini Dataset with a compact baseline model, referred to here as the EchoNext-Mini Model, which reached an AUROC of 0.820 for composite SHD [16]. These models were trained directly on echocardiography-derived labels for the task at hand.

ECG foundation models such as ECGFounder, pretrained on more than 10 million ECGs [17], are attracting growing attention because they learn broad representations that transfer across many diagnostic tasks [18]. They are, however, trained on general diagnostic statements, whereas SHD abnormalities are defined by echocardiography. Full fine-tuning adapts tens of millions of parameters to the task but is costly and can distort what was learned during pretraining [19]. Keeping the backbone frozen preserves those representations but adapts only through small task-specific layers. Alternatively, a network designed for the task can learn features tailored to TTE-derived abnormalities with far fewer parameters, but only from the labelled ECGs available, raising the question of whether a foundation model is needed at all. We hypothesized that a task-specific network and a frozen foundation model are complementary: the specialist learns the task, while the frozen model contributes

representations learned at scale. Combining pretrained and task-adapted models has improved accuracy and robustness in general-domain transfer learning [20].

Comparing these strategies also requires a common evaluation setting. Reported performance depends on how ECGs are linked to TTEs, how findings from repeated TTEs are combined, and whether patients overlap between training and test data, and SHD definitions and prevalence differ even between EchoNext and rECHOmmend [2]. Public benchmarks such as PTB-XL have made such choices explicit for ECG classification [21, 22]. Evaluation should also look beyond a composite endpoint, which can appear strong even when less common abnormalities are missed: the area under the precision–recall curve (AUPRC) is more informative than the AUROC when positive cases are few [23, 24], and overall metrics do not show how many affected patients a model would find when only a fixed number of TTEs can be performed.

We therefore developed GS-ECG and constructed MIMIC-IV-Ext-SHD, a patient-separated SHD benchmark built from the public MIMIC-IV database of Beth Israel Deaconess Medical Center and its linked ECG and echocardiography modules [25]. GS-ECG pairs a compact Dilated Gated Residual Network (DGRNet), designed for multi-label SHD prediction and trained from scratch on TTE-derived labels, with a frozen ECGFounder generalist, and combines their calibrated outputs with a separate weight for each abnormality. The study makes four contributions. First, we address how an ECG foundation model should be used for echocardiography-defined targets: across two health systems, full fine-tuning of ECGFounder improved only slightly on keeping it frozen, a compact specialist outperformed full fine-tuning, and the frozen foundation model still improved the specialist when the two were combined. Second, GS-ECG provides a parameter-efficient generalist–specialist design. Its specialist, DGRNet, is designed around beat-level waveform morphology, with each feature spanning about two cardiac cycles, and with fewer than one million parameters it outperformed all generic waveform baselines. Combined with the frozen generalist, GS-ECG trained 4.9% as many parameters as full fine-tuning, which trained 31.5 million, and outperformed the EchoNext-Mini Model, including the authors' released checkpoint on its own dataset. Third, MIMIC-IV-Ext-SHD provides patient-separated training, validation, and test sets under a fixed protocol, so that multi-label SHD models can be developed and compared on public data. Fourth, we show how model scores can guide echocardiography referral, identifying both the patients to scan first and those in whom a scan would rarely find disease.

2 Results

We compared four ways of using an ECG foundation model for TTE-derived outcomes (Figure 1a). ECGFounder-FT fine-tuned all parameters of ECGFounder; ECGFounder-Frozen kept its backbone frozen and trained task modules only; DGRNet replaced the foundation model with a compact specialist trained from scratch; and GS-ECG combined the frozen generalist with the specialist through label-wise weights fitted on held-out training patients (Figure 1b). The three single predictors received the same inputs and training patients and were selected by the same validation criterion. We also compared the EchoNext-Mini Model and six reference models that

used structured variables only or waveforms only. We constructed MIMIC-IV-Ext-SHD to compare these strategies under a fixed, patient-separated protocol. Models were developed and tested in MIMIC-IV-Ext-SHD and, with the same protocol, in the EchoNext-Mini Dataset from a second health system, where the authors' released EchoNext-Mini Model allowed a direct comparison with a published model (Figure 1c and Table 1). The two test sets comprised 7,385 and 5,442 patients with one ECG each, and component-label prevalence ranged from 0.4% to 28.7% and from 0.4% to 19.5%, respectively. The primary endpoint was mean AUPRC across all outputs, including composite SHD, with 95% confidence intervals (CIs) from 1,000 patient-level bootstrap resamples.

Table 1 MIMIC-IV-Ext-SHD cohort characteristics for all paired ECGs and the model-ready splits.

Variable	All paired ECGs	Training	Validation	Test
Cohort size				
ECG records, n	254,966	173,504	7,384	7,385
Patients, n	49,229	34,460	7,384	7,385
Demographics				
Age (years)	68 (57, 79)	68 (57, 79)	69 (57, 80)	69 (57, 80)
Male sex, n (%)	136,612 (53.6%)	92,878 (53.5%)	3,844 (52.1%)	3,898 (52.8%)
Race or ethnicity, n (%)				
White	174,883 (68.6%)	117,846 (67.9%)	5,071 (68.7%)	5,016 (67.9%)
Black	38,716 (15.2%)	26,610 (15.3%)	905 (12.3%)	894 (12.1%)
Hispanic	13,333 (5.2%)	9,419 (5.4%)	295 (4.0%)	304 (4.1%)
Other	8,628 (3.4%)	5,983 (3.4%)	240 (3.3%)	264 (3.6%)
Asian	7,473 (2.9%)	5,098 (2.9%)	191 (2.6%)	225 (3.0%)
Unknown	11,933 (4.7%)	8,548 (4.9%)	682 (9.2%)	682 (9.2%)
TTE-derived SHD labels, n (%)				
LVEF $\leq$ 45%	70,422 (27.6%)	45,812 (26.4%)	1,564 (21.2%)	1,538 (20.8%)
LV wall thickness $\geq$ 1.3 cm	77,587 (30.4%)	49,235 (28.4%)	2,044 (27.7%)	2,121 (28.7%)
Aortic stenosis	7,886 (3.1%)	4,902 (2.8%)	199 (2.7%)	215 (2.9%)
Aortic regurgitation	1,425 (0.6%)	1,003 (0.6%)	25 (0.3%)	29 (0.4%)
Mitral regurgitation	14,118 (5.5%)	9,215 (5.3%)	276 (3.7%)	277 (3.8%)
Tricuspid regurgitation	17,384 (6.8%)	11,203 (6.5%)	368 (5.0%)	382 (5.2%)
Pulmonary regurgitation	7,382 (2.9%)	4,776 (2.8%)	169 (2.3%)	157 (2.1%)
RV systolic dysfunction	28,711 (11.3%)	18,099 (10.4%)	552 (7.5%)	580 (7.9%)
Pericardial effusion	7,631 (3.0%)	4,850 (2.8%)	126 (1.7%)	103 (1.4%)
TR maximum velocity $\geq$ 3.2 m/s	50,171 (19.7%)	32,314 (18.6%)	1,204 (16.3%)	1,142 (15.5%)
Composite SHD	151,253 (59.3%)	97,159 (56.0%)	4,135 (56.0%)	4,136 (56.0%)
ECG measurements				
Ventricular rate (bpm)	77 (65, 93)	77 (65, 93)	77 (65, 92)	76 (65, 91)
PR interval (ms)	160 (142, 182)	160 (142, 182)	158 (141, 180)	160 (141, 180)
QRS duration (ms)	95 (86, 111)	94 (86, 110)	94 (86, 108)	94 (86, 108)
QTc (ms)	450 (428, 477)	450 (428, 476)	446 (425, 473)	447 (426, 473)
TTE measurements				
LVEF (%)	55 (45, 60)	55 (45, 60)	55 (50, 62)	55 (50, 62)
Interventricular septal thickness (cm)	1.1 (1.0, 1.3)	1.1 (1.0, 1.2)	1.1 (1.0, 1.3)	1.1 (1.0, 1.3)
Inferolateral/posterior wall thickness (cm)	1.1 (1.0, 1.2)	1.1 (1.0, 1.2)	1.1 (1.0, 1.2)	1.1 (1.0, 1.2)
LV wall thickness (cm)	1.1 (1.0, 1.3)	1.1 (1.0, 1.3)	1.1 (1.0, 1.3)	1.1 (1.0, 1.3)
TR maximum velocity (m/s)	2.7 (2.4, 3.1)	2.7 (2.4, 3.0)	2.6 (2.3, 3.0)	2.6 (2.3, 3.0)

Note. Values are medians with interquartile ranges for continuous variables and counts with percentages for categorical variables. The "All paired ECGs" column comprises 254,966 records before split-specific record selection, and the split columns describe the model-ready records, with one ECG per validation or test patient. Valvular disease is moderate or greater; pericardial effusion is moderate or large. LV wall thickness is the larger of the interventricular septal and inferolateral/posterior wall thicknesses on each TTE.
Abbreviations: bpm, beats per minute; ECG, electrocardiogram; LV, left ventricular; LVEF, left ventricular ejection fraction; QTc, corrected QT interval; RV, right ventricular; SHD, structural heart disease; TR, tricuspid regurgitation; TTE, transthoracic echocardiography.

