Automated detection of severe aortic stenosis using single-view echocardiography: A self-supervised ensemble learning approach

Gregory Holste BA[1,2*], Evangelos K. Oikonomou MD DPhil[2*], Bobak J. Mortazavi PhD[3,4], Kamil F. Faridi MD MSc[2], Edward J. Miller MD PhD[2], John K. Forrest MD[2], Robert L. McNamara MD MHS[2], Harlan M. Krumholz MD SM[2,4,5], Zhangyang Wang PhD[1], Rohan Khera MD MS[2,4,6]

[1] Department of Electrical and Computer Engineering, The University of Texas at Austin, TX, USA

[2] Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, CT, USA

[3] Department of Computer Science & Engineering, Texas A&M University, College Station, TX, USA

[4] Center for Outcomes Research and Evaluation, Yale-New Haven Hospital, New Haven, CT, USA

[5] Department of Health Policy and Management, Yale School of Public Health, New Haven, CT

[6] Section of Health Informatics, Department of Biostatistics, Yale School of Public Health, New Haven, CT

*Contributed equally as co-first authors

Manuscript type: Article

Word count: 2,565

Display items: 7

Brief title: Automated severe aortic stenosis detection

Address for correspondence:

Rohan Khera, MD, MS

195 Church St, 5th Floor, New Haven, CT 06510

203-764-5885; rohan.khera@yale.edu; @rohan_khera

Abstract

Early diagnosis of aortic stenosis (AS) is critical to prevent morbidity and premature mortality but requires skilled examination with Doppler imaging. We hypothesized that self-supervised learning of 2-dimensional parasternal long axis (PLAX) videos from transthoracic echocardiography (TTE) without Doppler imaging could extract discriminative features to identify severe AS suitable for point-of-care ultrasonography. In a training set of 5,311 studies (17,601 videos) from 2016-2020, we performed self-supervised pretraining based on contrastive learning of PLAX videos, then used those learned weights to initialize a convolutional neural network to predict severe AS in an external set of 2,040 studies from 2021. Our model achieved an AUC of 0.97 (95% CI: 0.96-0.99) for detecting severe AS with 95.8% sensitivity and 90% specificity. The models were interpretable with saliency maps identifying the aortic valve as the predictive region. Among non-severe AS cases, predicted probabilities were associated with worse quantitative metrics of AS suggesting association with AS severity. We propose an automated approach for severe AS detection using single-view 2D echocardiography, with implications for point-of-care screening.

INTRODUCTION

Aortic stenosis (AS) is a chronic, progressive disease, and associated with premature morbidity and mortality.[1,2] With advances in both surgical and transcatheter aortic valve replacement,[3] there has been an increasing focus on early detection and management.[4,5,6] The non-invasive diagnosis of AS can be done with hemodynamic measurements using Doppler echocardiography,[2,7,8] but it requires dedicated equipment and skilled acquisition and interpretation. On the other hand, even though two-dimensional (2D) cardiac ultrasonography is increasingly available with handheld devices that can visualize the heart,[9] it cannot yet be used for efficient and accurate screening of AS presence and severity. With an estimated prevalence of 5% among individuals aged 65 years or older,[8] there is a growing need for user-friendly screening tools which can be used in everyday practice by people with minimal training.

Machine learning offers opportunities to standardize the acquisition and interpretation of medical images.[10] Deep learning algorithms have successfully been applied in echocardiograms, where they have shown promise in detecting left ventricular dysfunction,[11] and left ventricular hypertrophy.[12] With the expanded use of point-of-care ultrasonography,[9] developing user-friendly screening algorithms relying on single 2D echocardiographic views would provide an opportunity to improve AS screening. This is however limited by the lack of carefully curated, labelled datasets, as well as efficient ways to utilize the often noisy real-world data for model development.[13]

In the current study, we hypothesized that a deep learning model trained on 2D echocardiographic views of parasternal long axis (PLAX) videos can reliably predict the presence of severe AS without requiring Doppler input. The approach leverages self-supervised learning of PLAX videos along with two other neural network initialization methods to form a

diverse ensemble model capable of identifying severe AS from raw 2D echocardiograms. The model is trained based on a carefully curated dataset from different operators and machines, and its discriminatory performance for severe AS is then tested in a retrospective cohort of echocardiographic studies performed at a later chronological date. Combined with automated view classification, our approach serves as an end-to-end automated solution for deep learning applications in the field of echocardiography.

RESULTS

Study Population

This study included individuals who underwent transthoracic echocardiograms (TTE) between 2016 and 2021 across the Yale-New Haven Health System. From the studies performed between 2016 and 2020 (n=257,829), a stratified weighted sample of 10,000 studies was drawn that overweighted studies with AS (sampling probability weights of 1 for no AS, 5 for non-severe AS, 50 for severe AS). After removing 3,378 studies with no pixel data, de-identifying video frames, and using an automated view classifier to determine the PLAX view, our final derivation set (training and validation) consisted of 6,021 studies with 22,912 videos (1,269,764 frames) (mean age 70.2 ± 15.7 years, n=2950 (49.0%) women), with mild, moderate, and severe AS in 12.4% (n=747), 8.4% (n=503), and 22.3% (n=1,344) of studies, respectively. A held-out, randomly selected, sample of 1,063 studies from the same period was used for (internal) testing, whereas 2,040 randomly selected scans with a total of 6,530 videos performed between January 1st 2021 and December 15th 2021 (mean age 65.7 ± 16.4 years, n=997 (48.9%) women) were used for external testing. The external test was *not* oversampled, with mild, moderate, and severe

AS estimated in 4.1% (n=83), 2.9% (n=59), and 1.0% (n=20) of the studies, respectively (**Figure 1**). Further information on patient characteristics is presented in the **Methods** and **Table 1**.

Self-supervised, contrastive learning

To learn transferable representations of PLAX echocardiogram videos for downstream severe AS identification, we performed self-supervised pretraining on all training set videos. This pretraining step critically enables the model to learn representations of echocardiograms that are robust to standard variations in video acquisition, thus better generalizing when later fine-tuned on a specific downstream task. We have previously demonstrated that a more appropriate initialization for data-efficient classification tasks could be achieved by "in-domain" pretraining on echocardiograms,[17] as opposed to other standard approaches such as random initialization of weights and transfer learning.[11,15,16] To this end, we have designed a novel self-supervised learning algorithm specifically catered to echocardiogram videos.

For this, we selected different PLAX videos from the same patient as "positive samples" for multi-instance contrastive learning (**Figure 2**). While contrastive learning frameworks such as SimCLR[18] rely on image augmentation to create two synthetic "views" of the same image for contrastive learning, we instead leverage the fact that patients often already have multiple PLAX videos to form challenging positive pairs between authentically distinct views of the patient. This critically removed the need for heavy augmentation strategies that would adversely affect valuable signal. To enforce temporal coherence, we additionally used a frame re-ordering pretext task during self-supervised pretraining, where we randomly shuffled the frames of an echocardiographic video, then trained the model to predict the original order of frames.

We pretrained a 3D-ResNet18[19] model with this joint contrastive and frame re-ordering objective on all 30,008 unique pairs of distinct PLAX videos from the same patient for 300 epochs. We then used these learned weights to initialize a 3D-ResNet18 to predict severe AS, producing a model that achieved 0.934 AUROC (95% CI: 0.920, 0.947) and 86.0% specificity at 90% sensitivity (95% CI: 79.7%, 88.4%) on the internal testing set (**Table 2**, **Figure 3**).

Ensemble learning for severe AS identification

We formed an ensemble of three models trained to detect severe AS, with diversity injected by the dramatically different methods used to initialize each model's weights before training. Ensembling is known to improve predictive performance by aggregating the outputs of multiple independently trained models[20]. Moreover, statistical[21–23] and deep learning[24–26] studies have shown that ensembles of *diverse* constituent models are most effective.

