

simultaneously and uniformly¹⁵, offering more efficient and objective solutions. However, a comprehensive evaluation of whether an end-to-end deep learning approach can be used to analyze CMR data to screen for and diagnose a broad range of CVDs remains lacking¹⁶. The few applications of deep learning in CMR so far have focused on single aspects of CMR interpretation (for example, segmentation^{17–19} or wall thickness measurement²⁰) or have demonstrated limited diagnostic capabilities (for example, myocardial scarring or aortic valve malformations^{21–23}).

In this Article, we aimed to develop and validate a deep learning approach for automatic, computerized CMR interpretation and diagnosis consisting of a two-stage paradigm that mimics the clinical workflow: (1) screening for anomalies using nonenhanced cine magnetic resonance imaging (MRI) followed by (2) diagnosing CVDs using cine and late gadolinium enhancement (LGE) MRI as combined inputs. The initial stage, based on cine modality, enables a noninvasive cardiac screening. Compared with LGE, which requires the injection of a gadolinium contrast agent²⁴, cine MRI is safer and more easily acquired. The second stage provides classification of 11 types of CVDs covering most patients referred to the CMR examination²⁵ (ischemic heart disease, most types of nonischemic cardiomyopathy²⁶, pulmonary hypertension and congenital heart disease; Table 1). We propose video-based swin transformer (VST)²⁷—a cutting-edge advancement in computer vision—as our model backbone of choice instead of the conventional convolutional neural network (CNN) approach, and highlighted the superiority of the transformer model in modeling CMR sequences. The proposed automatic pipeline consists of two serial VST-based artificial intelligence (AI) models: the screening model and the diagnostic model (Fig. 1). Further, we examined which imaging modality (cine or LGE), view (four chamber or short axis) and their aggregation should be utilized for optimal classification performance. Finally, we compared the performance of the AI model with physicians of varying experience in CMR interpretation. This study creates an avenue for accurate CMR interpretation in real time, as well as bringing CMR into more widespread use in CVD screening and diagnosis.

Results

Datasets and study design

We curated a nationwide, large representative CMR dataset of 9,719 individuals (6,608 male and 3,111 female) from eight medical centers across China. The dataset was divided into the CVD cohort and the normal control cohort. The disease cohort comprised 8,066 patients with CVD (mean (±s.d.) age 47.2 ± 15 years, 70% male, admitted between 2016 and 2022). Eleven types of CVDs were incorporated with the following distribution: hypertrophic cardiomyopathy (HCM; 2,715), dilated cardiomyopathy (DCM; 1,639), coronary artery disease (CAD; 1,241), left ventricular noncompaction cardiomyopathy (LVNC; 321), restrictive cardiomyopathy (RCM; 377), cardiac amyloidosis (CAM; 358), hypertensive heart disease (HHD; 509), myocarditis (153), arrhythmogenic right ventricular cardiomyopathy (ARVC; 424), pulmonary arterial hypertension (PAH; 200) and Ebstein's anomaly (129). The baseline CMR scan (pretreatment) of each patient, with short-axis (SAX) cine, four-chamber (4CH) cine and SAX LGE all available, was collected to establish the disease cohort. In addition, the SAX cine and 4CH cine of 1,653 normal subjects (age 38 ± 15 years, 56% male, enrolled between 2016 and 2022) were collected to assemble the normal control cohort without CVDs, allowing us to develop and validate the noninvasive screening model. Table 1 and Extended Data Table 1 contain the summary statistics and the demographics of the datasets. The inclusion–exclusion cascade is summarized in Methods and Extended Data Fig. 1.

For the data acquisition, cardiac MRI was performed using three vendors with the following distribution: GE Healthcare (4,569), Philips (3,683) and Siemens (1,467). Cine sequence was performed in SAX orientation covering the whole left ventricle (LV) (SAX cine), as well as in long-axis covering the two-chamber, three-chamber and 4CH view.

All cine sequences were 25 frames (cardiac cycle). LGE images cover the LV from the apex to the base (SAX LGE). We report performance as assessed from two major views of cine examination: SAX cine and 4CH cine, as well as SAX LGE (Extended Data Fig. 2). Supplementary Videos 1–11 show video and image examples for each class.

We used the CMR data from the Beijing Fuwai Hospital²⁸ as the primary dataset for model development and data pooled from all the other medical centers as external test sets. For both screening and diagnostics, threefold cross-validation was performed within the primary dataset to further validate performance. This involved a total of 7,900 subjects and 6,650 CVD patients from the primary dataset contributing to the training of the screening and diagnostic models, respectively. Each fold of cross-validation employed 5,267 patients for screening model training and 4,433 for diagnostic model training. Overall, the screening and diagnostic models were tested with 9,719 and 8,066 patients (internal and external), respectively, and included patients from eight medical centers and CMR acquired from three different MRI vendors.

Evaluation of screening model

The screening model with cine MRI from two combined views (SAX cine and 4CH cine) achieved an area under the curve (AUC) of 0.986 (95% confidence interval (CI) 0.984–0.988) and F_1 score of 0.977 (95% CI 0.974–0.979) for screening on the threefold cross-validation upon the primary dataset ($n = 7,900$) (Fig. 2 and Extended Data Table 2). The sensitivity of 0.973 (95% CI 0.968–0.978) was achieved by the model for anomaly detection with specificity at 90%. All sensitivity and specificity pairs were >90%. It is worth noting that the primary dataset contained a wide spectrum of CVDs (11 types; Table 1), demonstrating the robustness of the screening model with respect to disease type.

In the evaluation of each view of cine for screening, the model derived from 4CH view received an AUC of 0.974 (95% CI 0.969–0.979) and the model derived from SAX view received an AUC of 0.971 (95% CI 0.965–0.976). The combination of SAX and 4CH cine together provided the best performance in comparison to models derived from single-view input (Extended Data Table 2). Note that greater than 95% sensitivity was achieved by both single-view models for anomaly detection with specificity at 90% (Extended Data Table 2). This demonstrates the potential of fast screening based on cine sequence from either SAX or 4CH view.

Evaluation of diagnostic model

Next, we developed the diagnostic model to classify 11 CVD classes. Cine from both views (SAX and 4CH cine) and SAX LGE are combined inputs to the diagnostic model to ensure that any piece of complementary information present in CMR is effectively used to improve the diagnostic accuracy. Upon threefold cross-validation in the primary dataset ($n = 6,650$), the model achieved a class-weighted average AUC of 0.991 and F_1 score of 0.906 (Fig. 2 and Extended Data Table 3). The model achieved an AUC of greater than 0.96 for all classes; for all classes, all but three (LVNC, HHD and myocarditis) had F_1 scores above 0.80. The model demonstrated high AUCs and F_1 scores for the most prevalent CVDs including HCM (AUC 0.998, 95% CI 0.997–0.999; F_1 0.975, 95% CI 0.971–0.980), DCM (AUC 0.988, 95% CI 0.986–0.990; F_1 0.896, 95% CI 0.884–0.907) and CAD (AUC 0.991, 95% CI 0.988–0.994; F_1 0.921, 95% CI 0.908–0.935). The PAH class also had a high AUC of 0.998 (95% CI 0.995–1.000) and F_1 score of 0.962 (95% CI 0.937–0.984).

We further examined the five input schemes: (1) SAX cine, (2) 4CH cine, (3) SAX and 4CH cine, (4) SAX LGE and (5) the combination of SAX cine, 4CH cine and SAX LGE. The all-input scenario achieved the highest AUC and F_1 across all 11 disease classes (Fig. 3 and Extended Data Table 3). We plotted receiver operating characteristic curves (ROCs) for the 11 disease classes. Figure 3 shows the ROCs of three input schemes (cine, LGE and cine + LGE). Notably, the combination of cine and LGE MRIs substantially outperforms models

Table 1 | Characteristics of the primary and external test datasets

	Primary dataset				External test dataset				Entire dataset
	No. of subjects	Sex		Age in years (range)	No. of subjects	Sex		Age in years (range)	
		Male	Female			Male	Female		
Total	7,900	5,380 (68%)	2,520 (32%)	45±16 (2–86)	1,819	1,228 (68%)	591 (32%)	47±16 (1–88)	9,719
Normal control cohort	1,250	700 (56%)	550 (44%)	37±14 (10–78)	403	230 (57%)	173 (43%)	41±16 (6–79)	1,653
CVD cohort	6,650	4,680 (70%)	1,970 (30%)	47±15 (2–86)	1,416	998 (71%)	418 (29%)	48±16 (1–88)	8,066
1 HCM	2,327	1,513 (65%)	814 (35%)	48±14 (7–86)	388	260 (67%)	128 (33%)	51±15 (9–86)	2,715
2 DCM	1,435	1,076 (75%)	359 (25%)	44±15 (4–82)	204	140 (69%)	64 (31%)	50±14 (8–76)	1,639
3 CAD	942	829 (88%)	113 (12%)	56±11 (8–83)	299	269 (90%)	30 (10%)	56±11 (24–88)	1,241
4 LVNC	291	192 (66%)	99 (34%)	39±16 (6–77)	30	18 (60%)	12 (40%)	40±14 (11–65)	321
5 RCM	355	170 (48%)	185 (52%)	50±20 (7–85)	22	13 (59%)	9 (41%)	38±24 (1–78)	377
6 CAM	220	156 (71%)	64 (29%)	56±11 (18–83)	138	92 (67%)	46 (33%)	59±9 (29–82)	358
7 HHD	402	366 (91%)	36 (9%)	42±13 (12–75)	107	88 (82%)	19 (18%)	45±14 (21–75)	509
8 Myocarditis	87	64 (74%)	23 (26%)	28±11 (14–69)	66	48 (73%)	18 (27%)	26±12 (8–68)	153
9 ARVC	370	245 (66%)	125 (34%)	39±14 (9–74)	54	37 (68%)	17 (32%)	40±14 (13–67)	424
10 PAH	134	36 (27%)	98 (73%)	32±12 (10–72)	66	22 (33%)	44 (67%)	38±17 (10–72)	200
11 Ebstein's anomaly	87	33 (38%)	54 (62%)	34±16 (2–63)	42	11 (26%)	31 (74%)	32±14 (6–61)	129

derived from any single modality, with 1.9% points improvement in the averaged AUC metric and 6.8% points improvement in the averaged F_1 metric (compared with SAX cine). All sensitivity and specificity pairs were >90% (Extended Data Table 4). The positive predictive value (PPV) and negative predictive value (NPV) scores are provided in Supplementary Table 1.

Generalization to external test set

To assess whether our models could be transferred to different institutions with varying data collection protocols, we validated the screening and diagnostic models on external test sets collected from seven medical centers ($n = 1,819$; 403 normal subjects and 1,416 patients with CVDs). Our screening model for anomaly detection attained an AUC of 0.990 (95% CI 0.986–0.992), F_1 score of 0.970 (95% CI 0.964–0.977), sensitivity of 0.959 (95% CI 0.936–0.974) with specificity at 90%, and specificity of 0.970 (95% CI 0.950–0.990) with sensitivity at 90% (Fig. 2 and Extended Data Table 2). The diagnostic model (with all-input scenario) for CVD classification achieved a class-weighted AUC of 0.991 and F_1 score of 0.884 (Fig. 2 and Extended Data Table 5). This indicates that the AI model can generalize across diverse data sources, including medical centers uninvolved during model development.

In addition, we examined the generalizability of models derived from a single imaging modality. The diagnostic models based on cine (SAX and 4CH views) film and LGE achieved cross-institution F_1 scores of 0.831 and 0.792, respectively (Extended Data Table 5). For the screening task, the cross-institution performance was 0.953 (95% CI 0.942–0.965) of AUC by the model derived from SAX cine and 0.980 (95% CI 0.972–0.986) by the model of 4CH cine (Extended Data Table 2). The findings were consistent with that of the primary dataset: the combination of SAX and 4CH cine provides the best performance for detecting cardiac anomalies; integrating cine and LGE yields the optimal diagnostic performance.

Model interpretability

We leveraged the guided gradient-weighted class activation mapping (Grad-CAM)²⁹ to display an informative set of features and distinct patterns used by the model for classification. Specifically, we extracted the Grad-CAM for representative subjects from 11 CVD categories. Figure 4 shows the AI model activations that contributed to a prediction of CVD.

The LV area shows higher saliency at the detection of HCM, DCM, CAD, LVNC, RCM, CAM, HHD and myocarditis (Fig. 4, yellow background); the right ventricle (RV) was highlighted as salient for the detection of ARVC, PAH and Ebstein's anomaly (Fig. 4, red background). This is consistent with the clinical diagnostic criteria: ARVC, PAH and Ebstein's anomaly are all primarily RV involvement whereas the abnormality for the rest of the classes is mainly present on LV³⁰. In addition, the LGE signal in CAD, CAM, myocarditis and ARVC (Fig. 4, myocardium in SAX LGE, red arrows), which represents myocardial fibrosis or amyloid, was correctly captured by the saliency maps. Furthermore, the model accurately identified the LVNC in the apex and septal leaflet displacement as distinctive features in detecting LVNC and Ebstein's anomaly (Fig. 4, 4CH cine, red arrows), respectively, which is consistent with the underlying pathophysiology of these conditions^{31,32}.

Comparison with human annotations

To compare the performance of the AI model with that of board-certified physicians, we formed a gold-standard test dataset with 500 patients covering 11 types of CVDs (Extended Data Table 6). Each patient was independently evaluated for CVD class by physicians with three levels of experience in CMR reading (3–5 years, 5–10 years and more than 10 years), along with the AI diagnostic model for comparison (Table 2). The AI model achieved comparable performance with physicians with more than 10 years of experience in CMR reading (F_1 score of 0.931 versus 0.927) with faster speed of interpretation (1.94 min versus 418 min for interpreting 500 subjects). In addition, our model exceeded the performance of the most experienced group of physicians (more than 10 years) for the PAH class by successfully identifying CMR-negative patients (F_1 score of 0.983 versus 0.931). This demonstrates the potential of AI to identify MRI features not readily detectable by humans³³, a finding consistent with previous works in oncology^{34–36}.