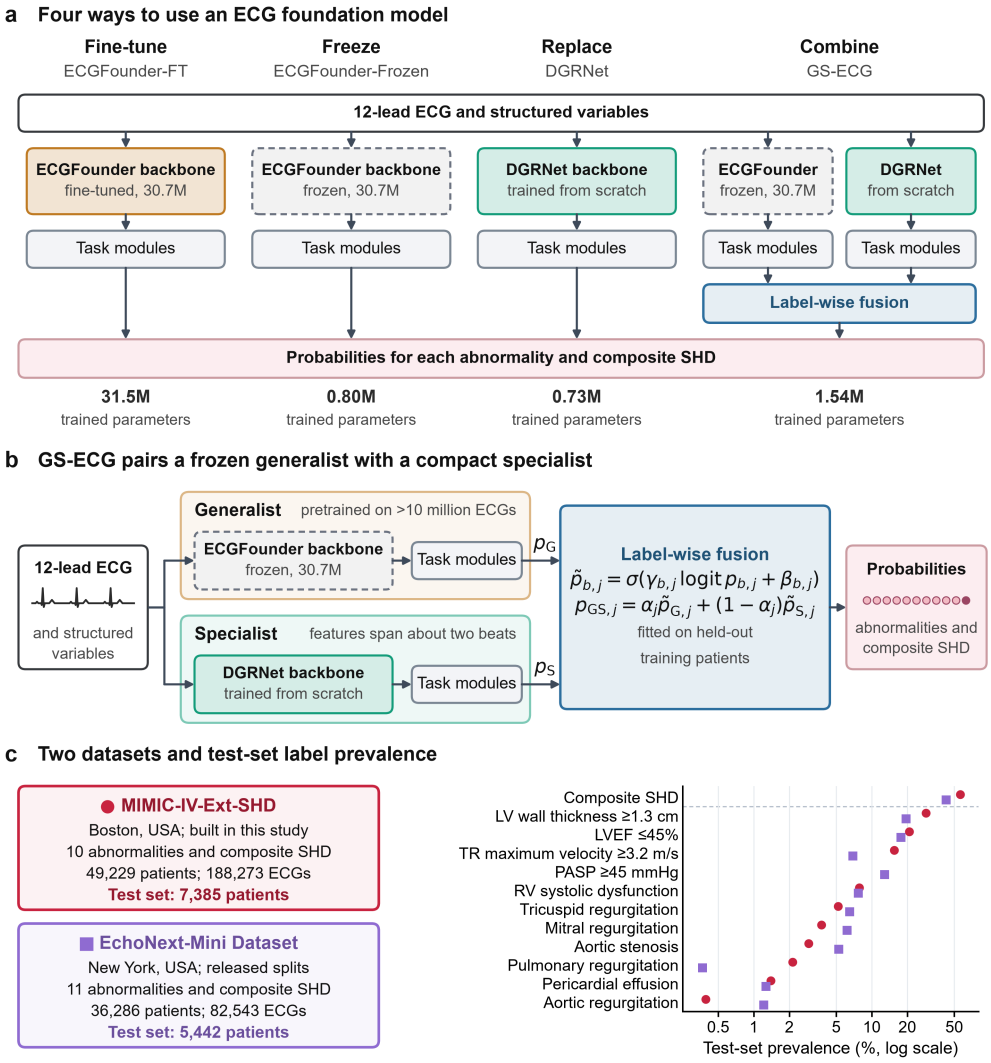


Fig. 1 Study overview. Panel a shows four ways of using an ECG foundation model for TTE-derived outcomes: fine-tuning all of its parameters, ECGFounder-FT; keeping its backbone frozen and training task modules only, ECGFounder-Frozen; a compact specialist trained from scratch, DGRNet; and the frozen generalist combined with the specialist, GS-ECG. Dashed boxes denote frozen weights, and trained-parameter counts are for MIMIC-IV-Ext-SHD. Panel b shows GS-ECG: the frozen ECGFounder backbone and DGRNet, each with its own task modules, predict every output, each predictor's probabilities are calibrated with a slope $\gamma_{b,j}$ and intercept $\beta_{b,j}$ for predictor $b \in \{\mathrm{G}, \mathrm{S}\}$ and output j, and the calibrated probabilities are combined with a separate weight α_j for each output, all fitted on held-out training patients. Panel c summarizes MIMIC-IV-Ext-SHD and the EchoNext-Mini Dataset and plots the test-set prevalence of each output on a log scale; PASP $\geq$45 mmHg is labelled only in the EchoNext-Mini Dataset. ECG, electrocardiogram; G, generalist; LV, left ventricular; LVEF, left ventricular ejection fraction; PASP, pulmonary artery systolic pressure; RV, right ventricular; S, specialist; SHD, structural heart disease; TR, tricuspid regurgitation; TTE, transthoracic echocardiography.

2.1 A compact specialist outperformed fine-tuning of an ECG foundation model

DGRNet, trained from scratch with fewer than one million parameters, had higher mean AUPRC than ECGFounder-FT, which trained 31.5 million, in both datasets (Figure 2). The differences were 0.0190 on MIMIC-IV-Ext-SHD and 0.0207 on the EchoNext-Mini Dataset, with 95% CIs of 0.0057–0.0325 and 0.0097–0.0369, and DGRNet also had higher mean AUROC in both datasets. Updating the backbone raised mean AUPRC over the frozen model by 0.0107 and 0.0040, less than DGRNet's margin over full fine-tuning, and did not raise mean AUROC on MIMIC-IV-Ext-SHD.

Training from scratch did not by itself explain this advantage. DGRNet had higher mean AUPRC than every other model trained from scratch in both datasets: the waveform-only CNN1D, CRNN, and Transformer models reached 0.2800 to 0.2924 on MIMIC-IV-Ext-SHD and 0.3017 to 0.3216 on the EchoNext-Mini Dataset, and the tabular-only models 0.2278 to 0.2362 and 0.2171 to 0.2337 (Supplementary Tables S1 and S2). On MIMIC-IV-Ext-SHD, ECGFounder-Frozen also exceeded all of these models, and its margin over the retrained EchoNext-Mini Model, which received the same inputs, was 0.0168 (95% CI 0.0084–0.0242).

2.2 The frozen foundation model added complementary information

Adding the frozen generalist improved DGRNet in both datasets (Figures 2 and 3). GS-ECG had the highest point estimate for each of the four overall metrics (Supplementary Tables S1 and S2), and its mean AUPRC of 0.3353 on MIMIC-IV-Ext-SHD and 0.3487 on the EchoNext-Mini Dataset exceeded that of DGRNet by 0.0060 and 0.0053, with 95% CIs of 0.0024–0.0103 and 0.0004–0.0090. It also exceeded ECGFounder-Frozen by 0.0358 and 0.0300 and ECGFounder-FT by 0.0251 and 0.0260, with all paired 95% CIs excluding zero. The generalist added 0.80 million trained parameters, so GS-ECG trained 1.54 million, 4.9% of the 31.5 million trained by ECGFounder-FT; at inference it also runs the frozen 30.7-million-parameter backbone, for 32.2 million parameters, close to the 31.5 million of ECGFounder-FT.

The gain was shared across outputs. GS-ECG had higher AUPRC and AUROC than DGRNet for all 11 MIMIC-IV-Ext-SHD outputs, including the two rarest, aortic regurgitation and pericardial effusion, with a median AUPRC gain of 0.0060, and for 10 of 12 EchoNext-Mini Dataset outputs on each metric. It exceeded ECGFounder-Frozen and ECGFounder-FT for every output in both datasets (Supplementary Tables S3 and S4).

The fusion assigned the generalist weights of 0.2 to 0.6 across MIMIC-IV-Ext-SHD outputs and 0.1 to 0.6 across EchoNext-Mini Dataset outputs, so no output was assigned to a single predictor. The generalist alone was stronger than DGRNet for pericardial effusion AUROC on MIMIC-IV-Ext-SHD and for mitral regurgitation on the EchoNext-Mini Dataset, and GS-ECG exceeded both predictors for these outputs; for mitral regurgitation, its AUPRC gain over DGRNet, 0.0170, was the largest on the EchoNext-Mini Dataset.

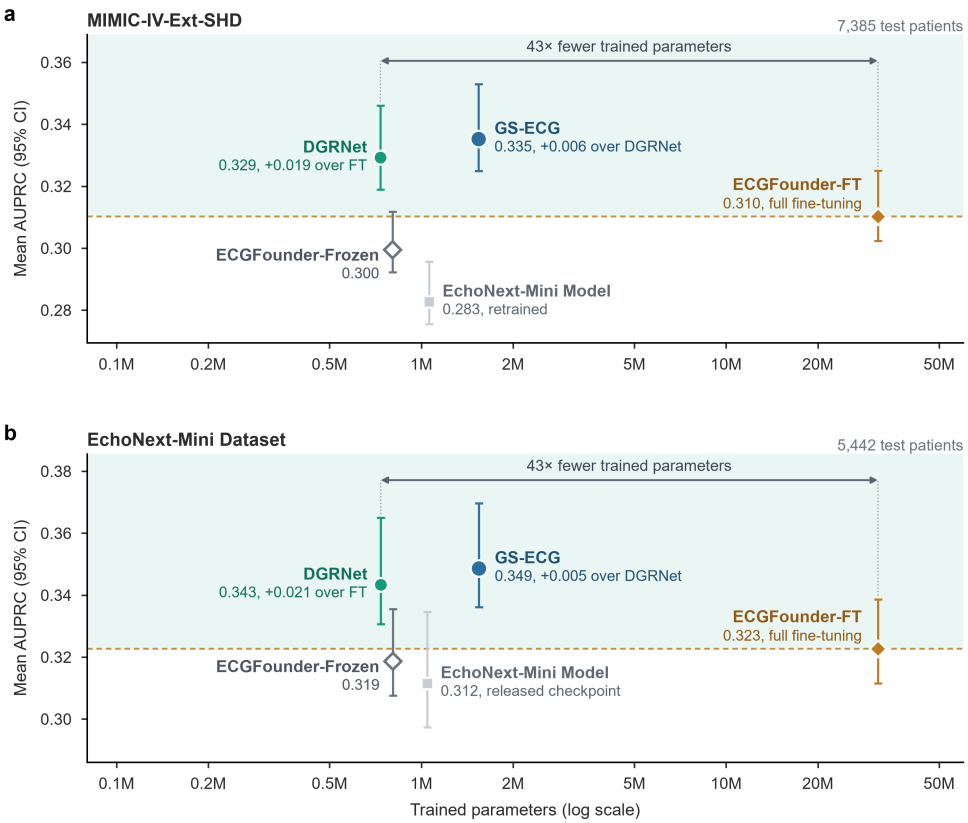


Fig. 2 Mean AUPRC against trained parameters on the MIMIC-IV-Ext-SHD test set, panel a, and the EchoNext-Mini Dataset test set, panel b. Points show mean AUPRC across all outputs and error bars 95% CIs from 1,000 patient-level bootstrap resamples; the horizontal axis gives the number of parameters each model trained, on a log scale. The dashed line marks the mean AUPRC of full fine-tuning, ECGFounder-FT, shading marks values above it, and the span gives the ratio of parameters trained by ECGFounder-FT and DGRNet. Gains beside DGRNet and GS-ECG are paired differences on the same resamples. GS-ECG and ECGFounder-Frozen also run the frozen 30.7-million-parameter backbone at inference. The EchoNext-Mini Model was retrained on MIMIC-IV-Ext-SHD and is the authors' released checkpoint on the EchoNext-Mini Dataset.