To do this, we first independently trained three models to predict severe AS: a randomly initialized model, a model initialized with weights from the human action classification dataset Kinetics-400[27] (representing a standard transfer learning approach), and an SSL-initialized model using the pretraining method described above. The outputs of these three models were then averaged to produce a powerful and diverse ensemble model, reaching 0.945 AUROC (95% CI: 0.933, 0.956) and 88.0% sensitivity at 90% specificity (95% CI: 84.3%, 90.1%) on the internal testing set and 0.974 AUROC (95% CI: 0.957, 0.989) and 95.8% sensitivity at 90% specificity (95% CI: 83.2%, 97.3%) on the external testing set (**Table 2**, **Figure 3**). Additionally, if multiple PLAX videos were present in a study, the predictions from each video were ensembled to form a single study-level AS prediction. Internal test results stratified by number of PLAX videos used

to form "per-study ensembles" can be found in **Extended Data Table 1**, and results *without* averaging predictions from multiple videos in the same study appear in **Extended Data Table 2**.

Explainable predictions through saliency maps

To increase explainability of our models, we used Gradient-weighted Class Activation Mapping (Grad-CAM) to identify the regions in each video frame that contributed the most to the predicted label.[28] These spatial attention maps were visualized based on the randomly initialized, Kinetics-400-pretrained, and SSL-pretrained AS models for five true positives, a true negative, and a false positive. In the examples shown in **Figure 4**, the first five columns represent the five most confident severe AS predictions, the sixth column represents the most confident "normal" (no severe AS) prediction, and the seventh column represents the most confident incorrect severe AS prediction. The saliency maps from our SSL approach demonstrated overall consistent and specific localization of the activation signal in the pixels corresponding to the aortic valve and annulus (bottom row). Relative to the saliency maps generated by the randomly initialized and Kinetics-400-pretrained models, the SSL attention maps more finely localized clinically relevant regions for AS detection.

Model identification of features of AS severity

Finally, we explored whether our model learned features of AS severity that could describe earlier stages of the disease's natural history. We observed that the predictions of the ensemble model correlated with continuous metrics of AS severity, including the peak aortic valve velocity ($r=0.60$, $P<0.001$), trans-valvular mean gradient ($r=0.69$, $P<0.001$) and the mean aortic valve area ($r=-0.51$, $P<0.001$). On the other hand, the model predictions were independent of the left

ventricular ejection fraction (LVEF), a negative control. In further sensitivity analysis, we stratified cases without AS or mild/moderate AS based on the predictions of our model as true negatives (TN) or false positives (FP). Compared to true negatives, false positive cases had significantly higher peak aortic velocities (FP: 3.4 [25th-75th percentile: 2.6-3.6] m/sec; TN: 1.4 [1.2-1.7] m/sec, P<0.001), trans-valvular mean gradients (FP: 27.0 [25th-75th percentile: 18.0-30.5] mmHg; TN: 5.0 [3.8-9.8] m/sec, P<0.001), and mean aortic valve area (FP: 0.94 [25th-75th percentile: 0.75-1.62] cm^2; TN: 1.97 [1.47-2.66] cm^2, P=0.001), but no significant difference in the LVEF (FP: 65.6% [52.3%-68.2%]; TN: 60.0% [55.0-65.1%], P=0.36) (**Figure 5**).

DISCUSSION

We have developed and validated an automated algorithm that can efficiently screen for and detect the presence of severe AS based on a single-view two-dimensional transthoracic echocardiographic video. The algorithm demonstrates excellent performance (0.974 AUROC), with high sensitivity (95.8%) and specificity (90%) confirmed in an external cohort of patients temporally distinct from the training set. We also present a novel self-supervised step leveraging multi-instance contrastive learning, which allowed our algorithm to learn key representations that define each patient's unique phenotype, independent of the expected technical variation in image acquisition, including differences in probe orientation, beam angulation and depth. Visualization of saliency maps introduces explainability to our algorithms and confirms the key areas of the PLAX view, including the aortic valve and annulus, that contributed the most to our predictions. Furthermore, features learned by the model generalize to lower severity cases, highlighting the potential value of our model in longitudinal monitoring of AS, a disease with a well-defined, progressive course.[2] Our approach has the potential to expand the use of echocardiographic

screening for suspected AS, shifting the burden away from dedicated echocardiographic laboratories to point-of-care screening in primary care offices, or low-resource settings. It may also enable operators with minimal echocardiographic experience to screen for the condition by obtaining simple two-dimension PLAX views without the need for comprehensive Doppler assessment, which can then be reserved for confirmatory assessment.

In the recent years a number of artificial intelligence applications have been described in the field of echocardiography,[29] ranging from automated classification of echocardiographic views,[30] video-based beat-to-beat assessment of left ventricular systolic dysfunction,[11] detection of left ventricular hypertrophy and its various subtypes,[12] diastolic dysfunction,[31] to expert-level prenatal detection of complex congenital heart disease.[32] Of note, machine learning methods further enable individuals without prior ultrasonography experience to obtain diagnostic TTE studies for limited diagnostic use.[33] Despite this and even though the diagnosis and grading of AS remains dependent on echocardiography,[2,34] most artificial intelligence solutions for timely AS screening have focused on alternative data types, such as audio files of cardiac auscultation,[35] 12-lead electrocardiograms,[36–38] cardio-mechanical signals using non-invasive wearable inertial sensors,[39] as well as chest radiographs.[40] For 12-lead electrocardiograms, AUC were consistently <0.90,[36–38] whereas for alternative data types, analyses were limited to small datasets without external validation.[35,39] Other studies have explored the value of structured data derived from comprehensive TTE studies in defining phenotypes with varying disease trajectories.[41] However, the value of AI-assisted AS detection through automated TTE interpretation has not been fully explored. In a recent study, investigators employed a form of self-supervised learning to automate the detection of AS, with their method however discarding temporal information by only including the first frame of each video loop, while also relying on the acquisition of images

from several different views.[42] The approach that relies on ultrasonography is also safer than the alternative screening strategies, such as those using chest computed tomography and aortic valve calcium scoring,[40,42] which expose patients to radiation.

In this context, our work represents an advance both in the clinical and methodological space. First, we describe a method that can efficiently screen for a condition associated with significant morbidity and mortality,[2,7] with increasing prevalence in the setting of an aging population.[43] Our method has the potential to shift the initial burden away from trained echocardiographers and specialized core laboratories, as part of a more cost-effective screening and diagnostic cascade.[9,33] In this regard, major strengths of our model include its reliance on a single echocardiographic view that can be obtained by individuals with limited experience and minimal training,[33] and its ability to process temporal information through analysis of videos rather than isolated frames. The overarching goal is to develop screening tools that can be deployed in a cost-effective manner, gatekeeping access to comprehensive TTE assessment, which can be used as a confirmatory test to establish the suspected diagnosis.

Second, our work describes an end-to-end framework to boost artificial intelligence applications in echocardiography. We present an algorithm that automatically detects echocardiographic views, then performs self-supervised representation learning of PLAX videos with a multi-instance, contrastive learning approach. This novel approach further enables our algorithm to learn key representations of a patient's cardiac phenotype that generalize and remain consistent across different clips and variations of the same echocardiographic views. By optimizing the detection of an echocardiographic fingerprint for each patient, this important pretraining step has the potential to boost AI-based echocardiographic assessment across a range of conditions. Furthermore, unlike previous approaches,[42] our method benefits from multi-

instance contrastive learning, which learns key representations using different videos from the same patient, a method that has been shown to improve predictive performance in the classification of dermatology images.[44]

Further to detecting severe AS, our algorithm learns features of aortic valvular pathology that generalize across different stages of the condition. Saliency maps demonstrate that the model focuses on the aortic valve area, possibly learning to detect aortic valve calcification and restricted mobility.[34] When restricting our analysis to patients without severe AS, the model's predictions strongly correlated with Doppler-derived, quantitative features of stenosis severity. This is in accordance with the known natural history of AS, a progressive, degenerative condition, the hallmarks of which are aortic valve calcification, restricted mobility, functional stenosis and eventual ventricular decompensation.[2,7] As such, our algorithm's predictions also carry significant value as quantitative predictors of the stage of AV severity and could theoretically be used to monitor the rate of AS progression.

Limitations of our study include the lack of prospective validation of our findings. To this end, we are working on deploying this method in a prospective cohort of patients referred for routine TTE assessment to understand its real-world implications as a screening tool. Second, even though our training set included a range of vendors, several different hospitals, and studies chronologically separated from the ones used for testing purposes, further external validation is needed to better understand the generalizability of our observations across healthcare systems. Third, our model is limited to the use of PLAX views, which often represent the first step of TTE or POCUS protocols in cardiovascular assessment. Though there is no technical restriction to expanding these methods to alternative views, increasing the complexity of the screening protocol is likely to negatively impact its adoption in busy clinical settings.