Comparison of video-based deep learning models

We compared the VST model and the conventional CNN–long short-term memory (LSTM)²¹ approach for modeling CMR sequences. Extended Data Fig. 3 illustrates the schematic overview of the two video-based deep learning algorithms in SAX cine film interpretation. The SAX cine-derived VST model notably outperformed CNN–LSTM with 3.5% points improvement in the AUC and 4.6% points improvement

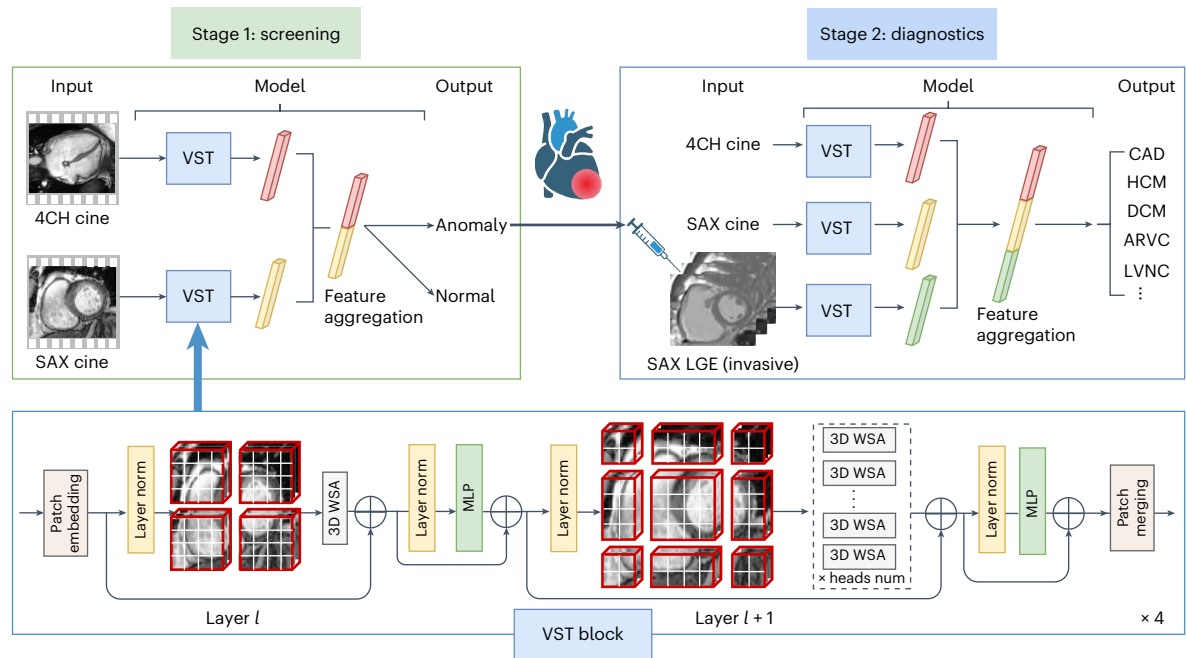


Fig. 1 | Workflow of the two-stage paradigm for automatic screening and diagnostics of CVDs. For each patient, the screening model takes cine movies as inputs and outputs the binary classification to detect cardiac anomaly. Second, the patient suspected of cardiac anomaly undergoes LGE imaging, while the diagnostic model integrates both cine and LGE to output their CVD

class. The AI models comprised four VST blocks to analyze the CMR sequences using the 3D-shifted window self-attention (WSA) mechanism. MLP, multilayer perceptron; norm, normalization; num, number; layer l and layer $l + 1$ indicate two consecutive layers.

in the F_1 score, tested on the primary dataset. This finding demonstrates the superiority of the VST algorithm in CMR analysis.

Validation on an independent consecutive test set

To further evaluate the performance of our developed AI model in a real-world clinical setting, we constructed a fresh independent testing set, consisting of 1,000 subjects consecutively admitted to Beijing Fuwai Hospital in 2023. This consecutive testing set was meticulously designed to be unselected, ensuring a representation of the authentic clinical prevalence and encompassing a diverse spectrum of cardiac disease phenotypes.

Evaluation of the AI screening model. From the 1,000 consecutively collected subjects, we formed a testing set for the screening model comprising 961 subjects with complete cine images, including 159 normal individuals and 802 patients with cardiac anomalies. Thirty-nine subjects were excluded based on the following criteria: (1) missing SAX cine or 4CH cine sequences (22 subjects), (2) SAX cine with fewer than five views (six subjects) and (3) inadequate imaging quality (11 subjects). Utilizing cine MRI from both SAX and 4CH views, the AI screening model demonstrated exceptional performance on the independent consecutive testing set ($n = 961$; Supplementary Table 2), achieving an AUC of 0.984 (95% CI 0.977–0.990) and an F_1 score of 0.962 (95% CI 0.953–0.972) for cardiac anomaly screening. The sensitivity of 0.946 (95% CI 0.930–0.964) was achieved by the screening model for cardiac anomaly detection with specificity at 90%. The screening model performance is detailed in Supplementary Table 2. Notably, the consecutive testing set encompassed a diverse range of CVDs, including mild/borderline cases and suspected phenocopies (for example, inherited metabolic cardiomyopathies), extending beyond the commonly identified 11 CVD classes. This underscores the robustness of the screening model with respect to both disease types and severity.

Evaluation of the AI diagnostic model. From the 1,000 consecutively collected subjects, we formed a testing set for the diagnostic

model, comprising 532 patients with CVD and complete sets of LGE and cine images. To ensure the integrity of the testing set, we established detailed exclusion criteria. Specifically, 159 normal individuals without cardiac anomalies were excluded, along with 222 patients lacking LGE images, which are essential inputs for our diagnostic model. It is crucial to note that LGE, an invasive examination requiring contrast injection, was not consistently performed for all admitted patients. Additionally, 48 patients with CVD, falling beyond the scope of the commonly identified 11 CVD classes, were excluded from the reported quantitative testing performance. Nevertheless, we have included and analyzed the AI screening and diagnostic results for these 48 patients in Supplementary Table 3.

With the established testing set ($n = 532$), our AI diagnostic model, utilizing cine and LGE images as combined inputs, demonstrated exceptional performance. It achieved a class-weighted average AUC of 0.986 and an F_1 score of 0.903 (Supplementary Table 4). Notably, the model exhibited high AUCs and F_1 scores for prevalent CVDs, including HCM (AUC 0.993, 95% CI 0.988–0.997; F_1 0.958, 95% CI 0.940–0.975), DCM (AUC 0.991, 95% CI 0.983–0.996; F_1 0.922, 95% CI 0.883–0.958) and CAD (AUC 0.997, 95% CI 0.994–0.999; F_1 0.915, 95% CI 0.855–0.966). Across all 11 CVD classes, the model achieved an AUC greater than 0.90, with F_1 scores above 0.80 for all except LVNC, HHD, RCM and myocarditis. The CAM class exhibited a high F_1 score of 0.947 and an AUC of 1.0.

Discussion

CMR has been considered the gold standard for assessing cardiac function; its contemporary application encompasses virtually all aspects of CVDs. It shows unique capabilities in the diagnostic workup of suspected CVD³⁷. However, CMR is also one of the most challenging radiologic imaging techniques to interpret due to the complexity of cardiac motion. In this study, we conducted a pioneering investigation in computerized CMR (cine and LGE) interpretation for screening and diagnostics. Our study of 8,066 patients with CVD and 1,653 normal individuals concluded that the screening model for anomaly detection and diagnostic model for CVD classification attained AUCs of

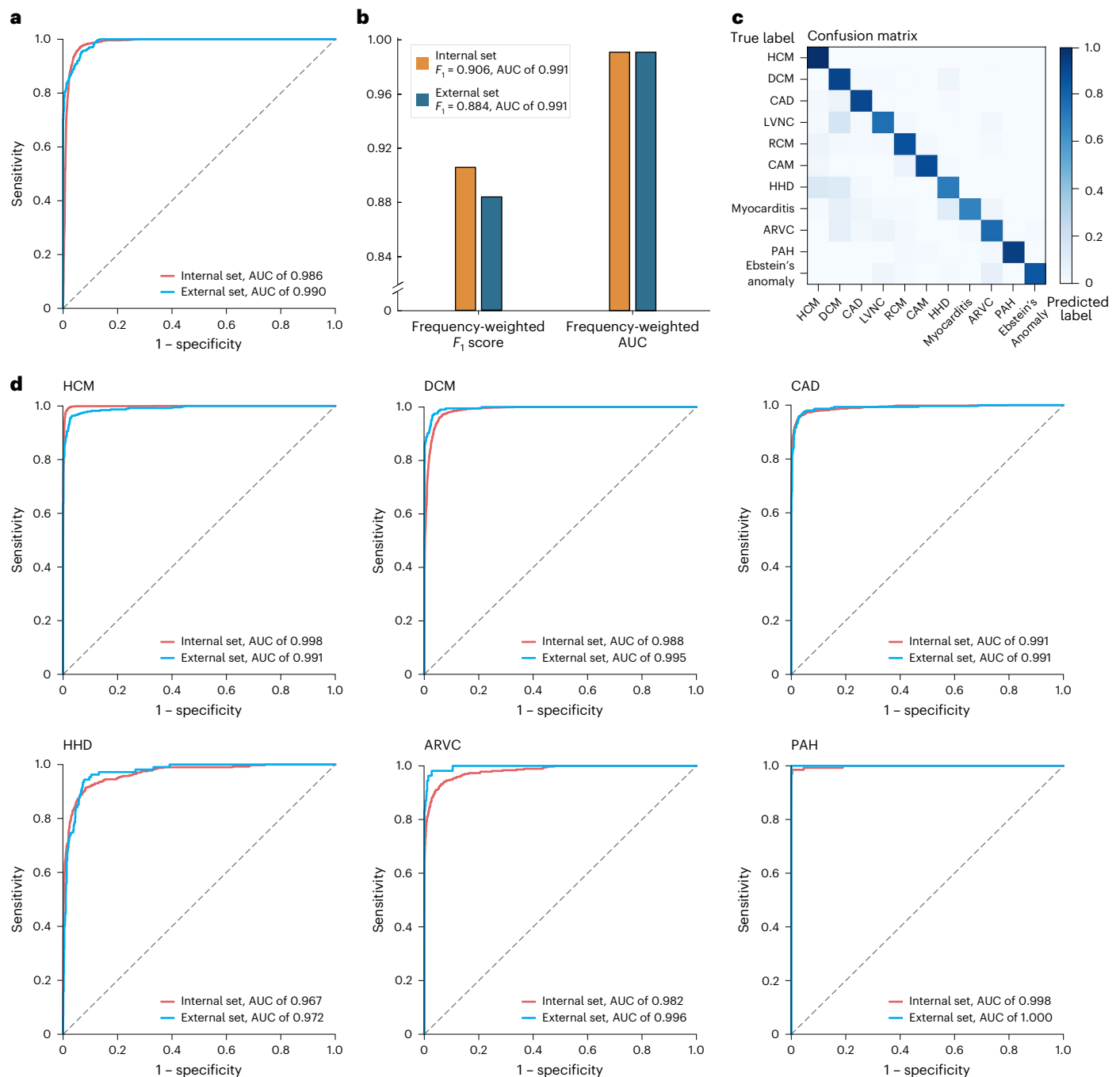


Fig. 2 | Performance of the screening and diagnostic models in internal and external testing. **a**, ROCs for the screening of cardiac anomalies for the primary internal test dataset (blue, $n = 7,900$) and external test dataset (red, $n = 1,819$). The screening model is derived from 4CH cine and SAX cine. **b**, The diagnostic performance for the internal test dataset (yellow, $n = 6,650$) and external test dataset (blue, $n = 1,416$). The diagnostic model takes cine (4CH and SAX) and LGE

as combined inputs. **c**, A confusion matrix for the predictions of the AI diagnostic model versus the ground truth over the entire CVD cohort ($n = 8,066$). The percentage of all possible predictions in each CVD class is displayed on a color gradient scale. **d**, ROCs for the diagnosis of CVD classes for the internal set and external set.

$0.988 \pm 0.3\%$ and $0.991 \pm 0.0\%$ (F_1 scores of $0.974 \pm 0.5\%$ and $0.895 \pm 1.6\%$; mean \pm s.d. of internal set and external set), respectively. These results demonstrate that video-based end-to-end deep learning approaches can reliably detect anomalies and classify various types of CVDs from CMR with high classification performance similar to or even superior to that of experienced cardiologists.

This proof-of-concept study shows an automatic pathway to CMR analysis. The standard clinical approach to CMR interpretation requires experts to (1) manually delineate the contours of the endocardium

and epicardium and (2) scan back and forth across cine film and LGE over a series of SAX and long-axis views. Specifically, a typical CMR examination consists of SAX cine films with nine parallel views (25 frames per view), a 4CH cine film (25 frames), a three-chamber cine film (25 frames), SAX LGEs (nine parallel views) and 4CH LGE, leading to at least 11 videos and 10 images to analyze in total. Hence, this procedure is extremely labor intensive, time consuming and susceptible to operator bias. In contrast, deep neural networks (DNNs) enable an approach that is fundamentally different since the automatic model can absorb

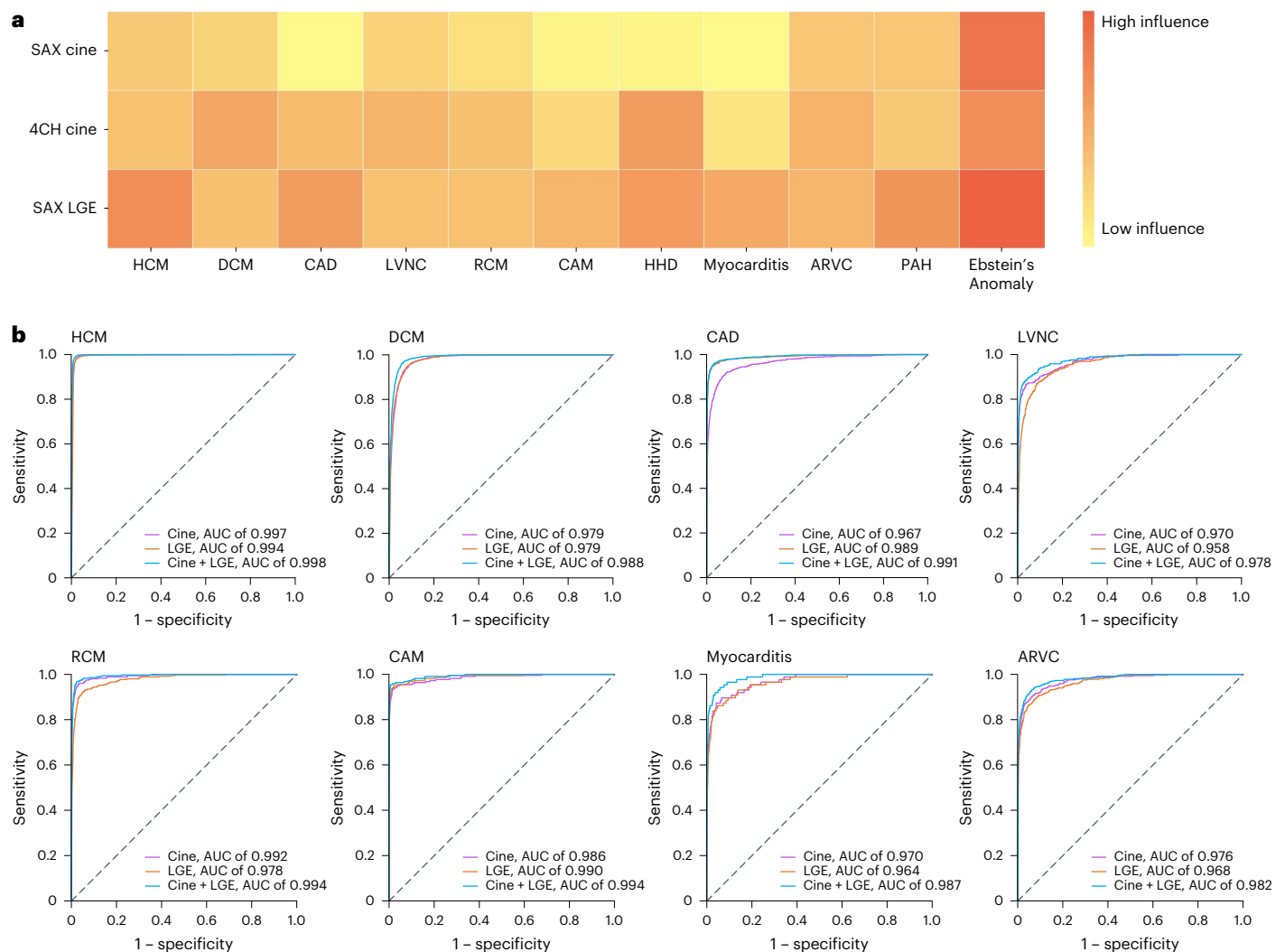


Fig. 3 | Influences of individual CMR modalities. a, Shapley values of SAX cine, 4CH cine and SAX LGE, derived from the diagnostic model (cine and LGE as combined inputs) for the prediction of each CVD class. Shapley values are displayed on a color gradient scale, with red indicating the CMR modality with the greatest influence for each CVD classification. The CMR modalities, exhibiting characteristic features for the diagnosis of the CVD class, demonstrate a consistently strong impact on their model prediction: SAX LGE for the diagnosis of CAD (distinct feature: the endomyocardial or transmural LGE matching

the area of coronary artery dominance); SAX LGE for HCM (hypertrophy and RV insertion point LGE); SAX LGE for myocarditis (epicardial LGE); 4CH cine for LVNC (LV noncompaction in the apex) and 4CH cine for RCM (bi-atrial enlargement on the 4CH view). **b**, ROCs from the diagnostic models based on cine (purple), LGE (yellow) and cine + LGE as combined inputs (blue). Combining cine and LGE yielded the optimal diagnostic performances for all CVD classes. The performance was based on the internal test set.

all pieces of information present in CMR ‘end-to-end’ without requiring manual tracing, calculation of cardiac function or class-specific feature extraction. In other words, the proposed DNN model accepts the raw CMR data as input, learns all of the important features—both previously manually derived and as-yet-unrecognized—in a data-driven way and outputs final diagnostic probabilities.