2.3 GS-ECG outperformed the published EchoNext-Mini Model

On the EchoNext-Mini Dataset, GS-ECG found more patients with aortic stenosis, mitral regurgitation, and tricuspid regurgitation than the authors' released EchoNext-Mini Model at every referral capacity from 1% to 20% (Figure 4). These were the valvular abnormalities with at least 200 positive test patients in both datasets, with prevalences of 5.3%, 6.2%, and 6.5% on the EchoNext-Mini Dataset, and moderate or greater disease of these valves can prompt intervention [6]. At 5% capacity, or 273 referrals, GS-ECG found 98, 104, and 116 affected patients against 84, 83, and 98, or

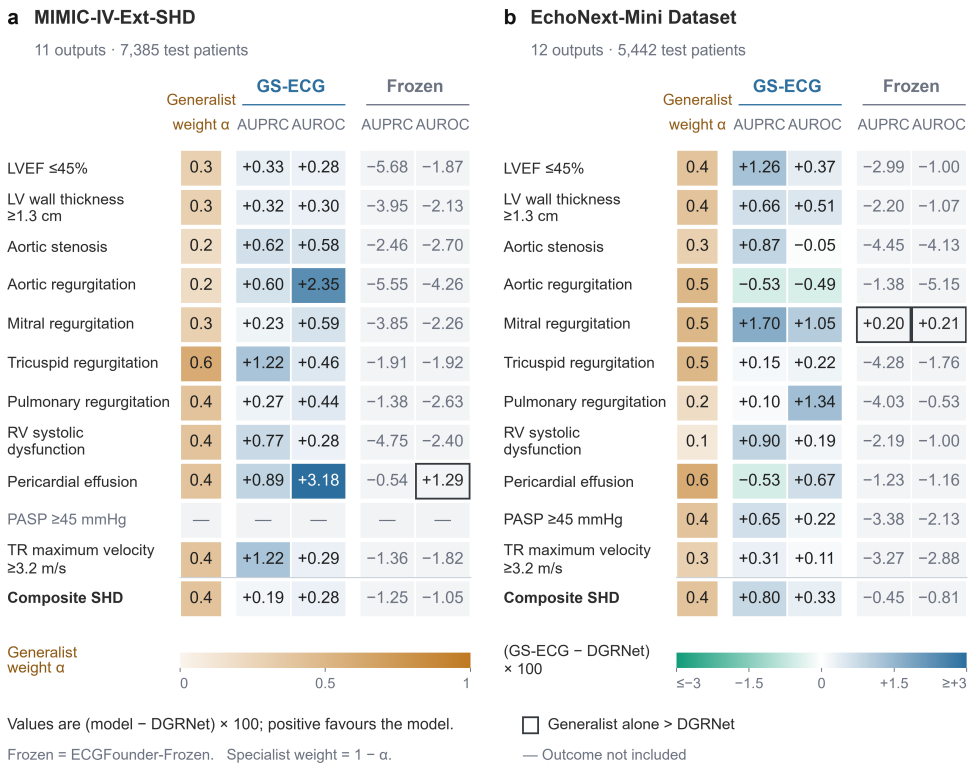


Fig. 3 Label-wise fusion weights and differences from DGRNet on MIMIC-IV-Ext-SHD, panel a, and the EchoNext-Mini Dataset, panel b. Columns give the weight assigned to the generalist, then the AUPRC and AUROC of GS-ECG and of ECGFounder-Frozen minus those of DGRNet, multiplied by 100; only the GS-ECG differences are coloured. Values are point estimates. Outlined cells mark outputs for which the generalist alone exceeded DGRNet.

17% to 25% more. At 20% capacity, it captured 66.5% to 74.8% of affected patients, against 61.1% to 71.0% for the EchoNext-Mini Model and 20% expected from random referral.

For composite SHD, both models were near ceiling among the highest-scoring patients: at 1% capacity, all 55 patients referred by the EchoNext-Mini Model and 54 of those referred by GS-ECG had SHD. The advantage of GS-ECG emerged as capacity widened, to 13 additional patients at 10% and 34 at 20% capacity, and was maintained up to 50% capacity, where it found 37 more.

Across all outputs, GS-ECG had 11.9% higher mean AUPRC than the released model, 0.3487 against 0.3115, with paired 95% CIs excluding zero for all four overall metrics. It had higher AUPRC for 10 of 12 outputs and higher AUROC for 11 (Supplementary Table S4); the released model kept a higher AUPRC for pulmonary regurgitation and pericardial effusion, with 20 and 69 positive test patients, and a higher AUROC for pericardial effusion.

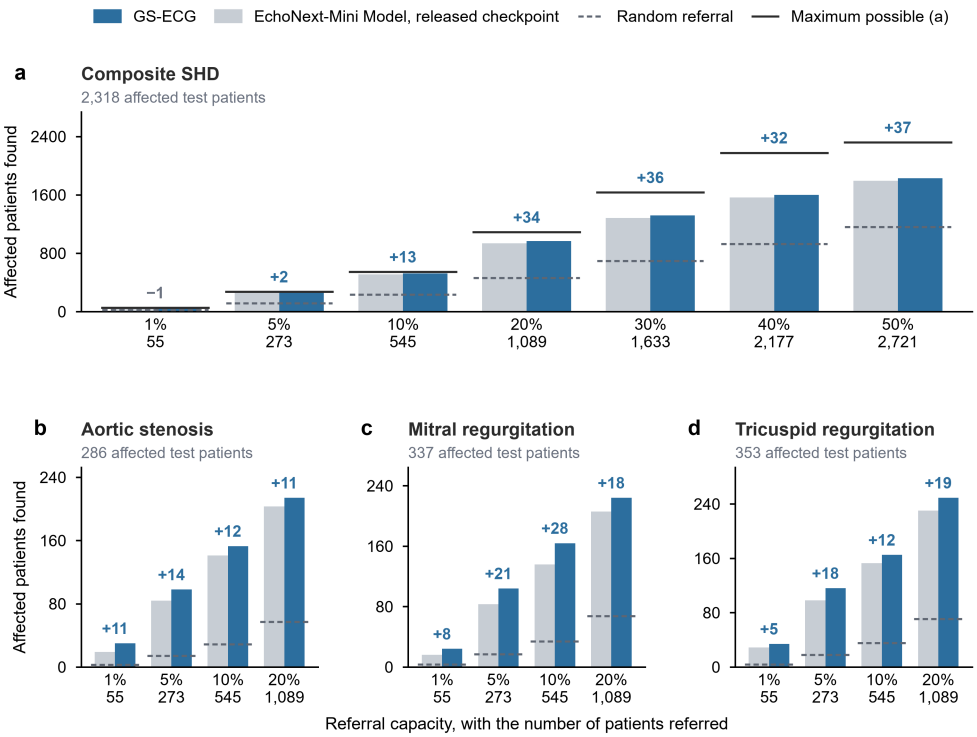


Fig. 4 Patients found by GS-ECG and the authors' released EchoNext-Mini Model at fixed referral capacities on the EchoNext-Mini Dataset. Panel a shows composite SHD at seven capacities from 1% to 50%, and panels b–d show aortic stenosis, mitral regurgitation, and tricuspid regurgitation at four capacities from 1% to 20%. At each capacity, each model referred that percentage of the 5,442 test patients, rounded up and ranked separately for each output; the number referred is given below each capacity, and annotations give GS-ECG minus the EchoNext-Mini Model. Dashed lines give the number expected under random referral, and solid lines in panel a the most patients that could be found at each capacity. For the three valvular abnormalities, GS-ECG also found more affected patients at every intermediate capacity in one-percentage-point steps. Y-axes are scaled separately.

Results on MIMIC-IV-Ext-SHD, where the EchoNext-Mini Model was retrained on the training set with the same checkpoint-selection criterion, were consistent. GS-ECG had 18.6% higher mean AUPRC, 0.3353 against 0.2827, with paired 95% CIs excluding zero for all four overall metrics, and was higher for all 11 outputs on both metrics (Supplementary Table S5). The largest AUPRC gains were 0.1238 for left ventricular ejection fraction (LVEF) $\leq$45%, 0.1002 for left ventricular (LV) wall thickness $\geq$1.3 cm, and 0.0622 for right ventricular (RV) systolic dysfunction. Supplementary Figure S1 shows the label-wise receiver operating characteristic and precision–recall curves of GS-ECG. GS-ECG also found more patients with each of the three valvular abnormalities in all 60 label–capacity comparisons from 1% to 20% (Supplementary Figure S2), capturing 68.4% to 69.6% of them at 20% capacity against 60.5% to 63.9% for the retrained model.

2.4 Ranking by GS-ECG could guide echocardiography referral

GS-ECG scores also stratified patients within an echocardiography referral queue. Two cutoffs selected on validation data divided the MIMIC-IV-Ext-SHD test patients into three score groups with composite SHD prevalences of 90.9%, 50.3%, and 7.1%, against 56.0% in the test set (Figure 5 and Supplementary Table S6). The highest-scoring 23.3% contained 1,567 of the 4,136 affected patients, so referring this group first would require 1.10 TTEs per affected patient found, compared with 1.79 under random referral. The lowest-scoring 8.7% contained 46 affected patients, so referring all other patients would have captured 98.9% of all affected patients, and the remaining 67.9% formed an intermediate group.

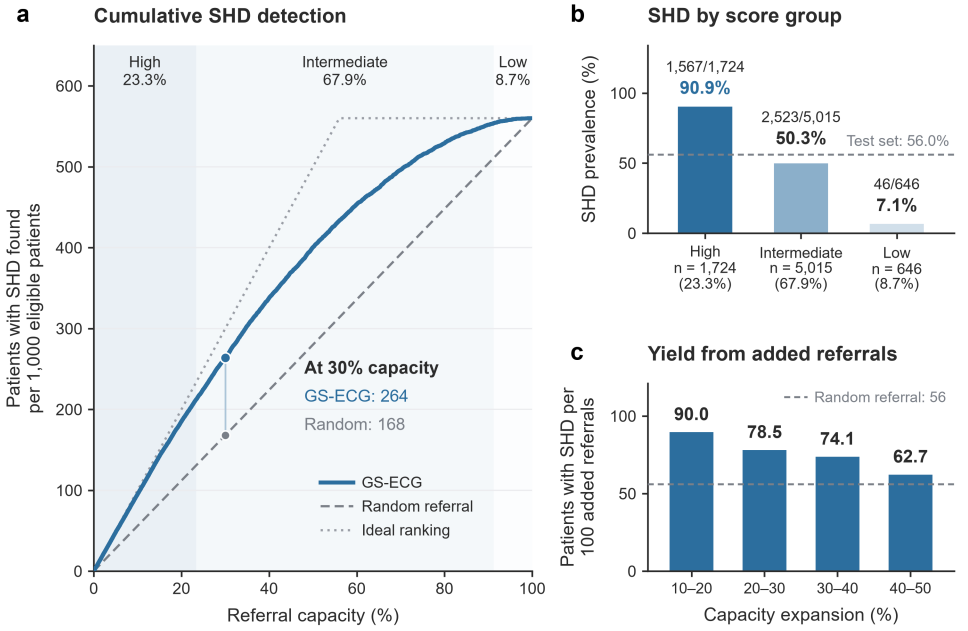


Fig. 5 Referral by GS-ECG score on the MIMIC-IV-Ext-SHD test set. Panel a shows cumulative patients with composite SHD found per 1,000 eligible patients as referral capacity increases, for ranking by GS-ECG, expected random referral, and ideal ranking. Shading marks the high, intermediate, and low score groups defined by two cutoffs selected on validation data. Panel b shows composite SHD prevalence within each group; annotations give affected/total patients and each group's share of the test cohort. Panel c shows patients with composite SHD found per 100 additional referrals for successive 10-percentage-point capacity expansions, using the actual integer numbers of additional referrals. Dashed reference lines in panels b and c show the test-set prevalence and expected random-referral yield, respectively. Values use all 7,385 test patients, of whom 4,136 had composite SHD.