In summary, we propose an efficient method to screen for severe AS using single-view (PLAX) TTE videos without the need for Doppler signals. More importantly, we describe an end-to-end approach for the deployment of artificial intelligence solutions in echocardiography, starting from automated view classification to self-supervised representation learning to accurate and explainable detection of severe AS. Our findings have significant implications for point-of-care ultrasound screening of AS as part of routine clinic visits and in limited resource settings and for individuals with minimal training.

METHODS

The study was reviewed by the Yale Institutional Review Board, which approved the study protocol and waived the need for informed consent as the study represents secondary analysis of existing data.

Echocardiogram interpretation

All studies were performed by trained echocardiographers or cardiologists and reported by board-certified cardiologists with specific training cardiac echocardiography. These reports were a part of routine clinical care, in accordance with the recommendations of the American Society of Echocardiography (ASE).[33,45] The presence of AS severity was adjudicated based on the original echocardiographic report. Doppler assessment was interpreted based on the parameters recommended by the ASE, which included peak aortic valve (stenosis) jet velocity, mean transaortic/trans-valvular gradient, and mean valve area, as assess by continuity equation. According to the guidelines, cut-offs of >4 m/sec, >40 mm Hg and less than <1.0 cm^2, respectively, were consistent with severe AS.[34] The left ventricular ejection fraction (LVEF) was

reported based on three-dimensional (3D) echocardiography, and in the absence of that, based on the Simpson's biplane method. In the absence of these measurements, we reported the lower end of the visually estimated LVEF.

Study cohort

Dataset Preparation. For this work, 12,500 studies were queried from the set of all complete TTE exams performed between 2016 and 2021 at the Yale New Haven Health System (**Figure 1**). For internal model development and evaluation, 10,000 studies from 2016-2020 were randomly queried with AS oversampled to mitigate class imbalance during model training. Specifically, this query sampled normal studies uniformly (including "no AS" and "sclerosis without stenosis"), oversampled non-severe AS studies 5-fold (including "mild AS", "mild-moderate AS", "moderate-severe AS", "low gradient AS", and "paradoxical AS"), and oversampled severe AS 50-fold. The 10,000 studies would later be split into an internal training, validation, and test set for deep learning model development. The remaining 2,500 studies of the query were all conducted a year later in 2021 with *no oversampling* to serve as a more challenging and clinically realistic "external" validation set, where severe AS only occurs under 1% of the time. All studies would undergo de-identification, view classification, and preprocessing (as described below) to curate a dataset of PLAX videos for deep-learned severe AS prediction.

*View classification***.** After excluding studies that were not properly extracted from the database, 10,865 studies first underwent de-identification. After loading the pixel data for each video with the pydicom library (https://pydicom.github.io/), pixels in the periphery of each video frame

were masked out to remove identifying information, and videos were converted to the Audio Video Interleave (AVI) format to enable fast loading for later preprocessing steps. The resulting 447,653 videos from 9,710 studies then underwent view classification. Using the pretrained TTE view classifier from Zhang *et al.*,[46] ten frames from each de-identified video were randomly selected, downsampled to 224 x 224 resolution, and fed through the pretrained VGG19 convolutional neural network. Video-level view predictions were then obtained by averaging each video's 10 frame-wise view probabilities, and videos that were most confidently predicted as PLAX were kept for further preprocessing. While the pretrained view classifier was capable of discriminating variants of the standard PLAX view such as "PLAX", "PLAX – remote," "PLAX – zoom of left atrium," and "PLAX – centered over left atrium," we elected to only proceed with videos most confidently classified as "PLAX."

Data preprocessing. After view classification, the 30,136 videos from 9,173 studies were prepared for deep learning model development. Given differences in AS severity measures across different domains, we excluded echocardiograms with low-flow, low-gradient & paradoxical aortic stenosis leaving 29,978 PLAX videos from 9,122 studies. Since severe AS detection was formulated as a binary classification task, all AS designations other than "severe AS" were binned into the "not severe AS" category. The 2,040 studies from 2021 were set aside for external validation, while the remaining 7,082 studies (with AS oversampled as described above) were then randomly split into derivation (training & validation) and test sets according to a 75%/10%/15% split (Table 1). All videos underwent a more thorough cleaning and de-identification process that involved binarizing each video frame with a fixed threshold, then masking out all pixels outside the convex hull of the largest contour in order to remove all

information outside the central image content. Finally, each video clip was spatially downsampled to 112 x 112 and saved to AVI format for fast loading during model training.

Self-supervised contrastive learning

Self-supervised representation learning was performed on the training set videos with a novel combination of (i) a multi-instance contrastive learning task and (ii) a frame re-ordering pretext task. We adopted the SimCLR framework[18] for contrastive learning, which traditionally generates two "views" of an input x by sending two copies of the input through a pipeline of random image augmentations, producing view $\tilde{x}_i$ and $\tilde{x}_j$. An encoder $f()$ is then used to learn representations of each view, $h_i = f(\tilde{x}_i)$ and $h_j = f(\tilde{x}_j)$, which are then projected to a lower dimensionality with a projector $g()$. The resulting learned embeddings of each view, $z_i = g(f(\tilde{x}_i))$and $z_j = g(f(\tilde{x}_j))$ are then "contrasted" via the temperature-normalized cross-entropy (NT-Xent) loss, which encourages the model to learn *similar* representations of views from the same original image (so-called "positive pairs") and *dissimilar* representations of views from all other images ("negative pairs") in a given minibatch.

While SimCLR has proven very successful for 2D natural images as well as in medical applications such as radiography and dermatological images, there are several barriers to its successful adaptation to echocardiogram videos. First, SimCLR requires extremely heavy image augmentation for effective representation learning, which would destroy valuable signal encoded in the brittle, noisy ultrasound images produced by echocardiography. Second, SimCLR was designed for 2D images, which would completely ignore the temporal dimension of echocardiography. To address the first issue, we utilized "multi-instance" contrastive learning –

borrowing language and key insights from Azizi *et al.*[44] – whereby we form positive pairs between *different* videos from the *same* patient. This critically removes the need to synthetically create two different "views" of a patient by heavily augmenting their echo video, instead leveraging the fact that almost all studies contain multiple distinct PLAX videos of a patient.

To address the second issue, we additionally included a frame re-ordering "pretext" task to our self-supervised learning method, where we randomly permuted the frames of each input echo, then trained the model to predict the original order of frames. Similar to the approach of Jiao *et al.*,[47] this frame re-ordering task is treated as a classification problem and was implemented with a simple fully-connected layer that minimizes the cross-entropy between the known and predicted original frame order; specifically, if an input echo clip has K frames, then the $K!$ possible permutations of frames served as the targets for classification. Then the loss function of our self-supervised learning method is simply the sum of the contrastive NT-Xent objective and the pretext frame re-ordering cross-entropy objective.

Self-supervised pretraining was performed on randomly sampled video clips of 4 consecutive frames from each of the training set echocardiogram videos. The encoder $f()$ was a randomly initialized 3D-ResNet18,[19] and the projector $g()$ projected each 512-dimensional learned representation down to a 128-dimensional representation with a hidden layer of 256 units followed by a ReLU activation, followed by an output layer with 128 units. The model was trained for 300 epochs on all unique pairs of *different* PLAX videos from the same patient with the Adam optimizer[48] and a learning rate of 0.1, a batch size of 392 (196 per GPU), and NT-Xent temperature hyperparameter of 0.5. The following augmentations were applied to each frame in a temporally consistent manner (same transformations for each frame of a given video clip): random zero padding by up to 8 pixels in each spatial dimension, a random horizontal flip with

probability 0.5, and a random rotation within -10 and 10 degrees with probability 0.5. After augmentation, each video clip was normalized so that the maximum pixel intensity was mapped to 1 and the minimum intensity to 0. The SSL model was trained on two NVIDIA RTX 3090 GPUs.