The high performance of the developed screening models derived from cine MRI suggests a fast, noninvasive and accurate screening technique for detecting CVDs. The screening model derived from 4CH cine achieved an AUC of $0.977 \pm 0.4\%$ (mean \pm s.d. of internal set and external set; Extended Data Table 2); the model derived from SAX cine achieved an AUC of $0.962 \pm 1.3\%$. The single-view schemes yielded similar performance as combined views (the model derived from 4CH and SAX cine received an AUC of $0.988 \pm 0.3\%$). Therefore, the finding that a single view can independently and reliably detect cardiac anomalies indicates that this method can be used to simplify CMR acquisition and improve clinical efficiency. Increased efficiency is beneficial, given the potential to decrease the cost of cine MRI acquisition and enhance

patient throughput. The shortened procedure time is also beneficial for patients who cannot tolerate longer scans. In addition, cine MRI provides high-resolution images for accurate quantitation of ventricular volume, cardiac function and motion estimation, along with detailed signals in myocardium, which together form the cornerstone of diagnosis³⁸. As such, the cine-based screening test can serve to improve the accuracy of anomaly detection in CVD, particularly since there is ample evidence to suggest that the most widely used screening examinations—electrocardiogram (ECG) and echocardiogram—capture only a fraction of the informative features for diagnosis^{39,40}.

CVD diagnosis is one of the most problematic and challenging tasks in cardiology. To address the challenge, this study introduced automatic diagnosis based on CMR. Cine and LGE MRIs together substantially outperformed the model derived from either cine or LGE alone. This finding is consistent with prior studies demonstrating that cine and LGE provide complementary information in CMR diagnosis⁴¹. The diagnostic model derived from cine and LGE yielded an average class-weighted AUC of 0.991 over 11 classes. The 11 classes account for most of the CVDs

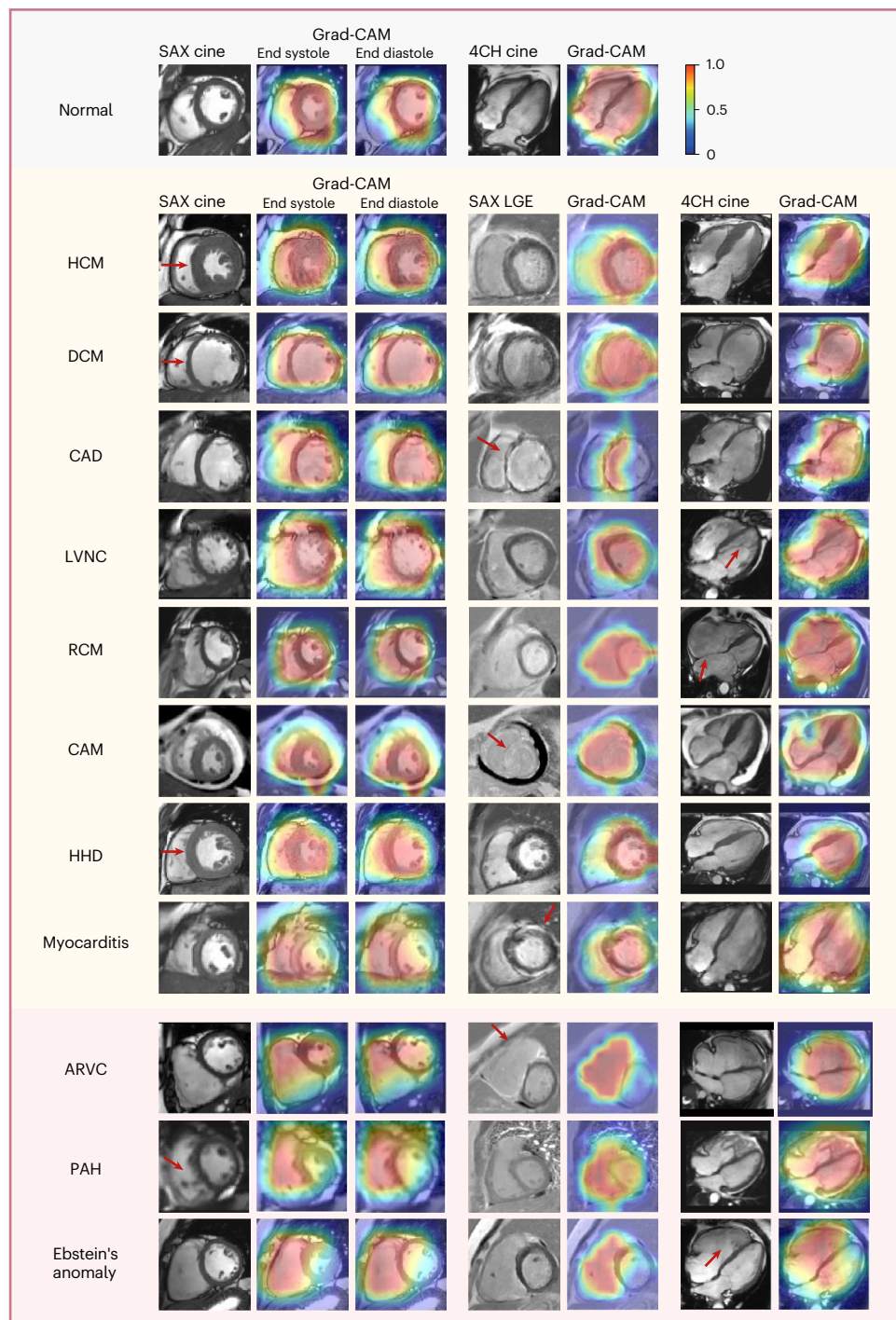


Fig. 4 | Saliency maps of CMR scans from representative patients of eleven CVD classes and the normal control. The saliency map (heat map) was generated using the guided Grad-CAM approach and reveals the region that contributes the most to the AI model's decision. The scale bar ranges from zero to one, with one indicating the highest influence provided by the normalized Grad-CAM value and zero indicating the lowest influence. The red arrows point to the characteristic features of each CVD class, which are consistently encompassed by the saliency maps of the diagnostic model: left ventricular hypertrophy, HCM; enlargement of the left ventricle and thinning of the left ventricular wall,

DCM; endocardial LGE in the ventricular septum and adjacent anterior of the left ventricular wall, CAD; left ventricular noncompaction in the apex, LVNC; bi-atrial enlargement, RCM; diffuse dust-like LGE of the left ventricular myocardium, CAM; symmetric left ventricular hypertrophy, HHD; subepicardial LGE of the left ventricular free wall, myocarditis; right ventricular enlargement with fibrosis, ARVC; enlargement of the RV and thickening of the right ventricular wall, PAH; apical displacement of the septal valve leaflet of the tricuspid valve, Ebstein's anomaly. The CVD classes with yellow background are primarily LV dysfunctions and the classes with red background are primarily RV dysfunctions.

referred for CMR examination (over 90% at Beijing Fuwai Hospital²⁸), making the model broadly applicable. This outcome effectively propels us toward making efficient and precise CVD diagnosis that has a significant clinical impact. As provider confirmation will still be needed in many

clinical settings and ambiguous cases, we expect the diagnosis model to complement, not replace, cardiologists. The AI model could expand the capability of a CMR-trained cardiologist in the clinical workflow by triaging the readings for which the model has the least 'confidence'.

Table 2 | Diagnostic performance of the AI model compared with physicians with varying experience (range from 3 to >10 years) in CMR reading

		No. of subjects (n=500)	F ₁ score			
			AI model	Physician (3–5 years)	Physician (5–10 years)	Physician (>10 years)
1	HCM	100	0.971	0.957	0.938	0.962
2	DCM	100	0.914	0.853	0.911	0.940
3	CAD	80	0.962	0.916	0.949	0.969
4	LVNC	30	0.877	0.667	0.778	0.885
5	RCM	30	0.933	0.578	0.760	0.800
6	CAM	30	0.947	0.667	0.931	0.931
7	HHD	30	0.833	0.615	0.667	0.896
8	Myocarditis	20	0.857	0.553	0.600	0.683
9	ARVC	30	0.897	0.451	0.814	0.983
10	PAH	30	0.983	0.061	0.929	0.931
11	Ebstein's anomaly	20	0.950	0.519	0.842	0.974
Frequency-weighted F ₁			0.931	0.734	0.872	0.927
Accuracy			0.932	0.746	0.868	0.928
Time cost (in total)			1.94 min	576 min	329 min	418 min

The physicians are categorized according to their number of years of experience in CMR interpretation. The bold font emphasizes the superior performance metric among subgroups, including the AI model and physicians with varying levels of experience.

Moreover, the AI model's ability to outperform cardiologists in diagnosing PAH by successfully identifying CMR-negative cases (that is, confirmed PAH without abnormal CMR findings) can have marked clinical impact by allowing for less invasive diagnosis of PAH. PAH is a progressive condition with high mortality, and timely diagnosis is vital for its treatment⁴². The current gold standard for diagnosis of PAH is right heart catheterization, which is an invasive procedure that can introduce serious surgical complications including hematoma, pneumothorax, arrhythmias and hypotensive episodes^{43–45}. The conventional CMR evaluation of the RV has been used to assess the severity of PAH and monitor its prognosis and therapy response⁴⁶. While CMR's diagnostic utility in PAH is largely underexplored due to its technical complexity⁴⁷, the AI-empowered CMR interpretation demonstrated in this study offers a timely and valuable perspective and pathway for an accurate, safe and rapid PAH diagnosis.

Of the CVD classes we examined, myocarditis is a clinically important CVD for which the diagnostic model derived from cine and LGE had a lower F₁ score compared with other CVD classes (internal set: 0.724; external set: 0.630). A manual review of the discordances revealed that the model misclassifications overall appear very reasonable. For example, some instances of mild myocarditis only present mild elevation of troponin with no remarkable myocardial necrosis, leading to an LGE-negative result. Meanwhile, the edema and functional ventricular impairment could be relieved if patients with myocarditis are not scanned in the appropriate time window, resulting in CMR negativity. This is consistent with the general findings: the sensitivity of myocarditis diagnosis based on the Lake Louise criteria—the diagnostic CMR imaging criteria for patients with suspected myocarditis—only reaches 0.780–0.875 (refs. 48,49). Moreover, for myocarditis diagnosis, the lack of T2-weighted images and parametric myocardial mapping⁵⁰ limited the conclusions that could reasonably be drawn from the cine and LGE MRI, making it more difficult to definitively ascertain whether the cardiologists and/or the AI model was correct.

We emphasize our use in this study of a CMR dataset representative enough (covering a wide spectrum of 11 types of CVDs, accounting for above 90% of the CVD patients referred for CMR examination and CMR acquired by three major vendors) to evaluate end-to-end deep learning approaches for screening and diagnostics and our comprehensive

internal and external validations of 9,719 subjects pooled from eight medical centers. We leveraged more than one million cardiac MRI images comprising 38,876 cine films and 72,594 LGE images. To the best of our knowledge, large pooled CMR databases containing both cine and LGE modalities that can be used to diagnose a wide range of heart conditions do not currently exist. As such, our collected cohort is unique in that it is the largest and first-ever complete CMR database with cine and LGE MRIs for AI-enabled studies.

We leveraged VST as our model backbone of choice in CMR interpretation. Transformer-based deep learning architectures very recently expanded to image and video processing and yield substantial improvements on a wide spectrum of high-level computer vision tasks. VST, a transformer adapted for video sequence processing, has shown impressive performance on the major video recognition benchmarks²⁷. However, few efforts have been made to explore its role in medical video analysis. As opposed to the conventional CNNs, which are limited by the small receptive field of the convolution operation, the global self-attention and shifted window mechanism inherent in VST broadens the receptive field and allows effective integration of temporal and spatial information from cardiac video and three-dimensional (3D) sequences. The superiority of VST confirmed in this study offers insight into the use of AI-enabled medical video analysis within and beyond CMR imaging.