Applied unchanged, the cutoffs met their 90% targets on the test set. The negative predictive value (NPV), the proportion of patients below the lower cutoff without composite SHD, reached 92.9%, and the positive predictive value (PPV), the proportion of patients at or above the upper cutoff with it, reached 90.9%, with its 95% CI extending to 89.5%; a third cutoff targeting 90% sensitivity, the proportion of affected patients referred, reached 91.3%.

When TTE capacity rather than a cutoff determined referral, ranking by GS-ECG concentrated composite SHD among the earliest referrals. The highest-scoring 10% had a PPV of 95.3%, and the highest-scoring 30% included 264 affected patients per 1,000 eligible patients, against 168 expected from random referral and 300 under ideal ranking, a PPV of 87.9% and 47.1% of all affected patients. Successive 10-percentage-point expansions from 10% to 50% capacity identified 90.0, 78.5, 74.1, and 62.7 affected patients per 100 additional referrals, against 56 under random referral.

3 Discussion

This study asked how an ECG foundation model should be used when the clinical target is defined by echocardiography. In two health systems, a compact specialist trained from scratch outperformed full fine-tuning of the foundation model, fine-tuning improved only slightly on keeping the model frozen, and the frozen model added most when it was combined with the specialist. DGRNet, with fewer than one million trained parameters, had higher mean AUPRC than ECGFounder-FT, which trained 31.5 million, and GS-ECG, which added the frozen generalist, ranked first on all four overall metrics in both datasets while training 4.9% as many parameters as full fine-tuning. GS-ECG also outperformed the published EchoNext-Mini Model, including the authors' released checkpoint on its own dataset.

Foundation models are built for breadth. ECGFounder was pretrained on more than 10 million ECGs to recognize a wide range of diagnostic statements, and its backbone of about 30.7 million parameters encodes general-purpose ECG representations [17, 18]. Adapting all of these parameters to a single multi-label task yielded limited returns: full fine-tuning increased mean AUPRC over the frozen model by 0.0107 and 0.0040, less than DGRNet's margin over full fine-tuning. DGRNet was instead designed for the task. Its stem operates at 50 Hz, and four dilated, depthwise gated blocks extend each feature to about two cardiac cycles without further downsampling, so that beat-level morphology is modelled with fewer than one million parameters. This design outperformed full fine-tuning and every other model trained from scratch in both datasets, so its advantage reflects fit to the task rather than training from scratch alone. At the scale of TTE-labelled data available here, a task-matched design contributed more than re-tuning a general model.

Adding the frozen generalist improved the specialist in both datasets. The two predictors received the same inputs and used task modules of similar form, except that the specialist's joint-context token also received a summary of its own features, so their differences arise mainly from the waveform encoders and their training: pretraining on more than 10 million ECGs versus training from scratch on TTE-derived labels; within-record normalization versus population-standardized amplitudes; 20

long-context tokens versus 125 features each spanning about two beats; and binary cross-entropy (BCE) with a ranking term versus BCE with an exponential moving average of weights. Predictors built on such different representations are expected to make partly different errors, and averaging them can improve accuracy even when one is weaker on average. Consistent with this, the fusion assigned weight to the generalist for every output, and GS-ECG exceeded both predictors for the outputs on which the generalist alone was stronger. These results suggest a division of labour: a compact specialist provides task-specific capacity, whereas the foundation model stays frozen and serves as a complementary predictor, so that only the task modules, the specialist, and the fusion parameters are trained, while the frozen backbone runs at inference. Combining a pretrained model with a task-adapted model has likewise improved accuracy and robustness in general-domain transfer learning [20].

GS-ECG also outperformed a model built specifically for SHD on its own data. Against the authors' released EchoNext-Mini checkpoint, it found more patients with aortic stenosis, mitral regurgitation, and tricuspid regurgitation at every referral capacity from 1% to 20%, 17% to 25% more among the highest-scoring 5% of patients, and the result held on MIMIC-IV-Ext-SHD with the model retrained. These valvular abnormalities are uncommon but clinically actionable, and their detection from the ECG is an established target [10]. For composite SHD, both models referred almost only affected patients at the smallest capacities, and the advantage of GS-ECG appeared among patients ranked lower, where the choice of whom to scan is less clear.

MIMIC-IV-Ext-SHD makes such comparisons repeatable on public data, for model development as well as evaluation. It fixes the design choices that shape reported performance, as PTB-XL did for ECG classification [21]: a 365-day linkage window anchored at the ECG, ECG-level aggregation of 10 abnormalities and composite SHD, and patient-level separation of the data splits. Its component-label prevalences in the test set, from 0.4% to 28.7%, preserve the imbalance that a composite endpoint can hide, so new models can be compared on common and rare abnormalities under the same protocol.

Because MIMIC-IV-Ext-SHD comprised patients with clinically indicated TTE, a potential application is to order an existing referral queue. Two validation-selected cutoffs separated a highest-scoring group in which 90.9% had composite SHD from a lowest-scoring group in which 7.1% did. These groups map onto a pathway that a prospective study could test: earlier TTE for the highest-scoring group, prioritization through usual clinical assessment for the intermediate group, and later TTE after clinical review for the lowest-scoring group. The lowest-scoring group contained 46 of the 4,136 test patients with composite SHD, and the safety of delayed imaging in this group remains to be established. Under a fixed capacity of 30%, ranking by GS-ECG referred 47.1% of affected patients, against at most 53.6% under ideal ranking. Prospective studies have used AI-ECG to increase diagnoses of low ejection fraction and cardiomyopathy [12, 14] and to screen for heart failure and SHD at the point of care [13, 26]; the referral yields estimated here could inform the design of such an evaluation.

Both datasets are retrospective and each comes from a single health system. The EchoNext-Mini Dataset is large and released with its own patient-level splits, which

allowed the full training and evaluation protocol to be repeated in a second health system; testing locked models across health systems is a next step. MIMIC-IV-Ext-SHD comprises patients referred for TTE, whose test ECGs preceded the linked TTE by a median of 1.55 days, so predictive values and referral yields reflect this case mix, and cutoffs would need to be selected locally in other populations, including screening settings.

In summary, when the clinical target is defined by echocardiography, a compact specialist outperformed full fine-tuning of an ECG foundation model, and the foundation model added most as a frozen complementary predictor: GS-ECG trained 4.9% as many parameters as full fine-tuning and outperformed the published EchoNext-Mini Model on its own dataset. On MIMIC-IV-Ext-SHD, a patient-separated benchmark built from public data, ranking patients by GS-ECG concentrated TTE referrals on those with SHD: 90.9% of the highest-scoring 23.3% had composite SHD, against 7.1% of the lowest-scoring 8.7%.

4 Methods

4.1 Data sources and ethics

MIMIC-IV-Ext-SHD was built from three de-identified PhysioNet resources from Beth Israel Deaconess Medical Center (Boston, MA, USA) [27]. Medical Information Mart for Intensive Care IV (MIMIC-IV) provided demographic data [25]. MIMIC-IV-ECG provided diagnostic 12-lead ECG waveforms and machine measurements acquired in emergency, inpatient, and outpatient settings between 2008 and 2019, and MIMIC-IV-ECHO provided structured measurements from TTEs performed between 2008 and 2022 [28–30]. Records were linked by patient identifier. Model inputs were restricted to information available at ECG acquisition, and outcome labels were derived from subsequent TTEs. The benchmark name follows the PhysioNet convention for resources derived from MIMIC.

The creation and sharing of MIMIC-IV were reviewed by the Institutional Review Board of Beth Israel Deaconess Medical Center, which waived the requirement for patient consent, and the MIMIC-IV-ECG and MIMIC-IV-ECHO modules were approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology with a waiver of individual patient consent, because the projects did not affect clinical care and all protected health information was de-identified. The EchoNext-Mini Dataset derives from retrospective data analysed under institutional review board-approved protocols at NewYork-Presbyterian Hospital/Columbia University Irving Medical Center that permit sharing of de-identified data for research. All datasets were obtained from PhysioNet under their respective licences and data use agreements, with credentialed access and the required training for MIMIC-IV and MIMIC-IV-ECHO.

4.2 ECG eligibility and structured variables

We retained diagnostic ECGs from patients aged 18 years or older with a valid acquisition time, demographic information, automated measurements, and a standard 10-second, 12-lead waveform sampled at 500 Hz. Before ECG–TTE pairing, we excluded records with machine measurements outside prespecified plausibility ranges or with machine-reported ventricular pacing or pacemaker rhythm, as in EchoNext [2], because ventricular pacing replaces intrinsic ventricular activation (Supplementary Table S7).

Six structured variables served as model inputs: sex, age at ECG, ventricular rate, PR interval, QRS duration, and Bazett-corrected QT interval (QTc). Age at ECG was derived from the ECG year and the MIMIC-IV anchor age and year. Ventricular rate was calculated from the RR interval, the PR interval from P-wave onset to QRS onset, QRS duration from QRS onset to QRS end, and the QT interval from QRS onset to T-wave end before Bazett correction. PR intervals with invalid timing or values outside 40–400 ms were set to missing. Race and ethnicity were reserved for cohort description, with unmapped entries assigned to an unknown category.

4.3 Echocardiographic outcome definitions

Outcomes were derived from structured TTE measurements and severity entries rather than diagnosis codes. To focus on native structural disease, we excluded TTEs indicating valve replacement or repair, annuloplasty, transcatheter aortic valve replacement, a valve clip, or a mechanical or bioprosthetic valve. Each eligible native-valve TTE was assigned 10 binary component labels adapted from EchoNext [2]. Two left ventricular labels captured LVEF $\leq$45% and LV wall thickness $\geq$1.3 cm. Five valvular labels captured moderate-or-greater aortic stenosis and moderate-or-greater regurgitation of the aortic, mitral, tricuspid, and pulmonary valves. The remaining three labels captured RV systolic dysfunction recorded as moderate, severe, or depressed; moderate or large pericardial effusion; and tricuspid regurgitation (TR) maximum velocity $\geq$3.2 m/s. EchoNext also used pulmonary artery systolic pressure (PASP) $\geq$45 mmHg, which MIMIC-IV-ECHO does not record. PASP is estimated from the TR jet, and a TR maximum velocity of 3.2 m/s corresponds to a gradient of about 41 mmHg and, with a right atrial pressure of 5 mmHg, to a PASP of about 46 mmHg, close to the EchoNext threshold [31].

The thresholds, the moderate-or-greater rule with mild-to-moderate grades counted as negative, and the exclusion of repaired or replaced valves follow EchoNext, which derived its definitions from echocardiography guidelines and chose cutoffs that correspond to clinically accepted moderate or greater disease and to left ventricular dysfunction eligible for guideline-directed therapy [2]. Adaptations specific to MIMIC-IV-Ext-SHD were the omission of the PASP label, the mapping of each label to MIMIC-IV-ECHO structured fields, use of a TR gradient of at least 40.96 mmHg when velocity was unrecorded, and retention of the dataset's significant pulmonary regurgitation grade. Labels followed a positive-evidence rule: a component label was positive when at least one structured measurement or severity entry met its criterion

and was otherwise set to zero (Supplementary Table S8). Composite SHD was positive when any component label was positive, providing an any-abnormality endpoint alongside the abnormality-specific component labels.