Deep neural network training for severe AS prediction

The same 3D-ResNet18 architecture was used to predict severe AS. Three different methods were used to initialize the parameters of this network: an SSL initialization, a Kinetics-400 initialization, and a random initialization. The SSL initialization directly used the learned weights of the encoder from the SSL pretraining step described in detail above. The Kinetics-400 initialization represents the "standard" transfer learning approach for 3D data, using the weights from a 3D-ResNet18 trained in a supervised fashion on the Kinetics-400 dataset, a large corpus of over 300,000 natural videos for human action classification; these weights are readily available through the torchvision API (https://pytorch.org/vision/stable/index.html) provided by PyTorch. The random initialization is the default when initializing a 3D-ResNet18 with PyTorch.[49]

All fine-tuning models were trained on randomly sampled video clips of 16 consecutive frames from training set echocardiograms, optionally padding with empty frames along the temporal axis if either the video is too short or the randomly chosen starting point of the clip is near the end of the video. The same augmentations were used as in self-supervised pretraining, and all video clips were min-max normalized; when fine-tuning from a Kinetics-400 initialization, video clips were further standardized using the channel-wise means and standard deviations from the Kinetics-400 training dataset, a standard preprocessing step when performing

transfer learning. All models were trained for a maximum of 30 epochs with early stopping – specifically, if the validation AUROC did not improve for 5 consecutive epochs, training was terminated and the weights from the epoch with maximum validation AUROC were used for final evaluation. Severe AS models were trained on a single NVIDIA RTX 3090 GPU with the Adam optimizer, a learning rate of 1×10^{-4} (except the SSL-pretrained model, which used a learning rate of 0.1), and a batch size of 88 in order to maximize GPU utilization. Since this problem was framed as a binary classification task, these models minimized a sigmoid cross-entropy loss. We additionally used class weights computed with the method provided by scikit-learn[50] to accommodate class imbalance in addition to label smoothing[51,52] with α=0.1, a method to improve model calibration and generalization. Learning curves depicting loss throughout training can be found in **Extended Data Figure 1**.

Ensemble learning

Since the fine-tuned severe AS models are trained on 16-frame video clips, yet AS labels describe each study, we aggregated clip-level predictions into study-level predictions for performance evaluation. When performing inference on an echo video, four evenly spaced clips (potentially with overlapping frames) of 16 consecutive frames were extracted and fed into the trained AS model. These clip-level predictions were then averaged to obtain a video-level prediction of severe AS. After repeating this process for all videos, the severe AS probabilities for videos in a given study were averaged to obtain study-level AS predictions that could be used to compute evaluation metrics. The final ensemble model is then formed by averaging the output probabilities of the SSL-pretrained model, the Kinetics-400-pretrained model, and the randomly initialized model after fine-tuning each ensemble member to classify severe AS.

These ensemble approaches of integrating information from multiple clips of a video and integrating information from multiple videos in a study are not required, but useful for improving predictive performance of study-level AS labels. See **Extended Data Table 1** for an examination of results by number of PLAX videos used to form per-study ensembles and **Extended Data Table 2** for results without averaging predictions from multiple videos in the same study. In this scheme, evaluation is performed at the *video* level, where all videos in a given study share the same label. Since no quality control is applied when selecting PLAX videos for this work, averaging results over multiple videos in the same study has a stabilizing effect that boosts predictive performance.

Internal and external validation

After fine-tuning the model to detect severe AS from TTE videos, the model checkpoint from the epoch with maximum AUROC on the validation set was used for evaluation on both the internal and external test sets. To evaluate the model's performance, we primarily use area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPR). These both measure a binary classifier's overall predictive performance across all possible decision thresholds, with AUPR included since it may be more informative when class imbalance is present.[53] In addition to ROC curve and precision-recall curve analysis, we use metrics that assess severe AS predictive performance at a specific decision thresholds such as F1 score, positive predictive value (PPV), specificity at 90% sensitivity, and PPV at 90% sensitivity. For these metrics, we proceed with the threshold that maximizes F1 score, the harmonic mean of precision and recall, on the given evaluation set. The latter two metrics – specificity and PPV at 90% sensitivity – were included to give a clinically realistic assessment of the model's

performance at a highly sensitive operating point that would be required for real-world clinical deployment.

Model explainability

Saliency maps were generated by leveraging the Grad-CAM method[28] for obtaining visual explanations from deep neural networks. Specifically, the heatmaps presented in Figure 4 were generated by applying Grad-CAM to a clip of the first 32 frames of an echo, using the last convolution block of the 3D ResNet18 to generate a 7 x 7 x 4 (height x width x time) heatmap displaying roughly where the model is attending to over the spatial and temporal dimensions. The Grad-CAM output was interpolated to the original input dimension of 112 x 112 x 32 with the scipy "zoom" function; this process produces a frame-by-frame "visual explanation" of *where* the model is focusing frame by frame in order to make its prediction. However, to generate a single 2D heatmap for a given echo clip as seen in Figure 4, the pixelwise maximum along the temporal axis was taken to capture the most salient regions for severe AS predictions across all timepoints.

Statistical analysis

All 95% confidence intervals for model performance metrics were computed by bootstrapping. Specifically, 10,000 bootstrap samples (samples with replacement having equal sample size to the original evaluation set) of the given evaluation set were drawn, metrics were computed on this set of studies, and nonparametric confidence intervals were constructed with the percentile method. Bootstrapping was performed at the study level since the severe AS labels are provided at the study level. For analysis of correlation between model outputs and quantitative measures

of AS, categorical variables were summarised as numbers (percentages), whereas continuous variables are reported as mean values with standard deviation and visualized using violin plots. Continuous variables between two groups were compared using the Student's *t*-test. Pearson's *r* was used to assess the pairwise correlation between continuous variables. All statistical tests were two-sided with a significance level of 0.05, unless specified otherwise. Analyses were performed using Python (version 3.8.5).

DATA AVAILABILITY

The data are not available for public sharing given the restrictions in our institutional review board approval. The deidentified test set may be available to researchers under a data use agreement after the study has been published in a peer-reviewed journal.

CODE AVAILABILITY

The code repository for this work can be found at https://github.com/CarDS-Yale/echo-severe-AS.

ACKNOWLEDGMENTS

The study was supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health (under award K23HL153775 to R.K.). A.H.R. is supported in part by CNPq (310679/2016-8 and 465518/2014-1) and by FAPEMIG (PPM-00428-17 and RED-00081-16). R.K. received support from the National Heart, Lung, and Blood Institute of the National Institutes of Health (under award K23HL153775) and the Doris Duke Charitable Foundation (under award, 2022060). B.J.M. reported receiving grants from the National Institute of

Biomedical Imaging and Bioengineering, National Heart, Lung, and Blood Institute, US Food and Drug Administration, and the US Department of Defense Advanced Research Projects Agency outside the submitted work.

AUTHOR CONTRIBUTIONS

G.H. and E.K.O. performed the analyses, G.H., E.K.O. and R.K. drafted the manuscript, and all other authors provided critical revisions. R.K. supervised the study and is the guarantor.

COMPETING INTERESTS

E.K.O. is a co-founder of Evidence2Health, and has served as a consultant for Caristo Diagnostics, Ltd. B.J.M. has a pending patent on predictive models using electronic health records (US20180315507A1). A.H.R. is funded by Kjell och Märta Beijer Foundation. J.K.F. has received grant support/research contracts and consultant fees/honoraria/Speakers Bureau fees from Edwards Lifesciences and Medtronic. H.M.K. works under contract with the Centers for Medicare & Medicaid Services to support quality measurement programs, was a recipient of a research grant from Johnson & Johnson, through Yale University, to support clinical trial data sharing; was a recipient of a research agreement, through Yale University, from the Shenzhen Center for Health Information for work to advance intelligent disease prevention and health promotion; collaborates with the National Center for Cardiovascular Diseases in Beijing; receives payment from the Arnold & Porter Law Firm for work related to the Sanofi clopidogrel litigation, from the Martin Baughman Law Firm for work related to the Cook Celect IVC filter litigation, and from the Siegfried and Jensen Law Firm for work related to Vioxx litigation; chairs a Cardiac Scientific Advisory Board for UnitedHealth; was a member of the IBM Watson

Health Life Sciences Board; is a member of the Advisory Board for Element Science, the Advisory Board for Facebook, and the Physician Advisory Board for Aetna; and is the co-founder of Hugo Health, a personal health information platform, and co-founder of Refactor Health, a healthcare AI-augmented data management company. R.K. receives research support, through Yale, from Bristol-Myers Squibb. He is also a coinventor of U.S. Pending Patent Applications. 63/177,117, and 63/346,610, unrelated to the current work. He is also a founder of Evidence2Health, a precision health platform to improve evidence-based cardiovascular care. The remaining authors have no competing interests to disclose.