Several limitations need to be considered when interpreting the presented results. Extensive evaluation through prospective studies and clinical trials is necessary before the models' clinical implementation. The reported algorithmic performance may not translate to real-world deployment, necessitating further validation. All participating institutions are from eastern Asia. The model generalizability across different ethnicities should be investigated in future work to ensure its broad utility. The number of health controls was limited compared with the overall study population. Owing to this, a more comprehensive assessment of the screening model based on a dataset with real-world CVD prevalence is warranted. While the screening model demonstrates robustness in handling abnormal cases outside the specified 11 commonly encountered CVD classes, the diagnostic model's ability to distinguish cases with unique phenocopies, such as Fabry disease, inherited metabolic cardiomyopathies and instances

34. Hiremath, A. et al. An integrated nomogram combining deep learning, Prostate Imaging–Reporting and Data System (PI-RADS) scoring, and clinical variables for identification of clinically significant prostate cancer on biparametric MRI: a retrospective multicentre study. *Lancet Digit. Health* **3**, e445–e454 (2021).
35. Zhang, M. et al. MRI radiogenomics of pediatric medulloblastoma: a multicenter study. *Radiology* **304**, 406–416 (2022).
36. Liu, Z. et al. Predicting distant metastasis and chemotherapy benefit in locally advanced rectal cancer. *Nat. Commun.* **11**, 4308 (2020).
37. Schulz-Menger, J. et al. Standardized image interpretation and post-processing in cardiovascular magnetic resonance-2020 update. *J. Cardiovasc. Magn. Reson.* **22**, 1–22 (2020).
38. Members, W. C. et al. ACCF/ACR/AHA/NASCI/SCMR 2010 expert consensus document on cardiovascular magnetic resonance: a report of the American College of Cardiology Foundation Task Force on Expert Consensus Documents. *Circulation* **121**, 2462–2508 (2010).
39. Valente, A. M. et al. Comparison of echocardiographic and cardiac magnetic resonance imaging in hypertrophic cardiomyopathy sarcomere mutation carriers without left ventricular hypertrophy. *Circ. Cardiovasc. Genet.* **6**, 230–237 (2013).
40. Capron, T. et al. Cardiac magnetic resonance assessment of left ventricular dilatation in chronic severe left-sided regurgitations: comparison with standard echocardiography. *Diagn. Interv. Imaging* **101**, 657–665 (2020).
41. Chatzantonis, G. et al. Diagnostic value of cardiovascular magnetic resonance in comparison to endomyocardial biopsy in cardiac amyloidosis: a multi-centre study. *Clin. Res. Cardiol.* **110**, 555–568 (2021).
42. Swift, A. J. et al. A machine learning cardiac magnetic resonance approach to extract disease features and automate pulmonary arterial hypertension diagnosis. *Eur. Heart J. Cardiovasc. Imaging* **22**, 236–245 (2021).
43. Hoepfer, M. M. et al. Complications of right heart catheterization procedures in patients with pulmonary hypertension in experienced centers. *J. Am. Coll. Cardiol.* **48**, 2546–2552 (2006).
44. D'Alto, M. et al. Right heart catheterization for the diagnosis of pulmonary hypertension: controversies and practical issues. *Heart Fail. Clin.* **14**, 467–477 (2018).
45. Taylor, C., Derrick, G., McEwan, A., Haworth, S. & Sury, M. Risk of cardiac catheterization under anaesthesia in children with pulmonary hypertension. *Br. J. Anaesth.* **98**, 657–661 (2007).
46. Alabed, S. et al. Cardiac magnetic resonance in pulmonary hypertension—an update. *Curr. Cardiovasc. Imaging Rep.* **13**, 1–9 (2020).
47. Johns, C. S., Wild, J. M., Rajaram, S., Swift, A. J. & Kiely, D. G. Current and emerging imaging techniques in the diagnosis and assessment of pulmonary hypertension. *Expert Rev. Respir. Med.* **12**, 145–160 (2018).
48. Kotanidis, C. P. et al. Diagnostic accuracy of cardiovascular magnetic resonance in acute myocarditis: a systematic review and meta-analysis. *JACC Cardiovasc. Imaging* **11**, 1583–1590 (2018).
49. Luetkens, J. A. et al. Comparison of original and 2018 Lake Louise criteria for diagnosis of acute myocarditis: results of a validation cohort. *Radiol. Cardiothorac. Imaging* **1**, e190010 (2019).
50. Friedrich, M. G. & Marcotte, F. Cardiac magnetic resonance assessment of myocarditis. *Circ. Cardiovasc. Imaging* **6**, 833–839 (2013).
51. Dvijotham, K. et al. Enhancing the reliability and accuracy of AI-enabled diagnosis via complementarity-driven deferral to clinicians. *Nat. Med.* **29**, 1814–1820 (2023).
52. Elliott, P. et al. Classification of the cardiomyopathies: a position statement from the European Society Of Cardiology Working Group on Myocardial and Pericardial Diseases. *Eur. heart J.* **29**, 270–276 (2008).
53. Limongelli, G. et al. Diagnosis and management of rare cardiomyopathies in adult and paediatric patients. a position paper of the Italian Society of Cardiology (SIC) and Italian Society of Paediatric Cardiology (SICP). *Int. J. Cardiol.* **357**, 55–71 (2022).
54. Mavrogeni, S. et al. T1 and T2 mapping in cardiology: 'mapping the obscure object of desire'. *Cardiology* **138**, 207–217 (2017).
55. Kidoh, M. et al. Myocardial tissue characterization by combining extracellular volume fraction and T2 mapping. *Cardiovasc. Imaging* **15**, 700–704 (2022).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024

¹School of Medicine, Stanford University, Stanford, CA, USA. ²Department of Magnetic Resonance Imaging, Fuwai Hospital and National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China. ³Changhong AI Research (CHAIR), Sichuan Changhong Electronics Holding Group, Mianyang, China. ⁴Department of Biomedical Engineering, University of Southern California, Los Angeles, CA, USA. ⁵Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA. ⁶School of Engineering, University of Science and Technology of China, Hefei, China. ⁷Department of Computer Science, Stony Brook University, New York, NY, USA. ⁸Paul C. Lauterbur Research Center for Biomedical Imaging, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. ⁹Department of Electrical Engineering, Stanford University, Stanford, CA, USA. ¹⁰Stanford Cardiovascular Institute, School of Medicine (Division of Cardiology), Stanford University, Stanford, CA, USA. ¹¹Peking Union Medical College Hospital, Beijing, China. ¹²Guangdong Provincial People's Hospital, Guangzhou, China. ¹³Beijing Anzhen Hospital, Beijing, China. ¹⁴Tongji Hospital, Wuhan, China. ¹⁵The Second Affiliated Hospital of Harbin Medical University, Harbin, China. ¹⁶Renji Hospital, Shanghai, China. ¹⁷The First Hospital of Lanzhou University, Lanzhou, China. ¹⁸The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, China. ¹⁹Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL, USA. ²⁰Mayo Clinic Alix School of Medicine, Phoenix, AZ, USA. ²¹These authors contributed equally: Yan-Ran (Joyce) Wang, Kai Yang. ²²These authors jointly supervised this work: Yan-Ran (Joyce) Wang, Joseph C. Wu, Shihua Zhao. ✉ e-mail: wangyanran100@gmail.com; kk.zhao@siat.ac.cn; cjrzhao@shihua2009@163.com

Methods

Ethics approval

The CMR datasets were acquired retrospectively under the approval of the institutional review boards (IRBs) at each participating institution, including Beijing Fuwai Hospital, Beijing Anzhen Hospital, Guangdong Provincial People's Hospital, the 2nd Affiliated Hospital of Harbin Medical University, the First Hospital of Lanzhou University, Renji Hospital, Tongji Hospital and Peking Union Medical College Hospital. Informed consent was waived by the IRBs. Before model training, testing and reader studies, all data underwent deidentification processes.

Datasets

The CMR database search was performed for all eight centers to identify CVDs and normal controls. All data were anonymized and deidentified, as per the Health Insurance Portability and Accountability Act Safe Harbor provision⁵⁶. Inclusion criteria were (1) patients with a definitive diagnosis of CVD and (2) patients with CMR scans at baseline before surgical treatment, if any. Exclusion criteria were (1) incomplete cine or LGE modalities, (2) SAX cine with fewer than five views, (3) CMR images with insufficient scan quality, (4) CVD patients missing clinical data and (5) CMR examinations that could not be interpreted and agreed upon by the committee cardiologists according to the diagnostic criteria (Methods). The detailed diagnostic criteria of the 11 types of CVDs and normal controls included in this study was described in Methods. Table 1 and Extended Data Table 1 present the detailed demographics and distribution of the primary dataset and the external validation sets collected from the other seven medical centers across China. To offer a comprehensive perspective on our primary development dataset, we went the extra mile by collecting the LV ejection fraction (LVEF) metric for all 7,900 subjects (including 1,250 normal controls and 6,650 patients with CVD) within the primary dataset. We meticulously summarized the distribution of demographics and LVEF across the 11 specified CVD classes and the normal control class in Supplementary Table 5. Additionally, we generated density plots to illustrate the distribution of LVEF for each class in the primary dataset, offering a more comprehensive representation (Supplementary Fig. 1).

The fresh consecutive testing set is designed to capture the genuine spectrum of disease phenotypes in the real-world clinical prevalence. To offer a thorough understanding of the severity of cases in alignment with real-world clinical prevalence, we have presented five key cardiac function metrics. These metrics include LVEF, LV mass, LVMI (LV mass index), LV end-diastolic volume and LV end-diastolic volume index. Supplementary Table 6 presents the distribution of demographics and the cardiac functions across 11 CVD classes and the normal control class in the fresh consecutive testing set. For improved visualization and clarity, we have depicted the prevalence of the 11 CVD classes in both the fresh consecutive testing set ($n = 532$ patients with CVD) and the primary discovery dataset ($n = 6,650$ patients with CVD) using pie charts in Supplementary Fig. 2. The fresh consecutive testing set offers a representation of the genuine clinical prevalence. Through direct comparison, it is evident that the primary dataset and the consecutive testing set exhibit very similar CVD prevalence and distribution. The top three most prevalent CVDs referred to the CMR examination remain HCM, DCM and CAD.

All images were acquired by breath-holding and electrocardiographic gating. A balanced steady-state free precession sequence was used for cine images with a continuous sampling from the basal to the apical levels on SAX views and two-chamber, three-chamber and 4CH long-axis views. We included cine MRI from two views in this study: the standard SAX cine and the long-axis 4CH cine. The SAX cine clearly depicts the RV and the LV. The 4CH cine shows the four chambers of heart: right atrium, left atrium, RV and LV.

LGE MRI has been established as the gold standard reference for myocardial viability and replacement fibrosis in the myocardium^{57,58}. In our CMR cohorts, the LGE images were obtained using phase-sensitive

inversion recovery sequence with a segmented FLASH readout scheme performed 10–15 min after injection of gadolinium-based contrast with 0.15 mmol kg⁻¹ per bolus. Gadolinium contrast agents can be used to detect areas of fibrosis, as the prolonged washout of the contrast correlates with a reduction in functional capillary density in the irreversibly injured myocardium⁵⁹. The SAX LGE used in the study was acquired from the SAX view with the same section thickness, covering the entire left ventricle from the base to the apex (nine parallel views for most cases). Note that LGE is an invasive examination that requires contrast injection and was therefore not performed for normal controls.

The typical CMR scan protocol and scanner parameters for the primary and external validation sets are presented in Supplementary Table 7. Extended Data Fig. 2 shows an illustration of cardiac MRIs (SAX cine, 4CH cine and SAX LGE) utilized in model development. Supplementary Videos 1–11 demonstrate example CMR of the 11 types of CVDs.

Annotation procedures

For each patient in the disease cohort, the textual description of the abnormalities in the CMR and the clinical report was extracted as the main reference. Besides that, all CMR records underwent additional annotation procedures. To annotate the disease cohort, a group of certified CMR experts reviewed all records and clinical reports. Every record was randomly assigned to be reviewed by a single physician specifically for this task, not for any other purpose. All annotators received specific instructions and training regarding how to annotate CMR data to improve labeling consistency. The diagnostic criteria we adopted in this study for each CVD class are described in Methods. CMR examinations that could not be interpreted by physicians received further annotation from a consensus committee of board-certified practicing cardiologists (with >15 years of experience in CMR reading) working in Fuwai Hospital. The CMR examinations that could not be interpreted or agreed upon by the committee were removed from our dataset.

For the independent gold-standard test dataset with 500 patients (Extended Data Table 6) for human–machine comparison, six physicians working in the MRI department at Fuwai Hospital contributed directly to its annotation (the six physicians were not involved in dataset annotation as described above). All participating physicians received specific instructions and training regarding how to annotate CMRs to ensure consistency. We divided the physicians into three groups according to their reading experience in CMR: 3–5 years, 5–10 years and more than 10 years. CMR physicians in each group reviewed a randomly selected set of the 500 CMRs in a nonrepetitive manner.

CMR preprocessing

The CMR preprocessing pipeline aimed to remove the additional burden of the deep neural network learning to find patterns between images for disease classification. All cardiac MRIs were preprocessed to (1) resample MRI images to the same spatial resolution and (2) localize the heart region of interest (ROI) to a crop image. We detailed the preprocessing step for cine and LGE MRI below and in Extended Data Fig. 4.

SAX cine comprises nine parallel views (for most cases) covering the apical to the basal levels of the LV. Each view contains 25 frames (cardiac phases), leading to 225 images in one single SAX cine record. We examined the representational power of different numbers of input views in developing the classification model. Balancing efficiency and effectiveness, the three-view input scheme achieved a greater representation of SAX cine and therefore is adopted throughout the rest of the study. The three-view input scheme includes the middle layer (the mid slice among the parallel layers spanning from the base to the apex), the second layer above the middle layer and the second layer below the middle layer (Extended Data Fig. 2). We extract the 'ImagePositionPatient' tag and the 'ImageOrientationPatient' tag from each Dicom header to locate the three layers. Then, three-spline interpolation provided by SimpleITK⁶⁰ library (<https://simpleitk.org/>) is applied to resample the raw cine MRIs to the same spatial resolution of 0.994 mm × 0.994 mm,

which is the most common spatial resolution across all subjects investigated in this study. We developed a heart ROI segmentation model (the following section) and used it to localize the region of heart for each cine MRI. The heart ROI segmentations predicted by the AI models were manually checked to ensure their accuracy. The extracted ROIs are padded to keep the aspect ratio the same without distortion, and then resized to 224×224 . The top and bottom 0.1% of the pixels in cine MRI images are clipped to avoid pixels that are outliers of the distribution. The cine images are scaled between 1 and 255, and then normalized by zero mean and unit variance before feeding them to the model. We sample a clip of 25 frames from each full-length cine sequence using a temporal stride of two, resulting in 13 frames as inputs to model development. The 4CH cine shares the same preprocessing pipeline as SAX cine, except that only one single layer (mid slice) is used to represent the 4CH view. For SAX LGE, all layers covering from the base to the apex of the heart are used for diagnostic model development. The preprocessing steps for SAX LGE are similar to that of cine MRI. We resampled SAX LGE along the z-axis to ensure that each LGE sequence contains nine slices because nine is the most common number of views for SAX LGE included in this study.

Heart ROI extraction

We developed heart detection DNN models to automatically extract the heart ROI regions (Extended Data Fig. 4). Three DNN models for SAX cine, 4CH cine and SAX LGE were trained and evaluated, respectively. We applied nnU-Net⁶¹ as our model backbone and generated the ground-truth segmentation masks for model supervision using a semi-automatic approach. (1) Automatic localization: for SAX cine and 4CH cine, we selected the pixel region with maximum standard deviation across all frames. These regions localize the heart ROI as heart is a beating organ with high standard deviation in its position. Specifically, for each cine movie sequence $s = \{x_1, \dots, x_n\}$, we computed a single pixel map of standard deviations across all frames $x_{\text{std}} = \sigma(\{x_1, \dots, x_n\})$. This map was used to compute an Otsu threshold to binarize and label regions with the greatest variation in cine modality²¹. For each cine sequence, a binary segmentation mask of the heart ROI is defined for the length of the cardiac cycle. All segmentation masks went through manual checking. The localization procedure captures the heart ROI in around 90% of cases. The rest of the cases are labeled manually. (2) Manual labeling: we manually drew the bounding box capturing the heart ROI, using 3D Slicer⁶² and ITK-SNAP⁶³. We used the Scissors tool provided by the Segment Editor in 3D Slicer and the Polygon Inspector in ITK-SNAP to locate heart ROI. A binary segmentation mask was saved for each CMR sequence. For SAX LGE, we manually drew the annotations as model supervision.

In terms of model architecture, the detection model shares the classic U-net⁶⁴ backbone with three small adjustments: (1) batch normalization is replaced with instance normalization⁶⁵, (2) rectified linear unit (ReLU) is replaced with leaky ReLU⁶⁶ as the activation function and (3) additional auxiliary losses are added in the decoder to all but the two lowest resolutions. The model outputs the binary bounding box that extracts the heart ROI. For model training, we adopted Adam optimizer and stochastic gradient descent (SGD) with Nesterov momentum ($\mu = 0.99$). The initial learning rate was set to be 0.01, and the decay of the learning rate followed the 'Poly' learning rate policy⁶⁷. Batch size was set to 36. Data augmentation included rotations, scaling, gamma correction and mirroring. The loss function is the sum of cross-entropy and Dice loss⁶⁸.