4.4 Cohort construction and data splits

Each ECG was paired with every eligible TTE of the same patient performed at or after the ECG acquisition time and within 365 days, following the EchoNext study [2, 16], and ECGs with at least one such TTE entered the cohort. All qualifying pairs were retained, so one ECG could be paired with more than one TTE. Labels were then aggregated to the ECG level: each component label was positive if any paired TTE documented the abnormality, and composite SHD was recomputed from the aggregated labels. This follows EchoNext, in which an ECG was positive if it preceded an echocardiogram with SHD by up to one year [2], so the task is to estimate TTE findings over the year after the ECG. For describing ECG–TTE intervals at the ECG level, one representative TTE was selected, prioritizing composite-positive status and then the shortest interval.

Patients were assigned to mutually exclusive training, validation, and test sets in approximate proportions of 70%, 15%, and 15%. Candidate assignments sampled validation and test patients, stratified by the composite SHD label of their latest eligible ECG, over a grid of target prevalences, and the assignment that minimized the spread of composite SHD prevalence across the three sets, computed on the records each set contributes, was retained and fixed before model training. Training patients contributed all eligible ECGs, so that every labelled recording was used for model development, whereas validation and test patients each contributed their latest eligible ECG, so that models were selected and tested in the same setting. Patient-level separation and one ECG per validation and test patient follow the design of the EchoNext-Mini Dataset [16]; the split-selection procedure and the plausibility checks in Supplementary Table S7 were introduced for MIMIC-IV-Ext-SHD.

Of 800,035 ECGs from 161,352 patients, 774,601 ECGs from 158,872 patients met the eligibility criteria; ventricular pacing accounted for most of the 25,434 exclusions. Of 179,928 TTEs from 85,292 patients, 8,696 were excluded because of a prosthetic valve or previous valve intervention, leaving 171,232 native-valve TTEs from 84,219 patients. Pairing yielded 398,802 ECG–TTE pairs from 49,229 patients, comprising 254,966 unique ECGs and 91,863 unique TTEs (Figure 6). ECG-level aggregation followed by split-specific record selection produced 188,273 model-ready records: 173,504 training ECGs from 34,460 patients and one ECG each from 7,384 validation and 7,385 test patients. Table 1 summarizes all paired ECGs and the three sets: median age at ECG was 68 to 69 years, and 52.1% to 53.6% of records were from male patients. Composite SHD prevalence was balanced across the three sets, whereas several component labels were less prevalent in the validation and test sets, which contain one ECG per patient, than in the training set, which retains repeated ECGs; for example, LVEF $\leq$45% was present in 21.2% and 20.8% of validation and test patients compared with 26.4% of training ECGs. The median ECG–TTE interval was 43.8 days across all pairs, and the median interval between each validation or test patient's ECG and its representative TTE was 1.46 and 1.55 days, respectively.

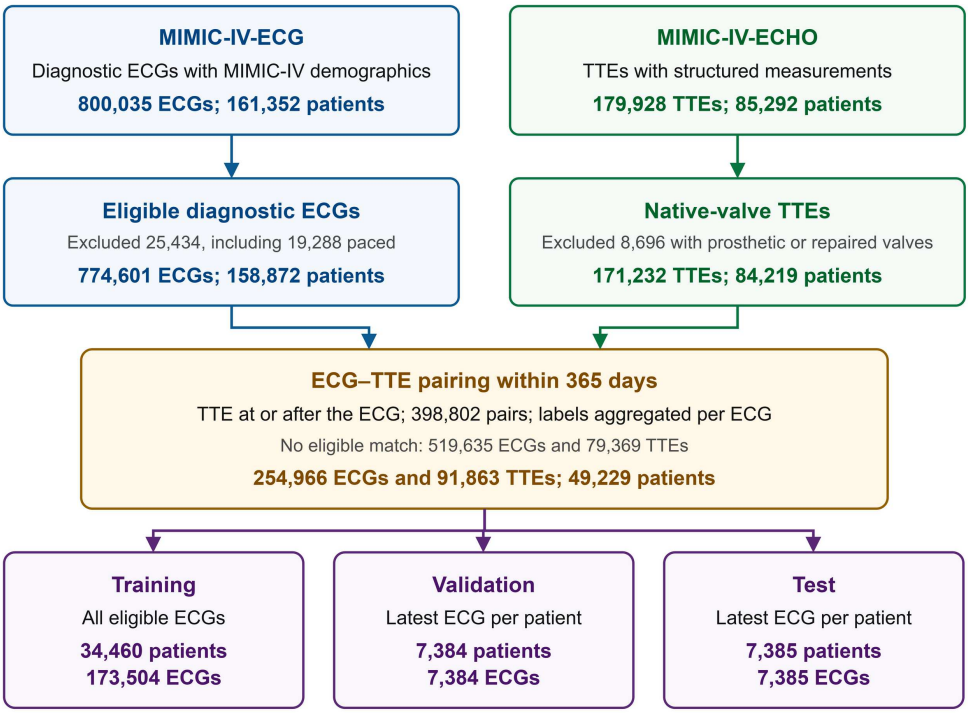


Fig. 6 MIMIC-IV-Ext-SHD cohort flow. ECG and TTE records were filtered separately, paired within patients over a 365-day window, aggregated into ECG-level labels, and assigned by patient to training, validation, and test sets. Bold numbers give the records and patients remaining after each step. ECG exclusions were counted in the order applied, and Supplementary Table S7 lists each criterion. Validation and test patients contributed their latest ECG only, so 66,693 earlier ECGs of these patients were not used. ECG, electrocardiogram; TTE, transthoracic echocardiography.

For GS-ECG development, training patients were further divided into a predictor-training subset, $\mathcal{D}_{\text{expert}}$, with 85% of training patients, and a fusion-training subset, $\mathcal{D}_{\text{fusion}}$, with the remaining 15%. The division used multi-label stratification of patient-level vectors indicating whether any of a patient's training ECGs was positive for each label. $\mathcal{D}_{\text{expert}}$ contained 29,291 patients and 148,284 ECGs, and $\mathcal{D}_{\text{fusion}}$ contained 5,169 patients and 25,220 ECGs. Reference models were trained on the full training set.

4.5 Input preprocessing

Each waveform was arranged in the lead order I, II, III, aVR, aVL, aVF, and V1–V6, giving a $12 \times 5{,}000$ array. For the ECGFounder-based predictors and DGRNet, waveforms were standardized lead by lead with means and standard deviations estimated from a sample of $\mathcal{D}_{\text{expert}}$ ECGs, and missing structured variables were imputed with medians before standardization. These parameters were estimated from $\mathcal{D}_{\text{expert}}$ and applied unchanged to $\mathcal{D}_{\text{fusion}}$, validation, and test data. Reference models used the

same imputation and standardization with parameters estimated from the full training set, after scaling waveform amplitudes by 1,000. No model trained in this study used stochastic waveform augmentation.

4.6 EchoNext-Mini Dataset

The EchoNext-Mini Dataset is a public, de-identified subset of the EchoNext cohort from NewYork-Presbyterian/Columbia University Irving Medical Center, released with the EchoNext-Mini Model [16] through PhysioNet [32], and is independent of the MIMIC-IV sources used for MIMIC-IV-Ext-SHD. The dataset was used as released, including its ECG waveforms, structured metadata, outcome labels, and patient-level training, validation, and test splits. The 82,543 ECGs assigned to the released splits comprised 26,218 training patients with 72,475 ECGs, 4,626 validation patients, and 5,442 test patients, with one ECG per validation and test patient. The released labels comprise 11 component labels, including PASP $\geq$45 mmHg, and composite SHD. Seven structured variables were taken as raw values from the released metadata: sex, age at ECG, ventricular rate, atrial rate, PR interval, QRS duration, and corrected QT interval.

Processing added three steps. The released 2,500-sample waveforms, recorded at 250 Hz, were resampled to 5,000 samples by linear interpolation, retaining the upstream waveform filtering and dataset-level normalization. Training patients were divided 85:15 by patient-level multi-label stratification, as for MIMIC-IV-Ext-SHD, into $\mathcal{D}_{\text{expert}}$ with 22,285 patients and 61,431 ECGs and $\mathcal{D}_{\text{fusion}}$ with 3,933 patients and 11,044 ECGs. For the ECGFounder-based predictors and DGRNet, missing structured variables were imputed with medians and standardized, and per-lead waveform statistics were taken from a fixed sample of 10,000 ECGs, all estimated from $\mathcal{D}_{\text{expert}}$.

4.7 GS-ECG architecture

GS-ECG is designed to combine two complementary ECG representations. The generalist reuses representations learned by ECGFounder during large-scale pretraining and keeps its backbone frozen. The specialist is trained from random initialization on the TTE-derived labels and learns task-specific waveform features with fewer than one million parameters. Each predictor integrates ECG features with the structured variables and outputs label probabilities, which are calibrated and then combined through label-wise weights (Figure 7).

Input representation

The standardized ECG, structured variables, and binary targets were represented as

$$X \in \mathbb{R}^{12\times 5{,}000}, \qquad x_{\text{tab}} \in \mathbb{R}^{d_{\text{tab}}}, \qquad y \in \{0,1\}^{M}, \tag{1}$$

with $d_{\text{tab}} = 6$ and $M = 11$ for MIMIC-IV-Ext-SHD, and $d_{\text{tab}} = 7$ and $M = 12$ for the EchoNext-Mini Dataset. Targets comprised the dataset-specific TTE-derived component labels and composite SHD. ECGFounder-based predictors further normalized each ECG within record, with the mean $\mu(X)$ and standard deviation $s(X)$ calculated