REFERENCES

1. Eugène Marc *et al.* Contemporary Management of Severe Symptomatic Aortic Stenosis. *J. Am. Coll. Cardiol.* **78**, 2131–2143 (2021).
2. Otto, C. M. & Prendergast, B. Aortic-valve stenosis--from patients at risk to severe valve obstruction. *N. Engl. J. Med.* **371**, 744–756 (2014).
3. Smith, C. R. *et al.* Transcatheter versus surgical aortic-valve replacement in high-risk patients. *N. Engl. J. Med.* **364**, 2187–2198 (2011).
4. Reardon, M. J. *et al.* Surgical or Transcatheter Aortic-Valve Replacement in Intermediate-Risk Patients. *N. Engl. J. Med.* **376**, 1321–1331 (2017).
5. Kang, D.-H. *et al.* Early Surgery or Conservative Care for Asymptomatic Aortic Stenosis. *N. Engl. J. Med.* **382**, 111–119 (2020).
6. The Early Valve Replacement in Severe Asymptomatic Aortic Stenosis Study. https://clinicaltrials.gov/ct2/show/NCT04204915.
7. Otto, C. M. *et al.* 2020 ACC/AHA Guideline for the Management of Patients With Valvular Heart Disease: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation* **143**, e72–e227 (2021).

8. Baumgartner, H. *et al.* 2017 ESC/EACTS Guidelines for the management of valvular heart disease. *Eur. Heart J.* **38**, 2739–2791 (2017).

9. Narula, J., Chandrashekhar, Y. & Braunwald, E. Time to Add a Fifth Pillar to Bedside Physical Examination: Inspection, Palpation, Percussion, Auscultation, and Insonation. *JAMA Cardiol* **3**, 346–350 (2018).

10. Dey, D. *et al.* Artificial intelligence in cardiovascular imaging: JACC state-of-the-art review. *J. Am. Coll. Cardiol.* **73**, 1317–1335 (2019).

11. Ouyang, D. *et al.* Video-based AI for beat-to-beat assessment of cardiac function. *Nature* **580**, 252–256 (2020).

12. Duffy, G. *et al.* High-Throughput Precision Phenotyping of Left Ventricular Hypertrophy With Cardiovascular Deep Learning. *JAMA Cardiol* **7**, 386–395 (2022).

13. Newgard, C. D. & Lewis, R. J. Missing data: How to best account for what is not known. *JAMA: the journal of the American Medical Association* vol. 314 940–941 (2015).

14. Holste, G., Oikonomou, E. K., Mortazavi, B., Wang, Z. & Khera, R. Self-supervised learning of echocardiogram videos enables data-efficient clinical diagnosis. *arXiv [cs.CV]* (2022).

15. Rajpurkar, P. *et al.* CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv [cs.CV]* (2017).

16. Gulshan, V. *et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **316**, 2402–2410 (2016).

17. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. in *Proceedings of the 37th International Conference on Machine Learning* (eds. Iii, H. D. & Singh, A.) vol. 119 1597–1607 (PMLR, 13--18 Jul 2020).

18. Tran, Wang, Torresani & Ray. A closer look at spatiotemporal convolutions for action recognition. *Proc. Estonian Acad. Sci. Biol. Ecol.* (2018).

19. Dietterich, T. G. Ensemble Methods in Machine Learning. in *Multiple Classifier Systems* 1–15 (Springer Berlin Heidelberg, 2000).

20. Dietterich, T. G. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Mach. Learn.* **40**, 139–157 (2000).

21. Opitz & Shavlik. Generating accurate and diverse members of a neural-network ensemble. *Adv. Neural Inf. Process. Syst.* (1995).

22. Brown, Wyatt, Tino & Bengio. Managing diversity in regression ensembles. *J. Mach. Learn. Res.* (2005).

23. Wasay, Hentschel & Liao. Mothernets: Rapid deep ensemble learning. *of Machine Learning …* (2020).

24. Lakshminarayanan & Pritzel. Simple and scalable predictive uncertainty estimation using deep ensembles. *Adv. Neural Inf. Process. Syst.* (2017).

25. Qummar, S. *et al.* A Deep Learning Ensemble Approach for Diabetic Retinopathy Detection. *IEEE Access* **7**, 150530–150539 (2019).

26. Kay, W. *et al.* The Kinetics Human Action Video Dataset. *arXiv [cs.CV]* (2017).

27. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *arXiv [cs.CV]* (2016).

28. Ghorbani, A. *et al.* Deep learning interpretation of echocardiograms. *NPJ Digit Med* **3**, 10 (2020).

29. Madani, A., Arnaout, R., Mofrad, M. & Arnaout, R. Fast and accurate view classification of echocardiograms using deep learning. *NPJ Digit. Med.* **1**, (2018).

30. Chiou, Y.-A., Hung, C.-L. & Lin, S.-F. AI-assisted echocardiographic prescreening of heart failure with preserved ejection fraction on the basis of intrabeat dynamics. *JACC Cardiovasc. Imaging* **14**, 2091–2104 (2021).

31. Arnaout, R. *et al.* An ensemble of neural networks provides expert-level prenatal detection of complex congenital heart disease. *Nat. Med.* **27**, 882–891 (2021).

32. Narang, A. *et al.* Utility of a Deep-Learning Algorithm to Guide Novices to Acquire Echocardiograms for Limited Diagnostic Use. *JAMA Cardiol* **6**, 624–632 (2021).

33. Baumgartner, H. *et al.* Echocardiographic assessment of valve stenosis: EAE/ASE recommendations

for clinical practice. *J. Am. Soc. Echocardiogr.* **22**, 1–23; quiz 101–2 (2009).

34. Voigt, I. *et al.* A deep neural network using audio files for detection of aortic stenosis. *Clin. Cardiol.* (2022) doi:10.1002/clc.23826.
35. Kwon, J.-M. *et al.* Deep learning-based algorithm for detecting aortic stenosis using electrocardiography. *J. Am. Heart Assoc.* **9**, e014717 (2020).
36. Cohen-Shelly, M. *et al.* Electrocardiogram screening for aortic valve stenosis using artificial intelligence. *Eur. Heart J.* **42**, 2885–2896 (2021).
37. Hata, E. *et al.* Classification of Aortic Stenosis Using ECG by Deep Learning and its Analysis Using Grad-CAM. in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* 1548–1551 (2020).
38. Yang, C., Ojha, B. D., Aranoff, N. D., Green, P. & Tavassolian, N. Classification of aortic stenosis using conventional machine learning and deep learning methods based on multi-dimensional cardio-mechanical signals. *Sci. Rep.* **10**, 17521 (2020).
39. Ueda, D. *et al.* Artificial intelligence-based detection of aortic stenosis from chest radiographs. *Eur Heart J Digit Health* **3**, 20–28 (2022).
40. Sengupta, P. P. *et al.* A machine-learning framework to identify distinct phenotypes of aortic stenosis severity. *JACC Cardiovasc. Imaging* **14**, 1707–1720 (2021).
41. Huang, Z., Long, G., Wessler, B. & Hughes, M. C. A New Semi-supervised Learning Benchmark for Classifying View and Diagnosing Aortic Stenosis from Echocardiograms. in *Proceedings of the 6th Machine Learning for Healthcare Conference* (eds. Jung, K., Yeung, S., Sendak, M., Sjoding, M. & Ranganath, R.) vol. 149 614–647 (PMLR, 06--07 Aug 2021).
42. Pawade, T. *et al.* Computed Tomography Aortic Valve Calcium Scoring in Patients With Aortic Stenosis. *Circ. Cardiovasc. Imaging* **11**, e007146 (2018).
43. Bonow, R. O. & Greenland, P. Population-wide trends in aortic stenosis incidence and outcomes. *Circulation* vol. 131 969–971 (2015).
44. Azizi, S. *et al.* Big self-supervised models advance medical image classification. in *2021 IEEE/CVF*

International Conference on Computer Vision (ICCV) (IEEE, 2021). doi:10.1109/iccv48922.2021.00346.