Video-based deep learning models and training details

Model architecture. For models based on cine sequence, we sampled a clip of 13 frames from each 25-frame cine video using a temporal stride of 2 and spatial size of 224×224 , resulting in $7 \times 56 \times 56$ input 3D tokens. The 3D patch partitioning layer obtains tokens, with each patch/token consisting of a 128-dimensional feature. In practice, 3D

convolution without overlapping is applied for this tokenization, and the number of output channels is set to be 128 to project the features of each token to a 128-dimension.

The developed model consists of four stages, that is, four video swin transformer blocks. Each stage, besides the last stage, performs $2 \times$ spatial downsampling in the patch merging layer. It is worth noting that we do not downsample along the temporal dimension. The patch merging layer concatenates the features of each group of 2×2 spatially neighboring patches and applies a linear layer to project the concatenated features to half of their dimension. The video swin transformer block consists of a 3D window-based multihead self-attention module and a 3D-shifted window-based multihead self-attention module, followed by a feedforward network, that is, a two-layer multilayer perceptron, with Gaussian error linear unit nonlinearity in between. Layer normalization is applied before each multihead self-attention module and multilayer perceptron, and a residual connection is applied after each module. We used the base version of VST. The number of heads for each stage is 4, 8, 16 and 32. Extended Data Fig. 3a shows the schematic overview of the VST-based framework for modeling SAX cine.

Data augmentation. Model performance improved with increasing training data sample size. For the screening model, we used random rotation, random color jitter and adding random number. During each step of SGD in the training process, we perturbed each training sample, cine video sequences, with a random rotation (between -45 and $+45$ degrees for SAX cine and between -20 to $+20$ degrees for 4CH cine), random color jitter and with adding a number sampled uniformly between -0.1 and 0.1 to image pixels (pixel values are normalized) to increase or decrease brightness of the images. For LGE, we used random rotation between -45 and $+45$ degrees, random color jitter and random flip along the z-axis. Data augmentation resulted in improvement for all models.

Multimodality fusion. First, we developed VST-based models for SAX cine, 4CH cine and SAX LGE, respectively. Then, to fuse information from different modalities, we added a global average pooling layer following the last self-attention module for each VST model. This resulted in a 1,024-dimension feature vector from each modality. We further concatenated the 1,024-dimension vectors and added a fully connected layer on top of that to aggregate the features. The final fully connected softmax layer produces a distribution over the output classes. In terms of training, we loaded and froze the pretrained weights of each VST branch from different modalities using transfer learning⁶⁹ and only finetuned the last fully connected layers for feature aggregation.

Implementation details. Following the classic VST configuration²⁷, we employed an AdamW optimizer using a cosine decay learning rate scheduler and 2.5 epochs of linear warmup. A batch size of 32 was used. The backbone VST is initialized from the ImageNet⁷⁰ and Kinetics-600 (ref. 71) pretrained model; the head is randomly initialized. Model pre-training plays a strikingly important role in VST-based CMR interpretation. We also found that multiplying the learning rate of the backbone by 0.1 improves performance. Specifically, the initial learning rates for the pretrained backbone and randomly initialized head were set to be 1×10^{-4} and 1×10^{-3} , respectively. The impact of learning rate modification on the VST backbone was systematically examined as below. We adopt 0.2 stochastic depth rate and 0.05 weight decay for the Swin base model used in this study. To prevent the models from becoming biased toward one class, we balanced the training datasets for both screening and diagnostics using the ClassBalancedDataset sampling strategy⁷². Each VST branch derived from the single modality was trained for 150 epochs and then fed into the fusion model, following with 20 epochs of finetuning particularly for the fusion layers. For inference, we set the batch size to be one and the number of workers to be four. The training time for model development using four NVIDIA GeForce RTX

3090 graphics processing units with 24 GB VRAM was about 77 h, and the inference time for each subject was only 0.233 s.

Learning rate on the VST backbone. The impact of learning rate modification on the VST backbone was systematically examined through a controlled experiment. The experiment encompassed a range of learning rates, from 1×10^{-2} to 1×10^{-6} , with a focus on their effects on the AI diagnostic model based on SAX cine. The investigation was conducted on the primary cohort (6,650 CVD patients), utilizing a twofold configuration for training and the remaining fold for testing. The model was trained for 150 epochs with five different learning rate initializations for the model backbone: 1×10^{-2} , 1×10^{-3} , 1×10^{-4} (as applied in this study), 1×10^{-5} and 1×10^{-6} . Other configurations were kept consistent for a fair and direct comparison, and the training loss for each scheme was plotted for analysis (Supplementary Fig. 3). From the depicted figure, several key observations emerge. When the learning rate is set too high (1×10^{-2} , curve in blue color), the model struggles to converge and the training loss fails to descend, in stark contrast to the more optimal setting of 1×10^{-4} (curve in green color). Notably, the model under the 1×10^{-2} learning rate incorrectly classified all samples into the HCM class during testing. Conversely, when the learning rate is set too low (1×10^{-6} , curve in purple color), the loss descends very slowly over the training period. As depicted in the figure, the loss curves for 1×10^{-5} and 1×10^{-6} remain at a relatively high level compared with the more effective setting of 1×10^{-4} . Further evaluation included the calculation of F_1 and area under the receiver operating characteristic curve scores for the testing fold under the aforementioned experimental settings (Supplementary Fig. 3). Notably, the model trained with a learning rate of 1×10^{-2} failed to converge and was consequently excluded from the quantitative metrics. According to the evaluation results, the initialized learning rate of 1×10^{-4} demonstrated superior performance compared with the other settings. Therefore, based on these comprehensive analyses, we selected 1×10^{-4} as the initialized learning rate for our experiment.

CNN-LSTM. We examined the conventional CNN-LSTM architecture in CMR interpretation. The CNN-LSTM consists of a DenseNet encoder with 40 layers and a growth rate of 12 for feature extraction and an LSTM for temporal feature aggregation. DenseNet encoder comprised a series of two-dimensional convolutions with kernel sizes 1×1 and 3×3 and global average pooling to extract the feature vector for each input frame. For LSTM, the feature vector for each input frame is fed into the LSTM module sequentially. LSTM fuses the feature vectors and produces the final classification score after one fully connected layer. For the training configuration of the CNN-LSTM model, we adopt the SGD optimizer with a learning rate of 0.001, a momentum of 0.9 and a weight decay of 0.001. A batch size of four is used for training and one is used for testing. The DenseNet encoder of the CNN-LSTM model is initialized from the pretrained model²¹ and the LSTM component is randomly initialized. We kept data augmentation, the input scheme and computational resources the same as VST models with the only difference: SAX cine inputs are resized to 64×64 due to CNN-LSTM memory constraints.

Quantitative assessment and statistical analysis

The performance of the AI models was evaluated by assessing their sensitivity, specificity, precision and F_1 score (harmonic mean of the predictive positive value and sensitivity), with two-sided 95% CIs, as well as the AUC of the ROC with two-sided CIs. The F_1 score is complementary to the AUC, which is particularly useful in the setting of multiclass prediction and less sensitive than the AUC in settings of class imbalance. For an aggregate measure of model performance, we computed the class frequency-weighted mean for the F_1 score and the AUC⁷³.

The cutoff value was set to 0.5 for screening; the CVD class with the highest probability was the diagnostic prediction. Precision,

sensitivity (recall), specificity, PPV, NPV and F_1 score of each class are related to true-positive (TP), true-negative (TN), false-positive (FP) and false-negative (FN) rates, with formulas as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}},$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}},$$

$$F_1\text{-score} = \frac{2 \times \text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}.$$

The ROC space is defined by $1 - \text{specificity}$ and sensitivity as the x axis and the y axis, respectively. It depicts relative trade-offs between true positive and false positive, as the classification threshold goes from zero to one. A random guess will give a point along the diagonal line from the bottom left to the top right. Points above the diagonal line represent good classification results and points below the line represent bad results. We applied the class frequency-weighted F_1 score and class frequency-weighted AUC to evaluate the performance of our diagnostic model, with the following formulas:

$$\text{Weighted } F_1\text{-score} = \sum_i^c \text{ratio}_i F_{1\text{-score}_i},$$

$$\text{Weighted AUC} = \sum_i^c \text{ratio}_i \text{AUC}_i,$$

where $F_{1\text{-score}_i}$ and AUC_i denote the F_1 score and AUC for class i , respectively, and ratio_i denotes a frequency ratio for each class i .

In addition, to improve the model interpretability and visualize the features used by the DNN model that determine the final prediction, we used Grad-CAM²⁹ to localize important regions—saliency regions—by visualizing class-specific gradient information. In Grad-CAM, the neuron importance weight α_k^c is estimated as

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k},$$

where y^c denotes the gradient score for class c before the softmax and A^k denotes the feature map activation of the k th layer. After computing the neuron importance weights for each feature map, we can generate a heat map indicating the significant regions related to class c by performing a weighted linear combination of the feature maps, followed with a ReLU activation function as

$$L_{\text{Grad-CAM}}^c = \text{ReLU} \left(\sum_k \alpha_k^c A^k \right).$$

We then used the Shapley values⁷⁴ to evaluate the influence of each input modality (SAX cine, 4CH cine and SAX LGE). The Shapley value is a principled attribution method used in AI to quantify the contribution of individual input features by assigning each input modality an

importance value for a particular prediction. The definition of the Shapley value⁷⁵ is given in equations below:

$$\phi_i(v) = \sum_{S \subset M \setminus \{i\}} \binom{n}{1, |S|, n - |S| - 1}^{-1} (v(S \cup \{i\}) - v(S)),$$

where $\phi_i(v)$ denotes the contribution value of input component i , namely the Shapley value of each input modality (player), N is the number of layers and v is a function mapping subsets of layers to the real numbers: $v : 2^N \rightarrow \mathbb{R}$, with $v(\emptyset) = 0$, where \emptyset denotes the empty set. A set of players is called a coalition. The function v is called a characteristic function: if S is a coalition of players, then $v(S)$, called the worth of coalition S , describes the total expected sum of payoffs the members of S can obtain by cooperation. The sum extends over all subsets S of N not containing input component i ; also note that $\binom{n}{a, b, c}$ is the multinomial coefficient. This formula can also be interpreted as

$$\phi_i(v) = \frac{1}{\text{Number of layers}} \sum_{\text{coalitions including } i} \frac{\text{Marginal contribution of } i \text{ to coalition}}{\text{Number of coalitions excluding } i \text{ of this size}}.$$

Diagnostic criteria of the CVDs and normal control

CAD or ischemic cardiomyopathy. The diagnosis of myocardial infarction or ischemic cardiomyopathy is based on the European Society of Cardiology, American College of Cardiology and American Heart Association committee criteria⁷⁶ with significant stenosis on invasive coronary angiography (CAG) or coronary computed tomography angiography, and CMR showed subendocardial or transmural LGE with matching coronary arteries. We excluded cases without available CAG present or inadequate image quality due to arrhythmia or respiratory motion artifact.

HCM. We followed the 2020 American Heart Association and American College of Cardiology guidelines for the diagnosis of patients with HCM⁷⁷. The clinical diagnosis of HCM was made by CMR showing a maximal end-diastolic wall thickness of ≥ 15 mm anywhere in the LV, in the absence of another cause of hypertrophy in adults. More limited hypertrophy (13–14 mm) can be diagnostic when present in family members of a patient with HCM or in conjunction with a positive genetic test.

We excluded cases with the following conditions:

1. Valvular heart disease (aortic valve stenosis, etc.)
2. Long-term uncontrolled hypertension
3. Inflammatory heart disease (sarcoidosis, etc.)
4. Infiltrative cardiomyopathy (amyloidosis, Fabry disease, etc.)
5. Septal myectomy or alcohol ablation before CMR
6. CMR images with poor quality

DCM. The diagnosis of DCM is based on the diagnostic criteria of the World Health Organization⁷⁸. Inclusion criteria were based on enlarged LV end-diastolic dimension (>60 mm) and reduced LVEF ($<45\%$). The exclusion criteria were as follows:

1. Significant stenosis of coronary artery ($>50\%$ stenosis, assessed on CAG or coronary computed tomography angiography)
2. Severe valvular disease, hypertension or congenital heart disease
3. Evidence of acute or subacute myocarditis (T2 weighted image and laboratory tests)
4. Any other metabolic disease through medical documentation
5. Inadequate CMR quality

LVNC. The diagnosis of LVNC is based on previous studies^{32,79}, as follows:

1. The presence of noncompacted and compacted LV myocardium with a two-layered appearance, with at least involvement of the LV apex
2. End-diastolic noncompaction/compaction ratio >2.3 on long-axis views and ≥ 3 on SAX views
3. Noncompacted mass $>20\%$ of the global LV mass
4. No pathologic (pressure/volume load, for example, hypertension) or physiologic (for example, pregnancy and vigorous physical activity) remodeling factors leading to excessive trabeculation

ARVC. The diagnostic standards for ARVC were based on the revised Task Force Criteria⁸⁰ score with either two major criteria, one major and two minor criteria or four minor criteria. The major criteria include regional RV akinesia or dyskinesia or dyssynchronous RV contraction, ratio of RV end-diastolic volume to body surface area >110 ml m⁻² (male) or >100 ml m⁻² (female) or RV ejection fraction $<40\%$; fibrous replacement of the RV free wall myocardium, with or without fatty replacement of tissue on endomyocardial biopsy; repolarization abnormalities and depolarization or conduction abnormalities on ECG test.

CAM. The diagnosis of CAM is based on endomyocardial biopsy or extracardiac biopsy specimens showing positive birefringence with Congo red staining under polarized light, and with native and enhanced CMR imaging in a pattern consistent with CAM: LV wall thickness of more than 12 mm shown by CMR without other known cause, with and without diffuse LGE⁸¹.

RCM. RCM is characterized by ventricular filling difficulties with increased stiffness of the myocardium. The restrictive cardiomyopathies are defined as restrictive ventricular physiology in the presence of normal or reduced diastolic volumes^{52,82}, as follows:

1. Nondilated LV or RV with diastolic dysfunction
2. Bi-atrial dilation
3. Preserved ejection fraction (LVEF $\geq 50\%$)

We excluded subjects that met the following criteria:

1. With a reduced LV systolic function
2. Severe atrial fibrillation
3. Severe valvular disease, hypertension or congenital heart disease
4. Significant stenosis of coronary artery.

PAH. The diagnosis of PAH is based on the results of right heart catheterization examination. Patients are included in this study if they were clinically diagnosed as PAH⁸³:

1. Mean pulmonary artery pressure (mPAP) ≥ 25 mmHg
2. Pulmonary capillary wedge pressure (PCWP) <15 mmHg
3. Pulmonary vascular resistance (PVR) >3 Wood units at rest

We excluded subjects with the following criteria:

1. Any evidence of cardiomyopathy, myocarditis, CAD, myocardial infarction, valvular disease, or constrictive pericarditis.
2. Any evidence of respiratory diseases.
3. History of cardiac surgery

Congenital heart disease—Ebstein's anomaly. The diagnosis of Ebstein's anomaly is based on apical displacement of tricuspid valve leaflets (≥ 8 mm m⁻²) with fibrous and muscular attachments to the underlying myocardium³¹. Patients with other concomitant malformation (for example, congenitally corrected transposition with Ebstein's anomaly) and history of cardiac surgery were excluded.