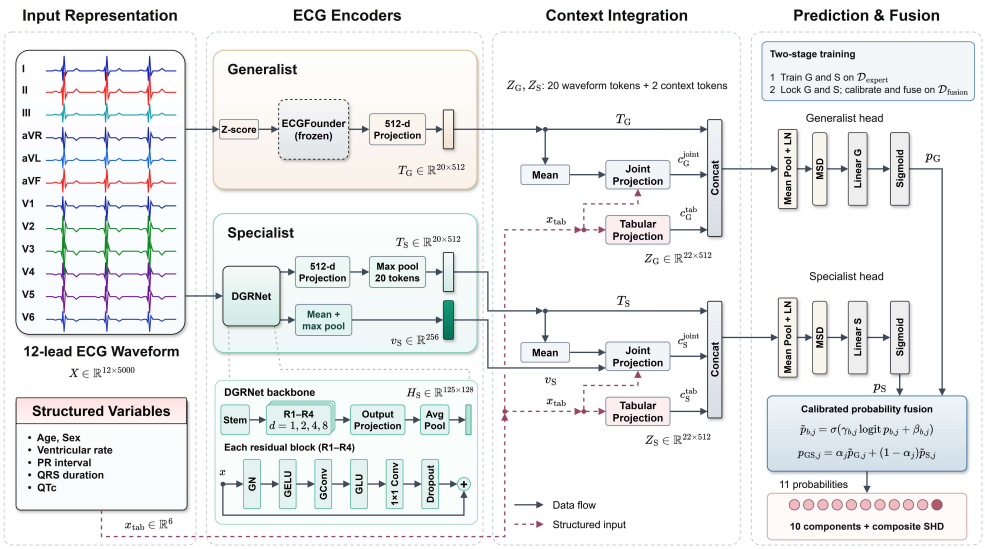


Fig. 7 GS-ECG architecture. A frozen ECGFounder generalist and a randomly initialized DGRNet specialist each map the 12-lead ECG and structured variables to label probabilities. Calibration slopes and intercepts $(\gamma_{b,j}, \beta_{b,j})$ for predictor $b \in \{\text{G}, \text{S}\}$ and label j, and label-wise fusion weights α_j, are fitted on the held-out fusion-training subset, $\mathcal{D}_{\text{fusion}}$, with both predictors fixed. The fusion panel shows the label-wise calibration and fusion rules (Equations 11 and 12). The inset expands the DGRNet backbone; dimensions are shown for MIMIC-IV-Ext-SHD. G, generalist; GConv, grouped dilated convolution; GELU, Gaussian error linear unit; GLU, gated linear unit; GN, group normalization; LN, layer normalization; MSD, multi-sample dropout; S, specialist; d, dilation rate.

across all leads and time points, whereas DGRNet received X directly:

$$\widetilde{X} = \frac{X - \mu(X)}{\max\{s(X), \varepsilon\}}, \qquad \varepsilon = 10^{-6}. \tag{2}$$

Throughout, $\text{LN}(\cdot)$ denotes layer normalization [33] and $\text{Lin}(\cdot)$ a fully connected linear layer; each instance has its own parameters.

ECGFounder generalist predictor

The generalist used the pretrained ECGFounder backbone $\mathcal{E}$ for waveform feature extraction [17]; used alone, it is referred to as ECGFounder-Frozen. Its original disease-classification layer was removed, and all backbone parameters and buffers remained frozen during SHD training. The backbone output a sequence of 20 feature vectors, $\mathcal{E}(\widetilde{X}) \in \mathbb{R}^{20\times1024}$, and a trainable projection $\Pi_{\text{G}}(\cdot) = \text{Lin}(\text{LN}(\cdot))$ mapped each vector to 512 dimensions. A second LN and dropout produced the generalist waveform tokens:

$$T_{\text{G}} = \text{Dropout}\Big(\text{LN}\Big(\Pi_{\text{G}}\Big(\mathcal{E}(\widetilde{X})\Big)\Big)\Big) \in \mathbb{R}^{K_{\text{G}}\times512}, \qquad K_{\text{G}} = 20. \tag{3}$$

DGRNet specialist predictor

The specialist was built on DGRNet, a dilated gated residual backbone formed by a convolutional stem, four residual blocks, an output projection, and adaptive average pooling (Figure 7, inset), followed by its own task modules, described below. Used alone, the specialist is referred to as DGRNet. The stem processed all 12 leads jointly with a convolution of kernel size 31 and stride 10, producing 192 channels at 50 Hz. Four residual blocks with dilation rates of 1, 2, 4, and 8 then enlarged the temporal receptive field without further downsampling [34, 35]. Each block applied group normalization (GN) with eight groups [36], a Gaussian error linear unit (GELU) activation [37], a grouped dilated convolution (GConv), a gated linear unit (GLU) [38], a 1×1 convolution, and dropout with probability 0.3, and a residual connection added the result to the block input. The GConv used kernel size 7 and one group per input channel, making it depthwise; it expanded the 192 channels to 384, which the GLU gated back to 192 before the channel-preserving 1×1 convolution. A final 1×1 output projection reduced the 192 channels to 128.

The backbone produced 125 temporal feature vectors after adaptive average pooling, and their temporal mean and maximum were concatenated to form a separate summary:

$$\begin{aligned} F^{(0)} &= \mathrm{Conv}_{\mathrm{stem}}(X) \in \mathbb{R}^{192\times 500}, \\ F^{(r)} &= F^{(r-1)} + \mathcal{B}_r\left(F^{(r-1)}\right), \qquad r = 1, \ldots, 4, \\ H_{\mathrm{S}} &= \mathrm{AAP}_{125}\left(\mathrm{Conv}_{\mathrm{out}}(F^{(4)})\right)^{\top} \in \mathbb{R}^{125\times 128}, \\ v_{\mathrm{S}} &= [\mathrm{mean}_t(H_{\mathrm{S}}); \max_t(H_{\mathrm{S}})] \in \mathbb{R}^{256}, \end{aligned} \tag{4}$$

where $\mathcal{B}_r$ is the rth residual block with dilation rate $d_r \in \{1, 2, 4, 8\}$, $\mathcal{B}_r = \mathrm{Dropout} \circ \mathrm{Conv}_{1\times 1} \circ \mathrm{GLU} \circ \mathrm{GConv}_{d_r} \circ \mathrm{GELU} \circ \mathrm{GN}$, and $\circ$ denotes function composition applied from right to left. Here, $\mathrm{Conv}_{\mathrm{out}}$ denotes the output projection and AAP_{125} adaptive average pooling to 125 time steps. The GLU splits its 384-channel input into halves A and B and returns $A \odot \sigma(B)$, where $\odot$ denotes element-wise multiplication and σ the logistic sigmoid. With these dilation rates, each of the 125 feature vectors in H_{S} spans 961 samples, about 1.9 s of ECG or roughly two cardiac cycles at typical heart rates.

These feature vectors were projected to 512 dimensions, adaptively max-pooled to 20 tokens to match the generalist, then layer-normalized and regularized with dropout:

$$T_{\mathrm{S}} = \mathrm{Dropout}(\mathrm{LN}(\mathrm{AMP}_{20}(\Pi_{\mathrm{S}}(H_{\mathrm{S}})))) \in \mathbb{R}^{K_{\mathrm{S}}\times 512}, \qquad K_{\mathrm{S}} = 20. \tag{5}$$

Here, $\Pi_{\mathrm{S}}(\cdot) = \mathrm{Lin}(\mathrm{LN}(\cdot))$ is an independent projection, and AMP_{20} denotes adaptive maximum pooling along time.

Context integration and multi-label prediction

Each predictor $b \in \{\mathrm{G}, \mathrm{S}\}$ had its own task modules: the waveform projection Π_b with the LN that follows it, the joint-context and tabular projections, and the classification head. The two projections produce context tokens that complement the waveform tokens: the joint-context token summarizes the waveform together with the structured

variables, whereas the tabular token represents the structured variables alone. Let $\bar{t}_b$ denote the temporal mean of T_b, with semicolons denoting concatenation. The joint-context inputs were

$$u_{\mathrm{G}} = [\bar{t}_{\mathrm{G}}; x_{\mathrm{tab}}], \qquad u_{\mathrm{S}} = [\bar{t}_{\mathrm{S}}; v_{\mathrm{S}}; x_{\mathrm{tab}}]. \tag{6}$$

Both projections took the form $\Phi(\cdot) = \mathrm{Dropout}(\mathrm{GELU}(\mathrm{Lin}(\mathrm{LN}(\cdot))))$, with 512 output dimensions:

$$c_b^{\mathrm{joint}} = \Phi_b^{\mathrm{joint}}(u_b), \qquad c_b^{\mathrm{tab}} = \Phi_b^{\mathrm{tab}}(x_{\mathrm{tab}}). \tag{7}$$

These two context tokens were appended to each predictor's waveform tokens to form

$$Z_b = [T_b; c_b^{\mathrm{joint}}; c_b^{\mathrm{tab}}] \in \mathbb{R}^{(K_b+2)\times 512}. \tag{8}$$

The classification head comprised mean pooling over Z_b and an LN, which produced the 512-dimensional representation h_b, followed by multi-sample dropout (MSD) [39], a linear classifier, and a sigmoid:

$$h_b = \mathrm{LN}\left(\frac{1}{K_b+2}\sum_{i=1}^{K_b+2} Z_{b,i}\right) \in \mathbb{R}^{512}. \tag{9}$$

During predictor training, MSD passed three independently masked versions of h_b through the linear classifier, and their logits were averaged before sigmoid transformation:

$$\begin{aligned} \ell_b^{(k)} &= W_b\,\mathrm{Dropout}_k(h_b) + a_b, \qquad k \in \{1,2,3\}, \\ \bar{\ell}_b &= \frac{1}{3}\sum_{k=1}^{3} \ell_b^{(k)}, \qquad p_b = \sigma\big(\bar{\ell}_b\big) \in [0,1]^M, \end{aligned} \tag{10}$$

where $W_b \in \mathbb{R}^{M\times 512}$ and $a_b \in \mathbb{R}^M$ are the weight and bias of the linear classifier. At inference, dropout was disabled and the linear classifier was applied once. Each predictor produced M probabilities, including a separate output for composite SHD.

Label-wise probability fusion

GS-ECG combined the outputs of the two predictors after label-wise calibration, which placed their probabilities on a comparable scale before averaging. With both predictor checkpoints fixed, all calibration and fusion parameters were fitted on $\mathcal{D}_{\mathrm{fusion}}$, and each ECG was weighted by the inverse of its patient's ECG count so that each patient contributed equal total weight. For predictor $b \in \{\mathrm{G}, \mathrm{S}\}$ and label j, an affine transformation of the predicted logit gave

$$q_{b,j} = \mathrm{clip}(p_{b,j}, 10^{-6}, 1-10^{-6}), \qquad \widetilde{p}_{b,j} = \sigma(\gamma_{b,j}\,\mathrm{logit}(q_{b,j}) + \beta_{b,j}), \tag{11}$$

where $\mathrm{logit}(q) = \log(q/(1-q))$, and $\gamma_{b,j}$ and $\beta_{b,j}$ are a fitted slope and intercept. They minimized weighted BCE with an L_2 penalty of 10^{-3} toward the identity transformation, with slopes bounded to $[0.05, 8]$ and intercepts to $[-12, 12]$. The final probability

was

$$p_{\mathrm{GS},j} = \alpha_j \widetilde{p}_{\mathrm{G},j} + (1-\alpha_j)\widetilde{p}_{\mathrm{S},j}, \qquad \alpha_j \in \{0, 0.1, \ldots, 1\}, \qquad j = 1, \ldots, M. \tag{12}$$

Each fusion weight α_j was shared across patients and selected from this grid to maximize the equal-weight average of weighted AUPRC and AUROC, minus $0.01(\alpha_j - 0.5)^2$, a small penalty that favoured balanced weights when scores were close. This stage estimated 44 calibration coefficients and 11 fusion weights for MIMIC-IV-Ext-SHD, and 48 calibration coefficients and 12 fusion weights for the EchoNext-Mini Dataset.