45. Mitchell, C. *et al.* Guidelines for performing a comprehensive transthoracic echocardiographic examination in adults: Recommendations from the American society of echocardiography. *J. Am. Soc. Echocardiogr.* **32**, 1–64 (2019).
46. Zhang, J. *et al.* Fully Automated Echocardiogram Interpretation in Clinical Practice. *Circulation* **138**, 1623–1635 (2018).
47. Jiao, J., Droste, R., Drukker, L., Papageorghiou, A. T. & Noble, J. A. Self-Supervised Representation Learning for Ultrasound Video. *Proc. IEEE Int. Symp. Biomed. Imaging* **2020**, 1847–1850 (2020).
48. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv [cs.LG]* (2014).
49. Paszke, Gross, Massa & Lerer. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* (2019).
50. Pedregosa, Varoquaux & Gramfort. Scikit-learn: Machine learning in Python. *the Journal of machine* (2011).
51. Müller, Kornblith & Hinton. When does label smoothing help? *Adv. Neural Inf. Process. Syst.* (2019).
52. Szegedy, Vanhoucke & Ioffe. Rethinking the inception architecture for computer vision. *Proc. Estonian Acad. Sci. Biol. Ecol.* (2016).
53. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* **10**, e0118432 (2015).

DISPLAY ITEMS

Table 1 | Table of baseline demographic and echocardiographic characteristics.

		Missing	Overall	1. Derivation (training & validation)	2. Internal testing	3. External testing
n			9124	6021	1063	2040
Year of study, n (%)	2016	0	1134 (12.4)	969 (16.1)	165 (15.5)	
	2017		1143 (12.5)	961 (16.0)	182 (17.1)	
	2018		1368 (15.0)	1167 (19.4)	201 (18.9)	
	2019		1894 (20.8)	1615 (26.8)	279 (26.2)	
	2020		1545 (16.9)	1309 (21.7)	236 (22.2)	
	2021		2040 (22.4)			2040 (100.0)
Age (years), mean (SD)		6	69.1 (16.0)	70.2 (15.7)	69.8 (15.8)	65.7 (16.4)
Gender, n (%)	Female	0	4468 (49.0)	2950 (49.0)	521 (49.0)	997 (48.9)
	Male		4656 (51.0)	3071 (51.0)	542 (51.0)	1043 (51.1)
Race, n (%)	Asian	6799	32 (1.4)	20 (1.2)	3 (1.1)	9 (2.0)
	Black		268 (11.5)	177 (11.0)	27 (10.2)	64 (14.2)
	Hispanic		118 (5.1)	79 (4.9)	15 (5.7)	24 (5.3)
	Other		37 (1.6)	28 (1.7)	3 (1.1)	6 (1.3)
	Unknown		18 (0.8)	14 (0.9)		4 (0.9)
	White		1852 (79.7)	1291 (80.2)	216 (81.8)	345 (76.3)
BMI (kg/m^2), mean (SD)		45	29.5 (16.3)	29.5 (19.3)	29.4 (8.2)	29.4 (7.3)
BP Systolic (mm Hg), mean (SD)		1054	132.0 (138.9)	133.0 (168.7)	130.6 (20.6)	129.8 (19.6)
BP Diastolic (mm Hg), mean (SD)		1062	73.6 (24.7)	73.5 (27.6)	73.1 (23.9)	74.4 (11.7)
LVIDd Index (cm/m^2), mean (SD)		1114	2.4 (0.4)	2.4 (0.4)	2.4 (0.4)	2.4 (0.4)
LA Vol Indexed. (cm2/m^2), mean (SD)		1092	36.1 (16.5)	37.1 (16.8)	37.5 (17.6)	32.3 (14.3)
RVSP (mmHg), mean (SD)		2512	32.3 (13.5)	32.9 (13.7)	33.0 (13.9)	29.8 (12.0)
EF (%), mean (SD)		108	59.4 (10.8)	59.5 (10.9)	59.0 (11.1)	59.1 (10.2)
AV continuity VTI (cm), mean (SD)		4718	1.4 (0.8)	1.3 (0.8)	1.3 (0.8)	2.1 (0.9)
AV mean gradient (mmHg), mean (SD)		3652	20.6 (17.8)	22.8 (18.2)	22.3 (18.1)	9.0 (9.4)
AV peak velocity (m/s), mean (SD)		443	2.2 (1.2)	2.4 (1.3)	2.4 (1.3)	1.6 (0.6)

Table 2 | Internal and external performance of an automated algorithm for detection of aortic stenosis.

	SSL	Kinetics-400	Random	Ensemble
Internal testing set				
AUROC	0.934 (0.920, 0.947)	0.938 (0.926, 0.950)	0.925 (0.912, 0.938)	**0.945 (0.933, 0.956)**
AUPR	0.818 (0.779, 0.854)	0.816 (0.774, 0.855)	0.749 (0.699, 0.799)	**0.827 (0.786, 0.864)**
F1 Score	0.764 (0.737, 0.799)	0.765 (0.740, 0.804)	0.742 (0.715, 0.781)	**0.785 (0.758, 0.819)**
PPV	0.674 (0.635, 0.749)	0.712 (0.661, 0.808)	0.643 (0.620, 0.730)	**0.702 (0.664, 0.766)**
Sensitivity	0.882 (0.797, 0.928)	0.825 (0.741, 0.899)	0.878 (0.782, 0.902)	**0.890 (0.828, 0.930)**
Specificity at 90% Sensitivity	0.860 (0.797, 0.884)	0.835 (0.777, 0.879)	0.814 (0.780, 0.862)	**0.880 (0.843, 0.901)**
PPV at 90% Sensitivity	0.661 (0.571, 0.709)	0.622 (0.543, 0.699)	0.594 (0.544, 0.670)	**0.694 (0.627, 0.741)**
External testing set				
AUROC	0.947 (0.884, 0.988)	0.954 (0.919, 0.984)	0.976 (0.966, 0.984)	**0.974 (0.957, 0.989)**
AUPR	0.337 (0.189, 0.559)	0.340 (0.171, 0.535)	0.238 (0.132, 0.407)	**0.365 (0.198, 0.563)**
F1 Score	0.458 (0.343, 0.640)	0.386 (0.299, 0.571)	0.377 (0.264, 0.538)	**0.439 (0.327, 0.615)**
PPV	0.393 (0.286, 0.875)	0.297 (0.217, 1.000)	0.303 (0.188, 0.714)	**0.429 (0.250, 0.750)**
Sensitivity	0.550 (0.304, 0.720)	0.550 (0.263, 0.773)	0.500 (0.263, 0.750)	**0.450 (0.316, 0.760)**
Specificity at 90% Sensitivity	0.947 (0.312, 0.973)	0.910 (0.674, 0.970)	0.948 (0.914, 0.962)	**0.958 (0.832, 0.973)**
PPV at 90% Sensitivity	0.143 (0.012, 0.247)	0.090 (0.024, 0.222)	0.146 (0.089, 0.218)	**0.175 (0.048, 0.273)**

Results come from a 3D-ResNet18 when initialized with the proposed self-supervised learning (SSL) initialization, a standard transfer learning approach (Kinetics-400), and a random weight initialization. "Ensemble" denotes an ensemble of the three individual models described above. Values in parentheses represent 95% confidence intervals determined by bootstrapping. AUROC = area under the receiver operating characteristic curve; AUPR = area under the precision-recall curve; PPV = positive predictive value.