Acute myocarditis. The diagnosis of acute myocarditis is based on the diagnostic criteria for clinically suspected myocarditis, as recommended by the European Society of Cardiology Working Group on Myocardial and Pericardial Diseases⁸⁴, and is fulfilled by meeting the Lake Louise criteria⁸⁵ or by confirmation through endomyocardial biopsy.

Patients with clinically acute myocarditis had the following: acute chest pain, signs of acute myocardial injury (electrocardiographic changes and/or elevated troponin level) and increased laboratory markers of inflammation (for example, C-reactive protein level). CAD was excluded before cardiac MRI. Patients with preexisting CVD were excluded.

HHD. The diagnostic criteria for HHD include (1) a history of prolonged, uncontrolled arterial hypertension and (2) concentric hypertrophy with left ventricular maximal wall thickness ≥ 12 mm.

We excluded patients with the following conditions:

1. Any other causes of LV hypertrophy
2. Cardiomyopathy
3. Obstructive coronary heart disease
4. Severe valvular disease
5. Inflammatory heart disease
6. Severe ventricular arrhythmia such as ventricular tachycardia or left bundle branch block
7. Poor CMR imaging quality

Normal controls. Healthy controls were recruited as volunteers without CVDs (including cardiomyopathy, CAD, severe arrhythmia or conduction block, valvular disease, congenital heart disease and so on) and other organic or systemic diseases on the comprehensive evaluation by patient history, clinical assessment, ECG and echocardiography.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

IRB approval was obtained from all participating institutions for imaging and data collection: Beijing Fuwai Hospital, China (2023–1935). The need for informed consent was waived by the respective ethics committees and institutions. No publicly available datasets were used in this study. The deidentified data can be shared only for noncommercial academic purposes and will require a formal material transfer agreement and a data use agreement. Requests should be submitted by emailing the corresponding authors (S.Z., Y.-R.J.W. or K.Z.) at cjrzhao@shihua2009@163.com, wangyanran100@gmail.com or kk.zhao@siat.ac.cn. All requests will be evaluated based on institutional policies to determine whether the data requested are subject to intellectual property or patient privacy obligations. Generally, all such requests for access to CMR data will be responded to within 1 month. Example CMR data in this study are available in supplementary videos.

Code availability

An open-source version of the code base is available on GitHub at <https://github.com/MedAI-Vision/CMR-AI> with no restrictions.

References

56. Cohen, I. G. & Mello, M. M. HIPAA and protecting health information in the 21st century. *JAMA* **320**, 231–232 (2018).
57. Treibel, T., White, S. & Moon, J. Myocardial tissue characterization: histological and pathophysiological correlation. *Curr. Cardiovasc. Imaging Rep.* **7**, 1–9 (2014).
58. Nakamori, S. & Dohi, K. Myocardial tissue imaging with cardiovascular magnetic resonance. *J. Cardiol.* **80**, 377–385 (2022).
59. Paiman, E. H. & Lamb, H. J. When should we use contrast material in cardiac MRI? *J. Magn. Reson. Imaging* **46**, 1551–1572 (2017).
60. Lowekamp, B. C., Chen, D. T., Ibáñez, L. & Blezek, D. The design of SimpleTK. *Front. Neuroinform.* **7**, 45–58 (2013).
61. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2021).
62. Fedorov, A. et al. 3D Slicer as an image computing platform for the quantitative imaging network. *Magn. Reson. Imaging* **30**, 1323–1341 (2012).
63. Yushkevich, P. A., Gao, Y. & Gerig, G. ITK-SNAP: an interactive tool for semi-automatic segmentation of multi-modality biomedical images. in *Proc. 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 3342–3345 (IEEE, 2016).
64. Ronneberger, O., Fischer, P. & Brox, T. U-net: convolutional networks for biomedical image segmentation. in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention* 234–241 (2015).
65. Huang, X. & Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. in *Proc. IEEE International Conference on Computer Vision* 1501–1510 (2017).
66. Maas, A. L., Hannun, A. Y. & Ng, A. Y. Rectifier nonlinearities improve neural network acoustic models. in *Proc. ICML 30* (Citeseer, 2013).
67. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 834–848 (2017).
68. Drodzdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S. & Pal, C. in *Deep Learning and Data Labeling for Medical Applications* 179–187 (Springer, 2016).
69. Zhuang, F. et al. A comprehensive survey on transfer learning. *Proc. IEEE* **109**, 43–76 (2020).
70. Deng, J., et al. Imagenet: a large-scale hierarchical image database. in *Proc. 2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
71. Carreira, J. & Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. in *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 6299–6308 (2017).
72. Gupta, A., Dollar, P. & Girshick, R. Lvis: a dataset for large vocabulary instance segmentation. in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 5356–5364 (2019).
73. Hannun, A. Y. et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* **25**, 65–69 (2019).
74. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**, 4765–4774 (2017).
75. Ichiishi, T. *Game Theory for Economic Analysis* (Elsevier, 2014).
76. Thygesen, K. et al. Fourth universal definition of myocardial infarction. *Circulation* **138**, e618–e651 (2018).
77. Ommen, S. R. et al. 2020 AHA/ACC guideline for the diagnosis and treatment of patients with hypertrophic cardiomyopathy: executive summary: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *J. Am. Coll. Cardiol.* **76**, 3022–3055 (2020).
78. Richardson, P. et al. Report of the 1995 World Health Organization/International Society and Federation of Cardiology Task Force on the definition and classification of cardiomyopathies. *Circulation* **93**, 841–842 (1996).
79. Yu, S. et al. Correlation between left ventricular fractal dimension and impaired strain assessed by cardiac MRI feature tracking in patients with left ventricular noncompaction and normal left ventricular ejection fraction. *Eur. Radiol.* **32**, 2594–2603 (2022).
80. Marcus, F. I. et al. Diagnosis of arrhythmogenic right ventricular cardiomyopathy/dysplasia: proposed modification of the task force criteria. *Circulation* **121**, 1533–1541 (2010).

81. Gertz, M. A. et al. Definition of organ involvement and treatment response in primary systemic amyloidosis (AL): a consensus opinion from the 10th international symposium on amyloid and amyloidosis. *Am. J. Hematol.* **104**, 754 (2004).
82. Amaki, M. et al. Diagnostic concordance of echocardiography and cardiac magnetic resonance-based tissue tracking for differentiating constrictive pericarditis from restrictive cardiomyopathy. *Circ. Cardiovasc. Imaging* **7**, 819–827 (2014).
83. Callan, P. & Clark, A. L. J. H. Right heart catheterisation: indications and interpretation. *Heart* **102**, 147–157 (2016).
84. Caforio, A. L. et al. Current state of knowledge on aetiology, diagnosis, management, and therapy of myocarditis: a position statement of the European Society of Cardiology Working Group on Myocardial and Pericardial Diseases. *Eur. Heart J.* **34**, 2636–2648 (2013).
85. Ferreira, V. M. et al. Cardiovascular magnetic resonance in nonischemic myocardial inflammation: expert recommendations. *J. Am. Coll. Cardiol.* **72**, 3158–3176 (2018).

Acknowledgements

S.Z. expresses gratitude for the support received from the National Key R&D Program of China (nos. 2021YFF0501400 and 2021YFF0501404). Thanks are extended to the primary researchers (Y.-R.J.W., H.Z., Y.H., P.W., Yi Wen and J.C.W.) for their unwavering and sustained voluntary dedication to this project since its initiation, all carried out without any funding support. The icons depicting the heart and syringe in Fig. 1 and Extended Data Fig. 2 were created with [BioRender.com](https://www.biorender.com).

Author contributions

Y.-R.J.W. and J.C.W. conceptualized the study. Y.-R.J.W., P.W., Y.H., Y. Lai, Yi Wen, Yufeng Wang and S.T. conducted data processing,

developed and trained the deep learning algorithms, and performed the model analysis. S.Z., K.Y., K.Z., X.C., S.Y. and Z.D. collected and quality controlled the data and conducted clinical evaluations of the model's performance. Yining Wang, H.L., L.Z., L.H., Y. Li, L.W. and Z.C. facilitated external data acquisition. Y.H., P.W., Y. Lai, Yufeng Wang and Yi Wen contributed equally as co-second authors. Y.-R.J.W. drafted the primary paper and figures with input from all authors. S.Z., J.C.W. and Y.-R.J.W. jointly supervised the project. All authors provided feedback and ensured the integrity of the work.

Competing interests

The authors declare no competing interests.

Additional information

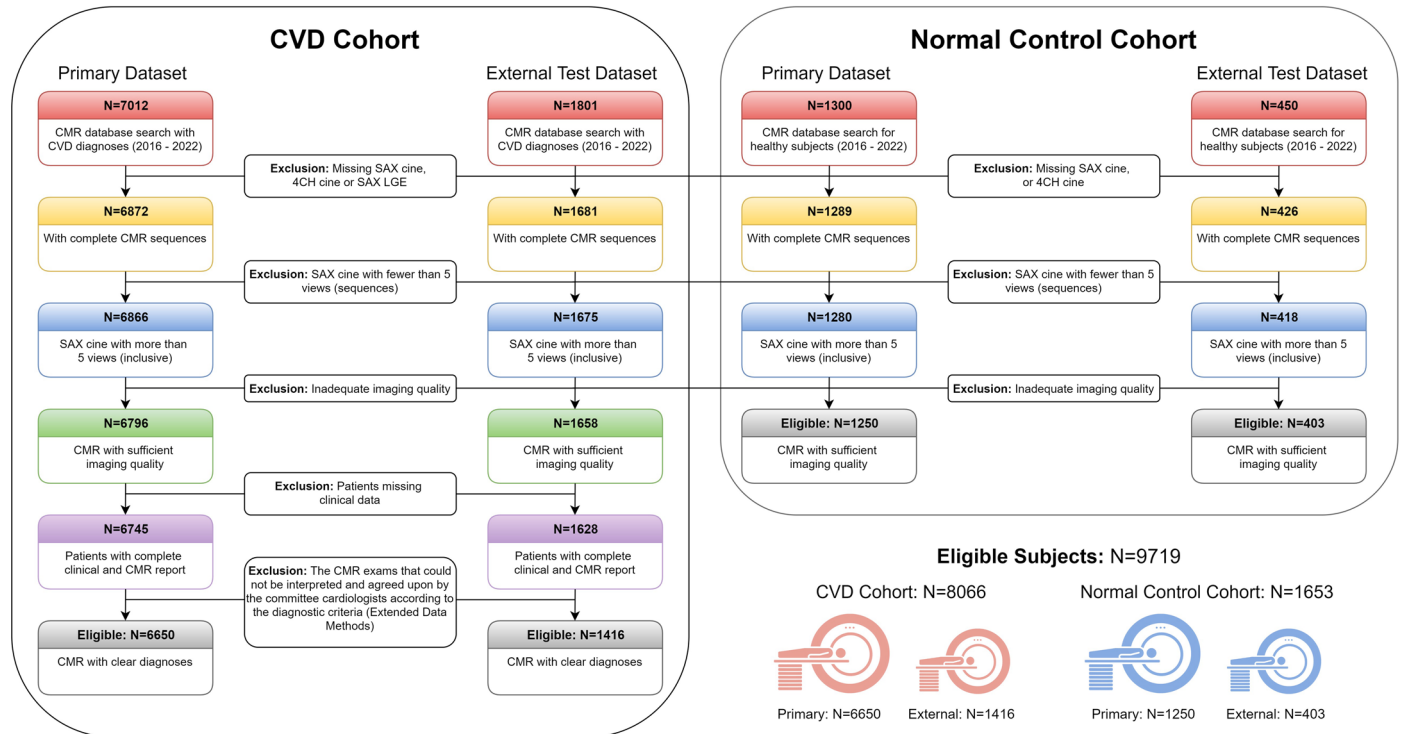
Extended data is available for this paper at <https://doi.org/10.1038/s41591-024-02971-2>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41591-024-02971-2>.

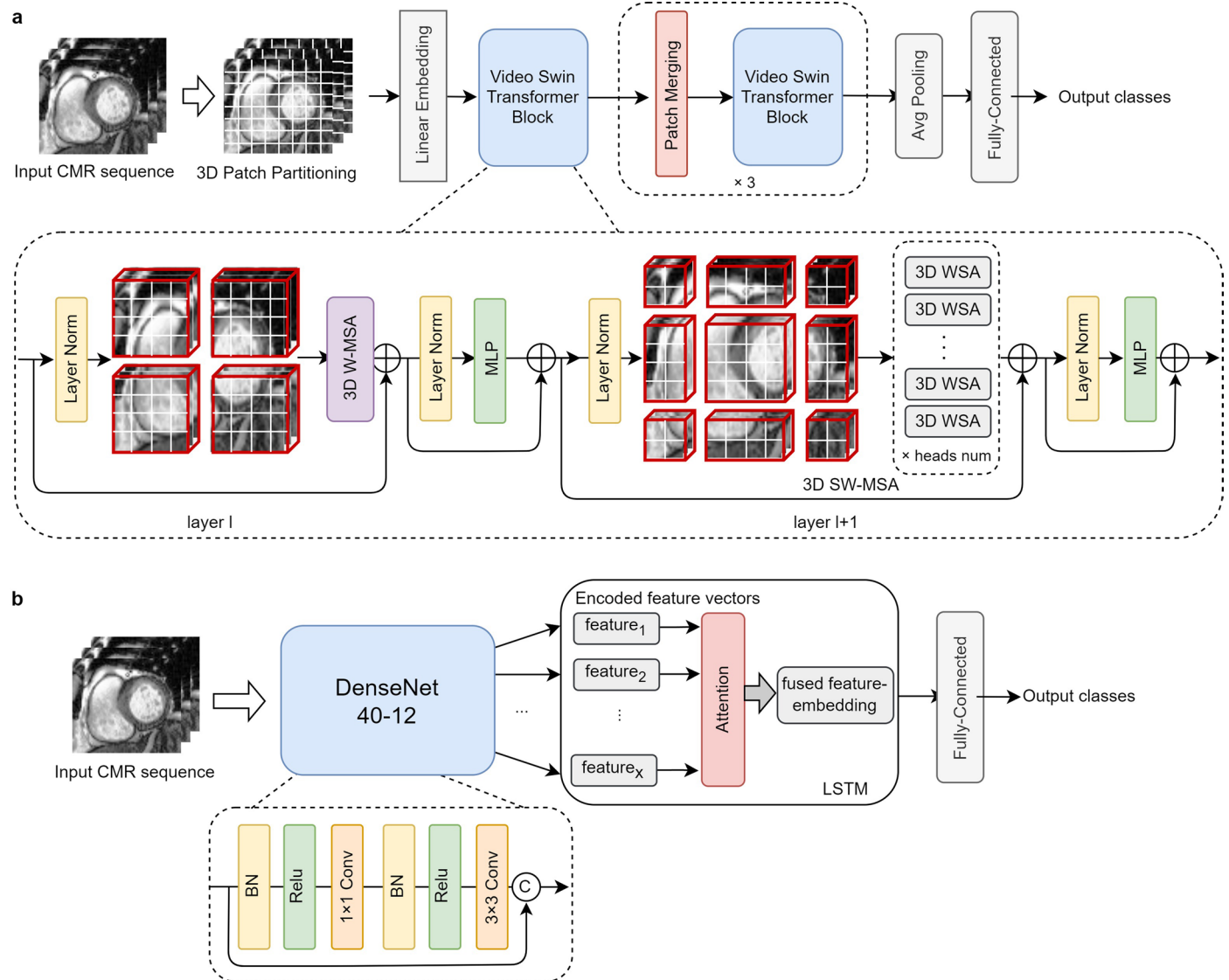
Correspondence and requests for materials should be addressed to Yan-Ran (Joyce) Wang, Kankan Zhao or Shihua Zhao.