4.8 Model training

Both GS-ECG predictors were trained with unweighted BCE in each dataset. For a mini-batch of N ECGs and M labels, let $\bar{\ell}_{nj}$ denote the logit of ECG n for label j (Equation 10) and $p_{nj} = \sigma(\bar{\ell}_{nj})$. Then

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{NM} \sum_{n=1}^{N} \sum_{j=1}^{M} \left[y_{nj} \log p_{nj} + (1 - y_{nj}) \log(1 - p_{nj}) \right]. \tag{13}$$

The generalist additionally used a pairwise ranking term over the 10 component labels of MIMIC-IV-Ext-SHD and the 11 of the EchoNext-Mini Dataset, excluding composite SHD. This encourages positive ECGs to score above negative ECGs for the same label. Let $\mathcal{P}_j$ and $\mathcal{N}_j$ denote the positive and negative ECGs for label j in the mini-batch, and let $\mathcal{J}$ denote the component labels with at least one of each. The ranking term was

$$\mathcal{L}_{\mathrm{rank}} = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \frac{\displaystyle\sum_{n \in \mathcal{P}_j} \sum_{m \in \mathcal{N}_j} \mathrm{softplus}(\bar{\ell}_{mj} - \bar{\ell}_{nj})}{|\mathcal{P}_j|\,|\mathcal{N}_j|}, \tag{14}$$

where $\mathrm{softplus}(u) = \log(1 + \exp(u))$; it was zero when $\mathcal{J}$ was empty. The generalist minimized $\mathcal{L}_{\mathrm{BCE}} + 0.1\mathcal{L}_{\mathrm{rank}}$, whereas the specialist minimized $\mathcal{L}_{\mathrm{BCE}}$.

ECGFounder-Frozen, ECGFounder-FT, and DGRNet were trained on the same $\mathcal{D}_{\mathrm{expert}}$ patients and inputs. Both ECGFounder variants were initialized from the official checkpoint pretrained on the Harvard–Emory ECG Database [17], with no MIMIC-IV-Ext-SHD records used for pretraining. ECGFounder-Frozen trained only its task modules, whereas ECGFounder-FT shared its objective and settings and additionally updated all backbone parameters, with a backbone learning rate of 1×10^{-5} selected from several candidate values on validation data. DGRNet was trained from random initialization. All predictors used AdamW [40] with batch size 128, weight decay 5×10^{-4}, dropout 0.3, and gradient clipping at norm 1, for up to 40 epochs with a minimum of 14 epochs and early-stopping patience of 10. Task modules used a learning rate of 3×10^{-4} and the DGRNet backbone 5×10^{-4}, and two warm-up epochs preceded cosine decay [41] to 1% of the initial rates. Forward passes used BF16 mixed precision, with loss and sigmoid calculations in FP32. Checkpoints maximized

the equal-weight average of validation mean AUPRC and mean AUROC, and DGRNet used an exponential moving average (EMA) of model weights with target decay 0.995 for validation and final prediction.

Hyperparameters, including learning rates, loss terms, and the fusion recipe, were tuned on MIMIC-IV-Ext-SHD validation data. On the EchoNext-Mini Dataset, all models other than the released EchoNext-Mini Model were trained on the released training split following the same procedure, and re-tuning on its validation data changed only the generalist task-module learning rate, to 5×10^{-4}. GS-ECG then reused the ECGFounder-Frozen and DGRNet checkpoints unchanged and fitted its calibration and fusion parameters on $\mathcal{D}_{\text{fusion}}$. All models trained in this study used training seed 2026, and all settings were locked before test inference.

GS-ECG trained 1.54 million parameters on MIMIC-IV-Ext-SHD, comprising the generalist task modules, DGRNet, and the calibration and fusion parameters, 4.9% of the 31.5 million trained by ECGFounder-FT; the proportion was also 4.9% on the EchoNext-Mini Dataset. Because GS-ECG also runs the frozen 30.7-million-parameter ECGFounder backbone at inference, it used 32.2 million parameters at inference, against 31.5 million for ECGFounder-FT and 0.73 million for DGRNet. Experiments were run on Ubuntu 24.04.3 LTS with Python 3.11.15, PyTorch 2.11.0, CUDA 12.8, and cuDNN 9.19.0 on NVIDIA RTX 6000 Ada GPUs, and statistical analyses used Python 3.11.16, NumPy 2.4.6, SciPy 1.17.1, and scikit-learn 1.9.0.

4.9 Comparator models

The *Tabular only* group comprised logistic regression, XGBoost, and a multilayer perceptron using structured variables [42–44]. The *Waveform only* group comprised CNN1D, CRNN, and a Transformer using ECG waveforms [45–47]. The *Tabular + Waveform* group contained ECGFounder-Frozen, ECGFounder-FT, the EchoNext-Mini Model, DGRNet, and GS-ECG. Logistic regression and XGBoost were fitted on all training ECGs. Neural reference models were trained with weighted BCE for up to 100 epochs with early stopping, and their checkpoints were selected by the same validation criterion as the three predictors.

On MIMIC-IV-Ext-SHD, the EchoNext-Mini Model was our task-adapted implementation of the released baseline architecture [16], a compact ResNet1D encoder followed by structured-feature concatenation and multi-label classification, adapted to six structured variables and 11 outputs and trained under the reference-model protocol. On the EchoNext-Mini Dataset, it was evaluated with the authors' released checkpoint, trained on the 72,475 ECGs of the 26,218 patients in the released training split, and the released preprocessing, giving a composite SHD AUROC of 0.8197, consistent with the reported 0.820 [16]. Within each dataset, all models used the same outcome definitions, patient-level splits, and test patients.

4.10 Statistical analysis

Mean AUPRC, calculated as the unweighted macro-average of label-wise average precision across all 11 MIMIC-IV-Ext-SHD or 12 EchoNext-Mini Dataset outputs,

including composite SHD, was the primary endpoint because it summarizes precision–recall performance across outcomes with widely different prevalence. Mean AUROC, composite SHD AUROC, and composite SHD AUPRC were secondary endpoints, and label-wise AUROC and AUPRC were also reported. Average precision was calculated with scikit-learn. All 95% CIs used 1,000 non-stratified patient bootstrap resamples drawn with replacement using seed 2026, applying the same resampled patients to all models and outputs within each dataset and taking the 2.5th and 97.5th percentiles; these intervals quantify test-patient sampling uncertainty for the fixed fitted models. Within each dataset, paired differences in mean AUPRC compared GS-ECG with ECGFounder-Frozen, ECGFounder-FT, and DGRNet, DGRNet with ECGFounder-FT, and ECGFounder-Frozen with the EchoNext-Mini Model; paired differences in all four overall metrics compared GS-ECG with the EchoNext-Mini Model. All paired comparisons used unadjusted 95% percentile intervals.

4.11 Operating-point and fixed-capacity analyses

On MIMIC-IV-Ext-SHD, composite SHD cutoffs were selected on validation data under five rules: the lowest referral fraction achieving NPV of at least 90%, the highest specificity achieving sensitivity of at least 90%, the maximum Youden index, the referral fraction closest to 30% with the lower fraction preferred when two were equally close, and the highest sensitivity achieving PPV of at least 90%. Each cutoff was applied unchanged to the test predictions [48], and test-set 95% CIs used 1,000 patient bootstrap resamples with thresholds held fixed. Number needed to echo (NNE) was defined as referred patients divided by true-positive patients. Score groups were defined below the NPV cutoff, between the two predictive-value cutoffs, and at or above the PPV cutoff.

Fixed-capacity analyses were performed in both datasets. Test patients were ranked by predicted probability separately for each model and output; at capacity f, the $\lceil fN_{\text{test}} \rceil$ highest-scoring patients were referred, with ties resolved by a fixed row order. Capacities ran in one-percentage-point steps from 1% to 50% for composite SHD and from 1% to 20% for aortic stenosis, mitral regurgitation, and tricuspid regurgitation, the valvular abnormalities with at least 200 positive test patients in both datasets. For the referral analysis on MIMIC-IV-Ext-SHD, cumulative composite SHD detection by GS-ECG was traced over all capacities from 0% to 100% and expressed per 1,000 eligible patients, and incremental yield was expressed per 100 additional referrals for successive 10-percentage-point expansions from 10% to 50%. GS-ECG was compared with the authors' released EchoNext-Mini Model on the EchoNext-Mini Dataset and with the retrained model on MIMIC-IV-Ext-SHD. Yields were compared with random referral, under which the expected proportion of affected patients among referred patients equals the test-set prevalence.

Data availability

MIMIC-IV, version 2.2 [28], and MIMIC-IV-ECHO, version 1.0 [30], are available on PhysioNet to credentialed users who complete the required training and sign the data use agreement. MIMIC-IV-ECG, version 1.0 [29], is openly available on PhysioNet.

The EchoNext-Mini Dataset, released on PhysioNet as EchoNext version 1.1.1 [32], is available to registered users who sign its data use agreement. Under these agreements, the derived MIMIC-IV-Ext-SHD labels and data splits can be shared only through PhysioNet.

Code availability

The code used to construct MIMIC-IV-Ext-SHD from the PhysioNet source records and to train and evaluate GS-ECG will be made publicly available upon publication.

Acknowledgements

H.W. has been supported by AIM-CISC (National Institute for Health Research (NIHR202639)); UK's Medical Research Council (MR/S004149/1, MR/X030075/1); British Council (UCL-NMU-SEU International Collaboration On Artificial Intelligence In Medicine: Tackling Challenges Of Low Generalisability And Health Inequality) and (Facilitating Better Urology Care With Effective And Fair Use Of Artificial Intelligence - A Partnership Between UCL And Shanghai Jiao Tong University School Of Medicine); EPSRC (UKRI2701: PAIR: Building a cloneable Pipeline for utilizing foundation AI on EHRs).