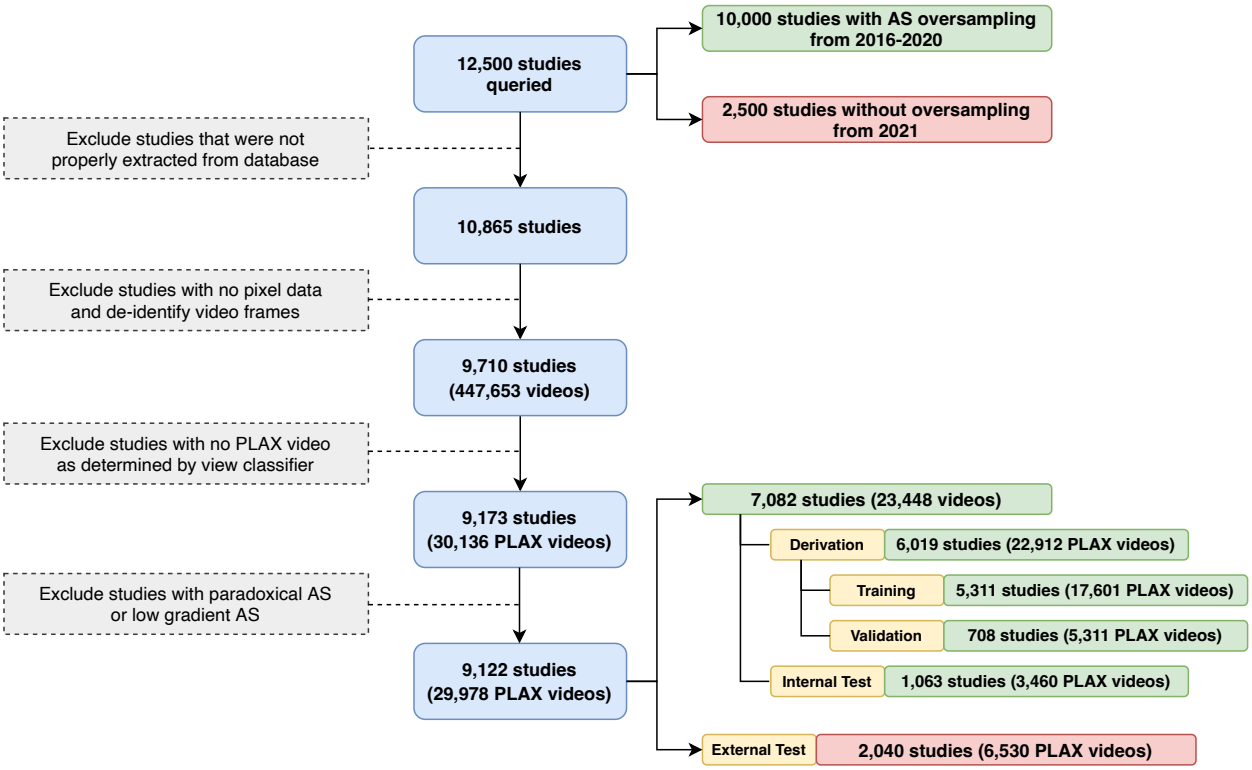

Figure 1 | Inclusion-exclusion flowchart for study population. Exclusion criteria for transthoracic echocardiogram (TTE) studies and videos included in this study. Studies with valid pixel data were de-identified frame by frame, and the parasternal long axis (PLAX) view was determined by an automated view classifier. A sample of 10,000 studies from 2016-2020 (with AS oversampled) were used for model development and internal testing, while a sample of 2,500 studies from 2021 (with no oversampling) were used as an "external" test set.

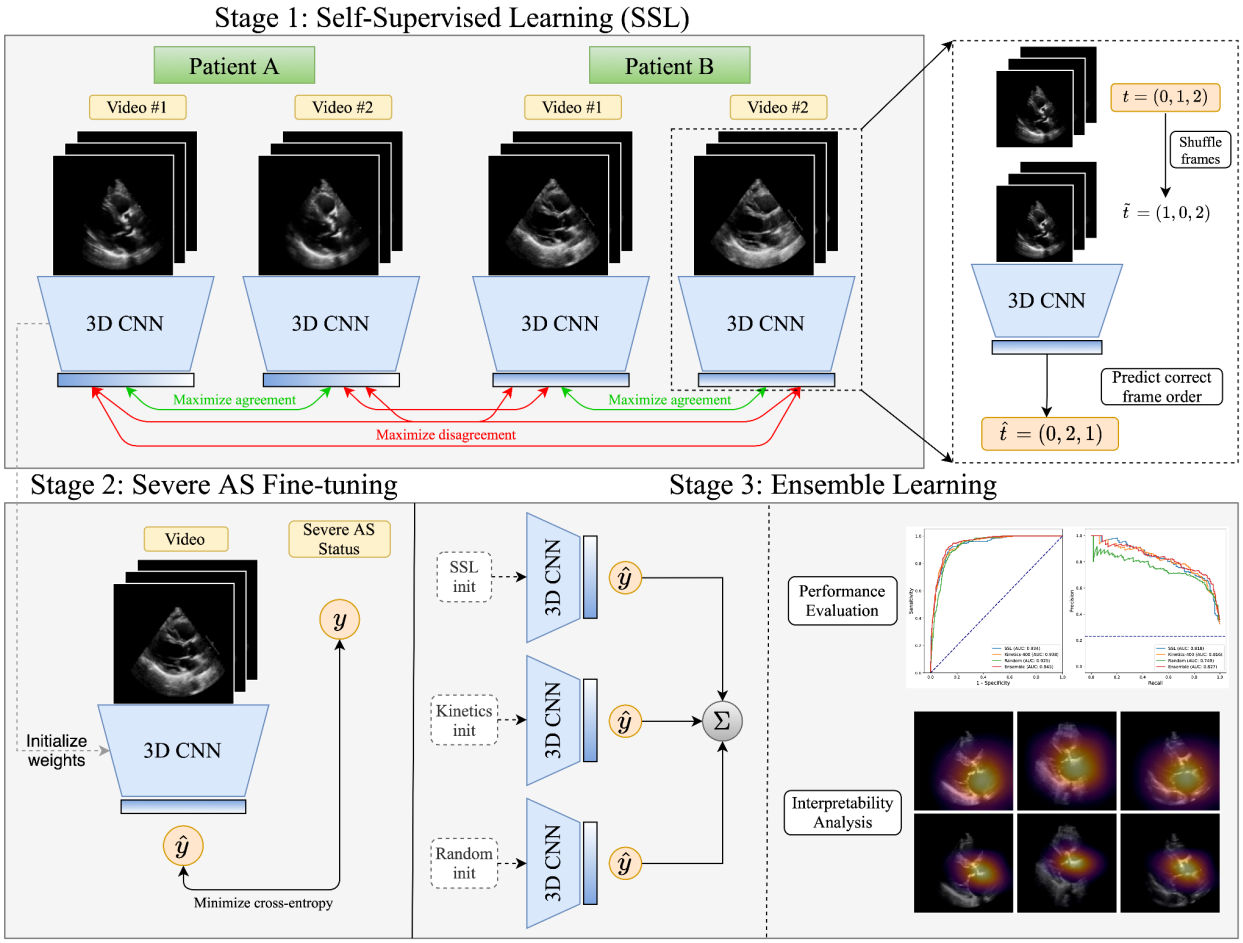

Figure 2 | Overview of proposed approach. We first perform self-supervised pretraining on parasternal long axis (PLAX) echocardiogram videos, selecting different PLAX videos from the same patient as "positive samples" for contrastive learning. After this representation learning step, we then use these learned weights as the initialization for a model that is fine-tuned to predict severe aortic stenosis (AS) in a supervised fashion.