Peer review information *Nature Medicine* thanks Tim Leiner, Declan O'Regan and Andrew Swift for their contribution to the peer review of this work. Primary Handling Editor: Michael Basson, in collaboration with the *Nature Medicine* team

Reprints and permissions information is available at www.nature.com/reprints.

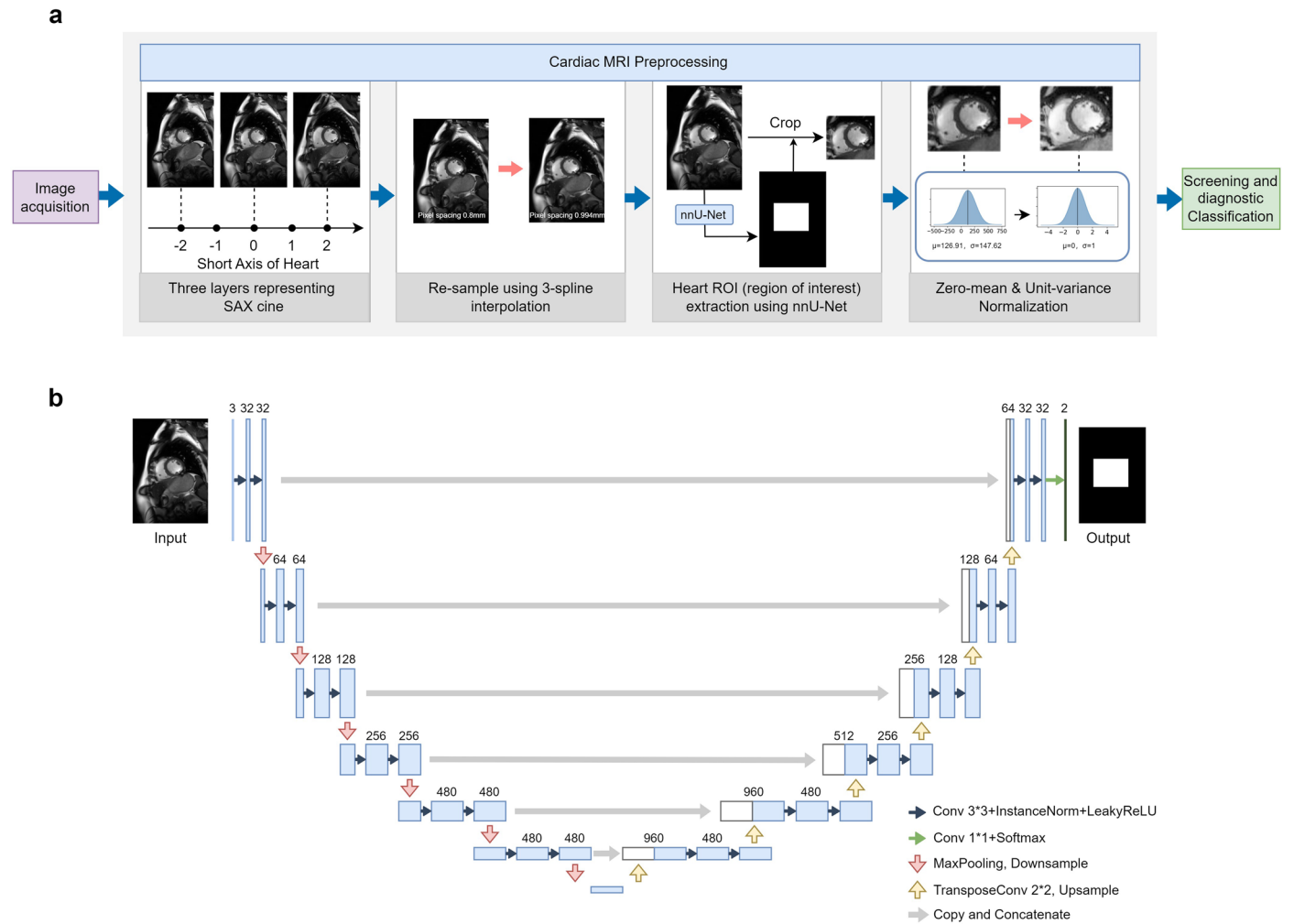


Extended Data Fig. 1 | Inclusion-exclusion cascade for the cardiovascular disease and the normal control cohorts. CVD, cardiovascular disease; CMR, cardiovascular magnetic resonance imaging; SAX, short-axis; 4CH, four-chamber; LGE, late gadolinium enhancement.



Extended Data Fig. 3 | Schematic overview of Video-based Swin Transformer and CNN-LSTM frameworks for CMR interpretation. **a**, the framework of video-based swin transformer (VST). The model takes cardiac MRI sequence as input, extracts distinct features by global self-attention and shifted window mechanism inherent in VST, and outputs the classification score. 3D W-MSA, 3D window based multi-head self-attention module; 3D SW-MSA, 3D-shifted window

based multi-head self-attention module; WSA, window self-attention module. **b**, the framework of the conventional CNN-LSTM (Long short-term memory). Each CMR slice is encoded by the DenseNet-40-12 (layer = 40; growth rate = 12) CNN into a feature vector $feature_i$. These feature vectors are sequentially fed into the LSTM encoder, which uses a soft attention layer to learn a weighted average embedding of all slices.



Extended Data Fig. 4 | The preprocessing pipeline and the framework of heart region of interest (ROI) automated detection. a, the preprocessing pipeline for SAX cine. **b**, the framework of ROI detection for SAX cine. The model takes a single cine sequence as the input and outputs the bounding box (binary mask) covering the heart ROI. SAX, short-axis; ROI, region of interest.

Extended Data Table 1 | The distribution of the external validation dataset pooled from seven medical centers across China

External Test Dataset	Total Numbers	Medical Centers							
		AZ	GD	HEB	LZ	RJ	TJ	XH	
Total	1819	574	205	77	428	76	225	234	
Normal Controls	403	217	10	29	79	0	19	49	
Cardiovascular Diseases	1416	357	195	48	349	76	206	185	
1 HCM	388	117	49	1	131	3	35	52	
2 DCM	204	40	16	21	37	9	25	56	
3 CAD	299	53	56	3	91	0	58	38	
4 LVNC	30	12	1	2	8	1	3	3	
5 RCM	22	7	3	1	5	2	3	1	
6 CAM	138	29	25	7	11	25	34	7	
7 HHD	107	26	8	7	30	7	12	17	
8 Myocarditis	66	21	11	2	16	4	9	3	
9 ARVC	54	17	8	4	15	1	9	0	
10 PAH	66	18	6	0	5	22	7	8	
11 Ebstein's Anomaly	42	17	12	0	0	2	11	0	
Sex	Male	1228 (68%)	379 (66%)	125 (61%)	56 (73%)	308 (72%)	62 (82%)	167 (74%)	129 (55%)
	Female	591 (32%)	195 (34%)	80 (39%)	21 (28%)	120 (28%)	14 (18%)	58 (26%)	105 (45%)
	Age \pm s.d.	47 \pm 16	45 \pm 17	45 \pm 17	50 \pm 14	48 \pm 16	59 \pm 13	45 \pm 17	45 \pm 17
	(Range)	(1-88)	(6-82)	(1-75)	(8-80)	(7-86)	(28-81)	(9-81)	(10-88)

*Percentages may total more than 100% due to rounding. AZ: Beijing Anzhen Hospital, Beijing; GD: Guangdong Provincial People's Hospital, Guangzhou; HEB: The 2nd Affiliated Hospital of Harbin Medical University, Harbin; LZ: The First Hospital of Lanzhou University, Lanzhou; RJ: Renji Hospital, Shanghai; TJ: Tongji hospital, Wuhan; XH: Peking Union Medical College Hospital, Beijing.

Extended Data Table 2 | Performance summary of the screening model for anomaly detection on the primary dataset (n = 7900; three-fold cross validation) and the external dataset (n = 1819) with different CMR input schemes

	SAX cine		4CH cine		SAX + 4CH cine	
	Internal	External	Internal	External	Internal	External
AUROC	0.971 (0.965-0.976)	0.953 (0.942-0.965)	0.974 (0.969-0.979)	0.980 (0.972-0.986)	0.986 (0.984-0.988)	0.990 (0.986-0.992)
PPV	0.976 (0.972-0.980)	0.940 (0.928-0.952)	0.977 (0.974-0.981)	0.953 (0.941-0.964)	0.979 (0.976-0.983)	0.955 (0.944-0.966)
Specificity with sensitivity at 90%	0.975 (0.966-0.983)	0.916 (0.873-0.949)	0.975 (0.966-0.983)	0.940 (0.914-0.964)	0.986 (0.978-0.993)	0.970 (0.950-0.990)
Sensitivity with specificity at 90%	0.956 (0.951-0.963)	0.909 (0.886-0.934)	0.967 (0.962-0.973)	0.941 (0.910-0.962)	0.973 (0.968-0.978)	0.959 (0.936-0.974)
F1-score	0.969 (0.966-0.972)	0.947 (0.939-0.955)	0.974 (0.971-0.977)	0.963 (0.955-0.970)	0.977 (0.974-0.979)	0.970 (0.964-0.977)

AUROC, area under the receiver operating characteristic curve; PPV, positive predictive value (precision); CI, confidence interval; SAX, short axis; 4CH, four chamber.

Extended Data Table 3 | Performance of the diagnostic models with different CMR input schemes over three-fold cross validation of the primary dataset (n = 6650)

Internal Testing		AUROC (95%CI)					F1 score (95%CI)				
		SAX cine	4CH cine	SAX+4CH cine	LGE	cine+LGE	SAX cine	4CH cine	SAX+4CH cine	LGE	cine+LGE
1	HCM	0.990 (0.988-0.992)	0.996 (0.995-0.997)	0.997 (0.996-0.998)	0.994 (0.993-0.996)	0.998 (0.997-0.999)	0.950 (0.944-0.956)	0.966 (0.960-0.971)	0.969 (0.964-0.974)	0.968 (0.962-0.973)	0.975 (0.971-0.980)
2	DCM	0.975 (0.971-0.978)	0.976 (0.973-0.979)	0.979 (0.976-0.982)	0.979 (0.976-0.982)	0.988 (0.986-0.990)	0.849 (0.836-0.863)	0.855 (0.841-0.868)	0.857 (0.843-0.871)	0.859 (0.844-0.871)	0.896 (0.884-0.907)
3	CAD	0.957 (0.950-0.965)	0.963 (0.956-0.969)	0.967 (0.960-0.973)	0.989 (0.986-0.993)	0.991 (0.988-0.994)	0.791 (0.767-0.812)	0.804 (0.783-0.823)	0.812 (0.791-0.831)	0.924 (0.912-0.936)	0.921 (0.908-0.935)
4	LVNC	0.943 (0.928-0.957)	0.961 (0.949-0.972)	0.970 (0.960-0.979)	0.958 (0.942-0.975)	0.978 (0.970-0.986)	0.657 (0.610-0.703)	0.760 (0.719-0.799)	0.784 (0.744-0.821)	0.637 (0.584-0.681)	0.778 (0.739-0.816)
5	RCM	0.959 (0.945-0.970)	0.984 (0.975-0.991)	0.992 (0.987-0.995)	0.978 (0.967-0.989)	0.994 (0.991-0.997)	0.732 (0.696-0.769)	0.865 (0.836-0.893)	0.870 (0.842-0.896)	0.769 (0.733-0.805)	0.873 (0.847-0.900)
6	CAM	0.980 (0.970-0.988)	0.981 (0.969-0.990)	0.986 (0.976-0.994)	0.990 (0.980-0.999)	0.994 (0.988-0.998)	0.787 (0.738-0.829)	0.868 (0.833-0.903)	0.869 (0.836-0.900)	0.884 (0.850-0.914)	0.918 (0.888-0.943)
7	HHD	0.929 (0.915-0.942)	0.947 (0.935-0.958)	0.955 (0.944-0.965)	0.955 (0.940-0.969)	0.967 (0.958-0.976)	0.662 (0.621-0.698)	0.672 (0.634-0.707)	0.676 (0.641-0.714)	0.696 (0.660-0.732)	0.723 (0.684-0.757)
8	Myocarditis	0.940 (0.918-0.960)	0.963 (0.941-0.980)	0.970 (0.951-0.984)	0.964 (0.937-0.992)	0.987 (0.978-0.995)	0.480 (0.375-0.578)	0.576 (0.488-0.651)	0.615 (0.526-0.697)	0.590 (0.503-0.674)	0.724 (0.638-0.795)
9	ARVC	0.965 (0.956-0.973)	0.969 (0.960-0.978)	0.976 (0.967-0.983)	0.968 (0.959-0.976)	0.982 (0.975-0.988)	0.721 (0.681-0.758)	0.775 (0.740-0.809)	0.780 (0.746-0.813)	0.757 (0.717-0.794)	0.816 (0.787-0.846)
10	PAH	0.997 (0.992-1.000)	0.995 (0.990-0.998)	0.998 (0.997-1.000)	0.994 (0.986-1.003)	0.998 (0.995-1.000)	0.939 (0.907-0.968)	0.893 (0.853-0.932)	0.923 (0.888-0.954)	0.951 (0.922-0.977)	0.962 (0.937-0.984)
11	Ebstein's Anomaly	0.976 (0.954-0.993)	0.985 (0.969-0.996)	0.987 (0.968-0.999)	0.992 (0.979-1.005)	0.997 (0.994-1.000)	0.833 (0.766-0.887)	0.824 (0.758-0.880)	0.852 (0.789-0.907)	0.830 (0.761-0.890)	0.892 (0.832-0.935)
Class frequency-weighted average		0.972	0.979	0.983	0.983	0.991	0.838	0.865	0.871	0.875	0.906

AUROC, area under the receiver operating characteristic curve; CMR, cardiac magnetic resonance imaging; SAX, short-axis; 4CH, four chamber; LGE, late gadolinium enhancement. The bold font emphasizes the optimal performance metric among various input schemes.