Author contributions

J.M. conceived the study, constructed MIMIC-IV-Ext-SHD, developed the models, performed the analyses, and wrote the manuscript. C.F.d.V., D.T.C., and H.W. reviewed and edited the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

References

[1] Ulloa-Cerna, A.E., Jing, L., Pfeifer, J.M., Raghunath, S., Ruhl, J.A., Rocha, D.B., Leader, J.B., Zimmerman, N., Lee, G., Steinhubl, S.R., *et al.*: rECHOmmend: an ECG-based machine learning approach for identifying patients at increased risk of undiagnosed structural heart disease detectable by echocardiography. Circulation **146**(1), 36–47 (2022)

[2] Poterucha, T.J., Jing, L., Ricart, R.P., Adjei-Mosi, M., Finer, J., Hartzel, D., Kelsey, C., Long, A., Rocha, D., Ruhl, J.A., *et al.*: Detecting structural heart disease from electrocardiograms using AI. Nature **644**(8075), 221–230 (2025)

[3] Lang, R.M., Badano, L.P., Mor-Avi, V., Afilalo, J., Armstrong, A., Ernande, L., Flachskampf, F.A., Foster, E., Goldstein, S.A., Kuznetsova, T., *et al.*: Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. European Heart Journal – Cardiovascular Imaging **16**(3), 233–271 (2015)

[4] McDonagh, T.A., Metra, M., Adamo, M., Gardner, R.S., Baumbach, A., Böhm, M., Burri, H., Butler, J., Čelutkienė, J., Chioncel, O., *et al.*: 2021 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure: Developed by the task force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) with the special contribution of the Heart Failure Association (HFA) of the ESC. European Heart Journal **42**(36), 3599–3726 (2021)

[5] Zoghbi, W.A., Adams, D., Bonow, R.O., Enriquez-Sarano, M., Foster, E., Grayburn, P.A., Hahn, R.T., Han, Y., Hung, J., Lang, R.M., *et al.*: Recommendations for noninvasive evaluation of native valvular regurgitation: a report from the American Society of Echocardiography developed in collaboration with the Society for Cardiovascular Magnetic Resonance. Journal of the American Society of Echocardiography **30**(4), 303–371 (2017)

[6] Vahanian, A., Beyersdorf, F., Praz, F., Milojevic, M., Baldus, S., Bauersachs, J., Capodanno, D., Conradi, L., De Bonis, M., De Paulis, R., *et al.*: 2021 ESC/EACTS guidelines for the management of valvular heart disease: Developed by the task force for the management of valvular heart disease of the European Society of Cardiology (ESC) and the European Association for Cardio-Thoracic Surgery (EACTS). European Heart Journal **43**(7), 561–632 (2022)

[7] Siontis, K.C., Noseworthy, P.A., Attia, Z.I., Friedman, P.A.: Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. Nature Reviews Cardiology **18**(7), 465–478 (2021)

[8] Lee, E., Ito, S., Miranda, W.R., Lopez-Jimenez, F., Kane, G.C., Asirvatham,

S.J., Noseworthy, P.A., Friedman, P.A., Carter, R.E., Borlaug, B.A., *et al.*: Artificial intelligence-enabled ECG for left ventricular diastolic function and filling pressure. npj Digital Medicine **7**(1), 4 (2024)

[9] Attia, Z.I., Kapa, S., Lopez-Jimenez, F., McKie, P.M., Ladewig, D.J., Satam, G., Pellikka, P.A., Enriquez-Sarano, M., Noseworthy, P.A., Munger, T.M., *et al.*: Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nature Medicine **25**(1), 70–74 (2019)

[10] Elias, P., Poterucha, T.J., Rajaram, V., Moller, L.M., Rodriguez, V., Bhave, S., Hahn, R.T., Tison, G., Abreau, S.A., Barrios, J., *et al.*: Deep learning electrocardiographic analysis for detection of left-sided valvular heart disease. Journal of the American College of Cardiology **80**(6), 613–626 (2022)

[11] Liang, Y., Sau, A., Zeidaabadi, B., Barker, J., Patlatzoglou, K., Pastika, L., Sieliwonczyk, E., Whinnett, Z., Peters, N.S., Yu, Z., *et al.*: Artificial intelligence-enhanced electrocardiography to predict regurgitant valvular heart diseases: an international study. European Heart Journal **46**(44), 4823–4837 (2025)

[12] Yao, X., Rushlow, D.R., Inselman, J.W., McCoy, R.G., Thacher, T.D., Behnken, E.M., Bernard, M.E., Rosas, S.L., Akfaly, A., Misra, A., *et al.*: Artificial intelligence–enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. Nature Medicine **27**(5), 815–819 (2021)

[13] Bachtiger, P., Petri, C.F., Scott, F.E., Park, S.R., Kelshiker, M.A., Sahemey, H.K., Dumea, B., Alquero, R., Padam, P.S., Hatrick, I.R., *et al.*: Point-of-care screening for heart failure with reduced ejection fraction using artificial intelligence during ECG-enabled stethoscope examination in London, UK: a prospective, observational, multicentre study. The Lancet Digital Health **4**(2), e117–e125 (2022)

[14] Adedinsewo, D.A., Morales-Lara, A.C., Afolabi, B.B., Kushimo, O.A., Mbakwem, A.C., Ibiyemi, K.F., Ogunmodede, J.A., Raji, H.O., Ringim, S.H., Habib, A.A., *et al.*: Artificial intelligence guided screening for cardiomyopathies in an obstetric population: a pragmatic randomized clinical trial. Nature Medicine **30**(10), 2897–2906 (2024)

[15] Dhingra, L.S., Aminorroaya, A., Sangha, V., Pedroso, A.F., Shankar, S.V., Coppi, A., Foppa, M., Brant, L.C., Barreto, S.M., Ribeiro, A.L.P., *et al.*: Ensemble deep learning algorithm for structural heart disease screening using electrocardiographic images: PRESENT SHD. Journal of the American College of Cardiology **85**(12), 1302–1313 (2025)

[16] Hughes, J.W., Jing, L., Finer, J., Hartzel, D., Kelsey, C., Long, A., Rocha, D., Ruhl, J., Poterucha, T., Elias, P.: EchoNext-mini: A dataset and baseline AI model for detecting structural heart disease from electrocardiograms. NEJM AI

3(5), AIdbp2500516 (2026)

[17] Li, J., Aguirre, A.D., Junior, V.M., Jin, J., Liu, C., Zhong, L., Sun, C., Clifford, G., Westover, M.B., Hong, S.: An electrocardiogram foundation model built on over 10 million recordings. NEJM AI **2**(7), AIoa2401033 (2025)

[18] Nolin-Lapalme, A., Sowa, A., Delfrate, J., Tastet, O., Corbin, D., Kulbay, M., Ozdemir, D., Noël, M.-J., Marois-Blanchet, F.-C., Harvey, F., *et al.*: Foundation models for electrocardiogram interpretation: clinical implications. European Heart Journal **47**(18), 2174–2186 (2026)

[19] Kumar, A., Raghunathan, A., Jones, R., Ma, T., Liang, P.: Fine-tuning can distort pretrained features and underperform out-of-distribution. In: International Conference on Learning Representations (2022)

[20] Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Gontijo Lopes, R., Hajishirzi, H., Farhadi, A., Namkoong, H., Schmidt, L.: Robust fine-tuning of zero-shot models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7959–7971 (2022)

[21] Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F.I., Samek, W., Schaeffter, T.: PTB-XL, a large publicly available electrocardiography dataset. Scientific Data **7**(1), 154 (2020)

[22] Strodthoff, N., Wagner, P., Schaeffter, T., Samek, W.: Deep learning for ECG analysis: Benchmarks and insights from PTB-XL. IEEE Journal of Biomedical and Health Informatics **25**(5), 1519–1528 (2021)

[23] Davis, J., Goadrich, M.: The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 233–240 (2006)

[24] Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE **10**(3), e0118432 (2015)

[25] Johnson, A.E.W., Bulgarelli, L., Shen, L., *et al.*: MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data **10**(1), 1 (2023) https://doi.org/10.1038/s41597-022-01899-x

[26] Killalea, M.M., Borgeson, J., Kane, C.J., Kovac, J., Greason, C., Gulati, M.A., Schonfeld, D.A., Crestanello, M.M., Zacher, J., Tsaban, G., *et al.*: Improving structural heart disease screening: AI-ECG and novice AI-guided focused cardiac ultrasound. npj Digital Medicine **9**, 712 (2026) https://doi.org/10.1038/s41746-026-02856-3

[27] Pollard, T., Moody, B.E., Lehman, L., Gow, B., Fernandes, C., Xie, C., Johnson,

A., Mark, R.G., Heldt, T.: PhysioNet as a global platform for biomedical research. Nature Health (2026) https://doi.org/10.1038/s44360-026-00096-z

[28] Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L.A., Mark, R.: MIMIC-IV. PhysioNet (2023) https://doi.org/10.13026/6mm1-ek67 . Version 2.2

[29] Gow, B., Pollard, T., Nathanson, L.A., Johnson, A., Moody, B., Fernandes, C., Greenbaum, N., Waks, J.W., Eslami, P., Carbonati, T., Chaudhari, A., Herbst, E., Moukheiber, D., Berkowitz, S., Mark, R., Horng, S.: MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset. PhysioNet (2023) https://doi.org/10.13026/4nqg-sb35 . Version 1.0

[30] Gow, B., Pollard, T., Greenbaum, N., Moody, B., Han, A., Waks, J.W., Johnson, A., Herbst, E., Eslami, P., Chaudhari, A., Carbonati, T., Berkowitz, S., Mark, R., Horng, S.: MIMIC-IV-ECHO: Echocardiogram Matched Subset. PhysioNet (2026) https://doi.org/10.13026/nrjh-5r77 . Version 1.0

[31] Mukherjee, M., Rudski, L.G., Addetia, K., *et al.*: Guidelines for the echocardiographic assessment of the right heart in adults and special considerations in pulmonary hypertension: recommendations from the American Society of Echocardiography. Journal of the American Society of Echocardiography **38**(3), 141–186 (2025) https://doi.org/10.1016/j.echo.2025.01.006

[32] Elias, P., Finer, J.: EchoNext: A Dataset for Detecting Echocardiogram-Confirmed Structural Heart Disease from ECGs. PhysioNet (2026) https://doi.org/10.13026/7cfw-d091 . Version 1.1.1

[33] Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)

[34] Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K.: WaveNet: A generative model for raw audio. arXiv preprint arXiv:1609.03499 (2016)

[35] Bai, S., Kolter, J.Z., Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271 (2018)

[36] Wu, Y., He, K.: Group normalization. In: Computer Vision – ECCV 2018. Lecture Notes in Computer Science, vol. 11217, pp. 3–19. Springer, Cham (2018)

[37] Hendrycks, D., Gimpel, K.: Gaussian error linear units (GELUs). arXiv preprint arXiv:1606.08415 (2016)

[38] Dauphin, Y.N., Fan, A., Auli, M., Grangier, D.: Language modeling with gated convolutional networks. In: Proceedings of the 34th International Conference on

Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 933–941 (2017)

[39] Inoue, H.: Multi-sample dropout for accelerated training and better generalization. arXiv preprint arXiv:1905.09788 (2019)

[40] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)

[41] Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. In: International Conference on Learning Representations (2017)

[42] Cox, D.R.: The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological) **20**(2), 215–232 (1958) https://doi.org/10.1111/j.2517-6161.1958.tb00292.x

[43] Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016). https://doi.org/10.1145/2939672.2939785

[44] Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**, 533–536 (1986) https://doi.org/10.1038/323533a0

[45] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998) https://doi.org/10.1109/5.726791

[46] Zihlmann, M., Perekrestenko, D., Tschannen, M.: Convolutional recurrent neural networks for electrocardiogram classification. In: 2017 Computing in Cardiology (CinC), pp. 1–4 (2017). https://doi.org/10.22489/CinC.2017.070-060

[47] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008 (2017)

[48] Collins, G.S., Moons, K.G.M., Dhiman, P., Riley, R.D., Beam, A.L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J.B., Smeden, M., *et al.*: TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ **385**, e078378 (2024) https://doi.org/10.1136/bmj-2023-078378

Supplementary Files

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

- supplementarymaterial.pdf