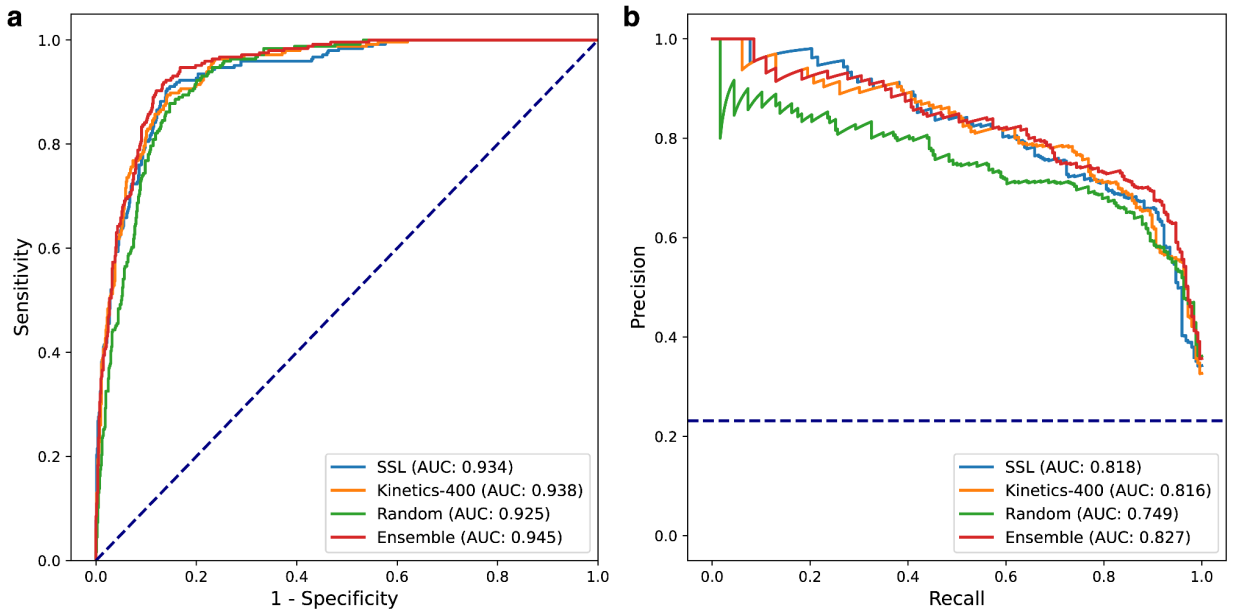

Figure 3 | Model performance in the internal testing set. Receiver operating characteristic (**a**) and precision-recall (**b**) curves for the proposed self-supervised learning approach, the standard transfer learning approach, and a model trained from scratch. AUC = area under the curve.

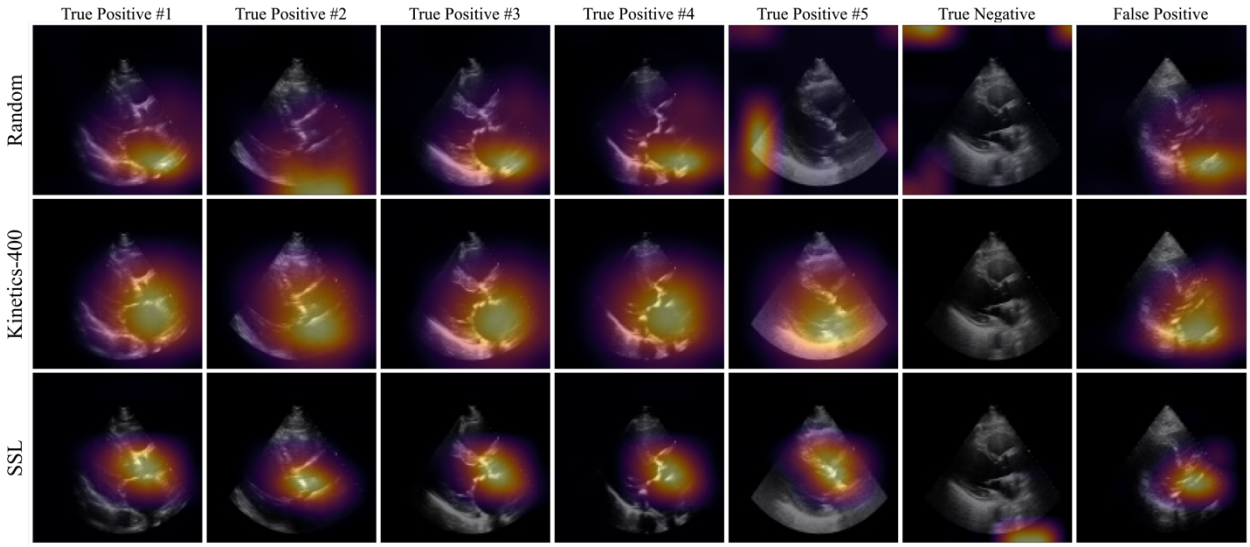

Figure 4 | Saliency map visualization. Spatial attention maps for the randomly initialized model (top row), Kinetics-pretrained model (middle row) and self-supervised learning (SSL) approach (bottom row) for five true positives (first five columns), a true negative (sixth column), and a false positive (last column). As determined by the Kinetics-pretrained model, the first five columns represent the five most confident severe AS predictions, the sixth column represents the most confident "normal" (no severe AS) prediction, and the seventh column represents the most confident *incorrect* severe AS prediction. Saliency maps were computed with the GradCAM method and reduced to a single 2D heatmap by maximum intensity projection along the temporal axis.

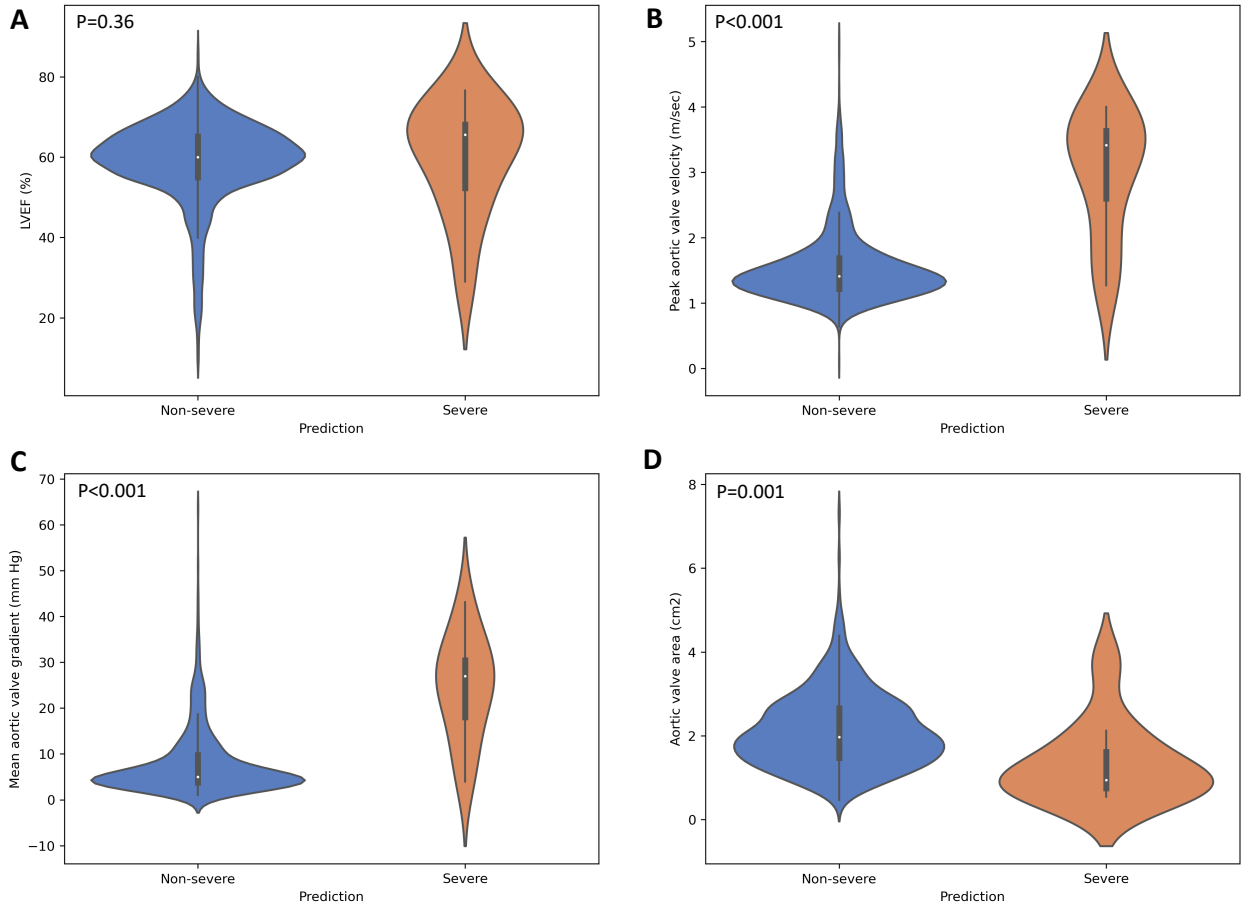

Figure 5 | Comparison between model predictions and echocardiographic left ventricular and aortic valve assessment among patients without severe aortic stenosis. Violin plots demonstrating the distribution of LVEF (left ventricular ejection fraction, **A**) peak aortic valve velocity (**B**), mean aortic valve gradient (**C**) and mean aortic valve area (**D**) for patients without severe AS, stratified based on the predicted class based on the final ensemble model.