Extended Data Table 4 | Sensitivity and specificity analysis of the diagnostic model derived from cine and LGE as combined inputs

		Sensitivity (Specificity = 0.9)		Specificity (Sensitivity = 0.9)	
		Internal	External	Internal	External
1	HCM	1.000 (0.999-1.000)	0.979 (0.965-0.993)	0.996 (0.994-0.997)	0.986 (0.973-0.994)
2	DCM	0.982 (0.975-0.989)	0.995 (0.984-1.000)	0.967 (0.961-0.973)	0.988 (0.973-0.999)
3	CAD	0.979 (0.968-0.988)	0.987 (0.970-0.997)	0.991 (0.987-0.995)	0.987 (0.977-0.996)
4	LVNC	0.938 (0.908-0.964)	1.000 (1.000-1.000)	0.948 (0.906-0.980)	1.000 (0.993-1.000)
5	RCM	0.986 (0.972-0.997)	1.000 (1.000-1.000)	0.991 (0.986-0.994)	0.994 (0.927-0.999)
6	CAM	0.973 (0.950-0.992)	0.978 (0.951-1.000)	0.998 (0.996-0.999)	0.986 (0.968-1.000)
7	HHD	0.920 (0.892-0.945)	0.953 (0.906-0.991)	0.923 (0.883-0.949)	0.934 (0.908-0.955)
8	Myocarditis	0.966 (0.923-1.000)	0.955 (0.885-1.000)	0.971 (0.922-0.993)	0.955 (0.885-1.000)
9	ARVC	0.949 (0.924-0.971)	0.981 (0.944-1.000)	0.963 (0.940-0.977)	0.992 (0.981-1.000)
10	PAH	0.993 (0.976-1.000)	1.000 (1.000-1.000)	1.000 (0.999-1.000)	1.000 (1.000-1.000)
11	Ebstein's Anomaly	0.989 (0.962-1.000)	1.000 (1.000-1.000)	0.999 (0.995-1.000)	1.000 (1.000-1.000)

*95% confidence interval in the brackets. HCM, hypertrophic cardiomyopathy; DCM, dilated cardiomyopathy; CAD, coronary artery disease; LVNC, left ventricular non-compaction; RCM, restrictive cardiomyopathy; CAM, cardiac amyloidosis; HHD, hypertensive heart disease; ARVC, arrhythmogenic right ventricular cardiomyopathy; PAH, pulmonary arterial hypertension.

Extended Data Table 5 | Performance of the diagnostic models with different CMR input schemes over the external test dataset (n = 1416)

External Testing		AUROC (95%CI)					F1 score (95%CI)				
		SAX cine	4CH cine	SAX+4CH cine	LGE	cine+LGE	SAX cine	4CH cine	SAX+4CH cine	LGE	cine+LGE
1	HCM	0.972 (0.961-0.980)	0.976 (0.966-0.984)	0.979 (0.971-0.987)	0.981 (0.974-0.988)	0.991 (0.986-0.995)	0.865 (0.840-0.889)	0.880 (0.855-0.902)	0.894 (0.870-0.914)	0.898 (0.872-0.919)	0.944 (0.928-0.960)
2	DCM	0.985 (0.979-0.991)	0.978 (0.969-0.985)	0.987 (0.981-0.992)	0.968 (0.958-0.977)	0.995 (0.992-0.998)	0.860 (0.821-0.892)	0.834 (0.795-0.872)	0.878 (0.844-0.910)	0.704 (0.658-0.747)	0.856 (0.821-0.887)
3	CAD	0.952 (0.940-0.963)	0.960 (0.949-0.970)	0.967 (0.955-0.977)	0.973 (0.962-0.983)	0.991 (0.984-0.996)	0.783 (0.743-0.819)	0.814 (0.775-0.848)	0.837 (0.804-0.868)	0.832 (0.794-0.865)	0.909 (0.882-0.932)
4	LVNC	0.962 (0.906-0.994)	0.994 (0.987-0.998)	0.997 (0.995-0.999)	0.962 (0.923-0.990)	1.000 (0.999-1.000)	0.638 (0.491-0.772)	0.691 (0.533-0.821)	0.814 (0.690-0.913)	0.512 (0.293-0.667)	0.824 (0.683-0.931)
5	RCM	0.951 (0.903-0.987)	0.997 (0.994-0.999)	0.997 (0.995-0.999)	0.914 (0.840-0.967)	0.995 (0.988-0.999)	0.433 (0.250-0.586)	0.667 (0.519-0.789)	0.688 (0.548-0.813)	0.333 (0.121-0.536)	0.737 (0.583-0.852)
6	CAM	0.951 (0.927-0.973)	0.973 (0.957-0.986)	0.977 (0.964-0.989)	0.977 (0.964-0.989)	0.992 (0.986-0.997)	0.782 (0.727-0.839)	0.852 (0.803-0.897)	0.859 (0.810-0.904)	0.827 (0.773-0.872)	0.915 (0.877-0.949)
7	HHD	0.927 (0.898-0.953)	0.926 (0.894-0.955)	0.937 (0.909-0.963)	0.917 (0.878-0.951)	0.972 (0.959-0.983)	0.687 (0.608-0.759)	0.690 (0.616-0.759)	0.694 (0.624-0.761)	0.654 (0.571-0.725)	0.718 (0.644-0.789)
8	Myocarditis	0.917 (0.876-0.950)	0.913 (0.871-0.948)	0.943 (0.909-0.971)	0.951 (0.921-0.974)	0.972 (0.950-0.989)	0.438 (0.316-0.547)	0.391 (0.259-0.519)	0.458 (0.327-0.574)	0.605 (0.492-0.697)	0.630 (0.514-0.735)
9	ARVC	0.985 (0.972-0.994)	0.973 (0.951-0.990)	0.982 (0.951-0.997)	0.965 (0.940-0.984)	0.996 (0.992-0.999)	0.723 (0.625-0.800)	0.706 (0.607-0.793)	0.748 (0.649-0.830)	0.729 (0.621-0.818)	0.887 (0.813-0.948)
10	PAH	0.999 (0.998-1.000)	0.969 (0.941-0.991)	0.999 (0.999-1.000)	0.993 (0.985-0.999)	1.000 (1.000-1.000)	0.930 (0.880-0.971)	0.814 (0.731-0.887)	0.900 (0.833-0.954)	0.880 (0.811-0.936)	0.969 (0.936-0.993)
11	Ebstein's Anomaly	0.999 (0.998-1.000)	0.999 (0.998-1.000)	1.000 (1.000-1.000)	0.999 (0.997-1.000)	1.000 (1.000-1.000)	0.941 (0.886-0.987)	0.848 (0.764-0.918)	0.921 (0.853-0.974)	0.889 (0.817-0.952)	0.988 (0.961-1.000)
Class frequency-weighted average		0.964	0.967	0.975	0.970	0.991	0.794	0.802	0.831	0.792	0.884

AUROC, area under the receiver operating characteristic curve; CMR, cardiac magnetic resonance imaging; SAX, short-axis; 4CH, four chamber; LGE, late gadolinium enhancement. The bold font emphasizes the optimal performance metric among various input schemes.

Extended Data Table 6 | The distribution of the gold-standard test for human-AI comparison

Gold-standard test	Numbers (n=500)	Sex		Age (Range)
		Male	Female	
Total	500	345 (69%)	155 (31%)	46 ± 344 (2-83)
1 HCM	100	68 (68%)	32 (32%)	48 ± 14 (13-73)
2 DCM	100	74 (74%)	26 (26%)	45 ± 15 (12-78)
3 CAD	80	70 (80%)	10 (13%)	56 ± 10 (36-83)
4 LVNC	30	20 (67%)	10 (33%)	38 ± 19 (9-71)
5 RCM	30	19 (63%)	11 (37%)	45 ± 21 (7-77)
6 CAM	30	21 (70%)	9 (30%)	58 ± 8 (42-81)
7 HHD	30	27 (90%)	3 (10%)	42 ± 14 (19-74)
8 Myocarditis	20	14 (70%)	6 (30%)	31 ± 14 (14-69)
9 ARVC	30	18 (60%)	12 (40%)	40 ± 16 (11-68)
10 PAH	30	9 (30%)	21 (70%)	35 ± 10 (19-56)
11 Ebstein's Anomaly	20	5 (25%)	15 (75%)	35 ± 14 (2-59)

HCM, hypertrophic cardiomyopathy; DCM, dilated cardiomyopathy; CAD, coronary artery disease; LVNC, left ventricular non-compaction; RCM, restrictive cardiomyopathy; CAM, cardiac amyloidosis; HHD, hypertensive heart disease; ARVC, arrhythmogenic right ventricular cardiomyopathy; PAH, pulmonary arterial hypertension.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.

For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection Cardiac magnetic resonance imaging (CMR) data was collected for this study. Snap-ITK (version 4.0.1) and 3Dslicer (version 4.11.20210226) were used to visualize the CMR data and annotate the cardiac region of interests (ROI).

Data analysis Our codes are available at <https://github.com/MedAI-Vision/CMR-AI>. We used Python 3.8 and PyTorch 1.10. A comprehensive list of dependencies and their version number is available at <https://github.com/MedAI-Vision/CMR-AI/blob/main/requirements.txt>.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

No publicly available datasets were used in this study. The de-identified data can be shared only for non-commercial academic purposes and will require a formal material transfer agreement and a data use agreement. Requests should be submitted by emailing the corresponding authors (S.Z. or Y.J.W.) at cjrzhaojihua2009@163.com or wangyanran100@gmail.com. All requests will be evaluated based on institutional policies to determine whether the data requested are subject to intellectual property or patient privacy obligations. Generally, all such requests for access to CMR data will be responded to within 1 month. Example CMR data in this study is available in Extended Data Movies.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender	All reported findings apply to patients of any sex or gender.
Reporting on race, ethnicity, or other socially relevant groupings	The findings of this study are applicable to individuals of all races and ethnicities, as race, ethnicity, or other socially relevant groupings were not inclusion or exclusion criteria. However, it is important to note that all participating institutions are located in eastern Asia. Future investigations should explore the generalizability of the model across diverse ethnicities to ensure its broad utility.
Population characteristics	Table 1 and Extended Data Table 1 detailed patient demographics.
Recruitment	The study included retrospective investigation of imaging studies from patients admitted during standard clinical care. Patients were not directly recruited in this study.
Ethics oversight	The CMR datasets were acquired retrospectively under the approval of the institutional review boards (IRBs) at each participating institution, including Beijing Fuwai Hospital (FW), Beijing Anzhen Hospital (AZ), Guangdong Provincial People's Hospital (GD), the 2nd Affiliated Hospital of Harbin Medical University (HEB), the First Hospital of Lanzhou University (LZ), Renji Hospital (RJ), Tongji Hospital (TJ), and Peking Union Medical College Hospital (XH). Informed consent was waived by the IRBs. Before model training, testing, and reader studies, all data underwent de-identification processes..

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was determined by the availability of high quality CMR and clinical data. No additional statistical method for sample size estimation was used.
Data exclusions	Data inclusion and exclusion criteria was described in Methods and Extended Data Figure 1. Exclusion criteria were (1) incomplete cine or LGE modalities; (2) SAX cine with fewer than 5 views; (3) CMR images with insufficient scan quality; (4) CVD patients missing clinical data; (5) CMR exams that could not be interpreted and agreed upon by the committee cardiologists according to the diagnostic criteria (Methods).
Replication	Three-fold cross validation was performed within the primary discovery cohort to further validate the model performance. We also included true external validation. We were able to replicate model performance results both internally and externally, covering eight medical centers. Hyperparameters were described in detail in Methods section for replication of model results. Confidence intervals when applicable represent variation of results or performance.
Randomization	For each three-fold cross validation, patients were randomly assigned to either the training or the validation set. For the annotation procedure, every CMR record was randomly assigned to be reviewed by a single physician (Method: Annotation procedures). For the generation of Table 2 (human-machine comparison), the 500 subjects with CMR data were randomly selected from the primary discovery cohort with a fixed CVD class ratio.
Blinding	During dataset collection, CMR experts meticulously reviewed all records and clinical reports to annotate the data with cardiovascular disease (CVD) labels, serving as ground truth for model training and evaluation. Consequently, they were not blinded to the clinical records during this process. However, in the human-machine comparison experiment involving six physicians, each physician was blinded to the interpretations of their peers and AI, as well as to the clinical and CMR reports of the patients.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	<input type="text"/>
Research sample	<input type="text"/>
Sampling strategy	<input type="text"/>
Data collection	<input type="text"/>
Timing	<input type="text"/>
Data exclusions	<input type="text"/>
Non-participation	<input type="text"/>
Randomization	<input type="text"/>

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	<input type="text"/>
Research sample	<input type="text"/>
Sampling strategy	<input type="text"/>
Data collection	<input type="text"/>
Timing and spatial scale	<input type="text"/>
Data exclusions	<input type="text"/>
Reproducibility	<input type="text"/>
Randomization	<input type="text"/>
Blinding	<input type="text"/>

Did the study involve field work? Yes No

Field work, collection and transport

Field conditions	<input type="text"/>
Location	<input type="text"/>
Access & import/export	<input type="text"/>
Disturbance	<input type="text"/>

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern
<input checked="" type="checkbox"/>	<input type="checkbox"/> Plants

Methods

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used	<input type="text"/>
Validation	<input type="text"/>

Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s)	<input type="text"/>
Authentication	<input type="text"/>
Mycoplasma contamination	<input type="text"/>
Commonly misidentified lines (See ICLAC register)	<input type="text"/>

Palaeontology and Archaeology

Specimen provenance	<input type="text"/>
Specimen deposition	<input type="text"/>
Dating methods	<input type="text"/>
<input type="checkbox"/> Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.	
Ethics oversight	<input type="text"/>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about [studies involving animals; ARRIVE guidelines](#) recommended for reporting animal research, and [Sex and Gender in Research](#)

Laboratory animals	<input type="text"/>
Wild animals	<input type="text"/>
Reporting on sex	<input type="text"/>
Field-collected samples	<input type="text"/>
Ethics oversight	<input type="text"/>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration	<input type="text"/>
Study protocol	<input type="text"/>
Data collection	<input type="text"/>
Outcomes	<input type="text"/>

Dual use research of concern

Policy information about [dual use research of concern](#)

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

- | No | Yes |
|--------------------------|--------------------------|
| <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> | <input type="checkbox"/> |
- Public health
- National security
- Crops and/or livestock
- Ecosystems
- Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

- | No | Yes |
|--------------------------|--------------------------|
| <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> | <input type="checkbox"/> |
- Demonstrate how to render a vaccine ineffective
- Confer resistance to therapeutically useful antibiotics or antiviral agents
- Enhance the virulence of a pathogen or render a nonpathogen virulent
- Increase transmissibility of a pathogen
- Alter the host range of a pathogen
- Enable evasion of diagnostic/detection modalities
- Enable the weaponization of a biological agent or toxin
- Any other potentially harmful combination of experiments and agents

Plants

Seed stocks	<input type="text"/>
Novel plant genotypes	<input type="text"/>
Authentication	<input type="text"/>

ChIP-seq

Data deposition

- Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).
- Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links <i>May remain private before publication.</i>	<input type="text"/>
Files in database submission	<input type="text"/>
Genome browser session (e.g. UCSC)	<input type="text"/>

Methodology

Replicates	<input type="text"/>
Sequencing depth	<input type="text"/>
Antibodies	<input type="text"/>
Peak calling parameters	<input type="text"/>
Data quality	<input type="text"/>

Software

Flow Cytometry

Plots

Confirm that:

- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Instrument

Software

Cell population abundance

Gating strategy

- Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type

Design specifications

Behavioral performance measures

Imaging type(s)

Field strength

Sequence & imaging parameters

Area of acquisition

Diffusion MRI

Used

Not used

Preprocessing

Preprocessing software

Normalization

Normalization template

Noise and artifact removal

Volume censoring

Statistical modeling & inference

Model type and settings

Effect(s) tested

Specify type of analysis: Whole brain ROI-based Both

Statistic type for inference

(See [Eklund et al. 2016](#))

Correction

Models & analysis

n/a | Involved in the study

 Functional and/or effective connectivity Graph analysis Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Graph analysis

Multivariate modeling and predictive analysis

