Article type: Original Article

Artificial Intelligence-Enhanced Comprehensive Assessment of the Aortic Valve Stenosis Continuum in Echocardiography

Jiesuck Park,[a,b†] Jiyeon Kim,[c,d†] Jaeik Jeon,[e*] Yeonyee E. Yoon,[a,b,e*] Yeonggul Jang,[e] Hyunseok Jeong,[c,d] Youngtaek Hong,[c,e] Seung-Ah Lee,[c,e] Hong-Mi Choi,[a,b] In-Chang Hwang,[a,b] Goo-Yeong Cho,[a,b] Hyuk-Jae Chang[c,e,f]

[a]Cardiovascular Center and Division of Cardiology, Department of Internal Medicine, Seoul National University Bundang Hospital, Seongnam, Gyeonggi, Republic of Korea

[b]Department of Internal Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea

[c]CONNECT-AI Research Center, Yonsei University College of Medicine, Seoul, Republic of Korea

[d]Department of Internal Medicine, Graduate School of Medical Science, Brain Korea 21 Project, Yonsei University College of Medicine, Seoul, Republic of Korea

[e]Ontact Health Inc., Seoul, Republic of Korea

[f]Division of Cardiology, Severance Cardiovascular Hospital, Yonsei University College of Medicine, Yonsei University Health System, Seoul, Republic of Korea

†The first two authors contributed equally to this work.

Running title: AI for Aortic Stenosis Assessment in Echocardiography

***Corresponding authors:**

Yeonyee E. Yoon, MD, PhD

Professor, Division of Cardiology, Cardiovascular Center

Seoul National University Bundang Hospital

82, 173 Beon-gil, Gumi-ro, Bundang-gu, Seongnam-si, Gyeonggi-do, Republic of Korea

Phone: 82-31-787-7072

Email: yeonyeeyoon@snubh.org

And

Jaeik Jeon, MSc

Ontact Health Inc., Seoul, Republic of Korea

50-5, Ewhayeodae-gil, Seodaemun-gu, Seoul, Republic of Korea

Phone: 82-2-362-9610

E-mail: jaeikjeon9919@gmail.com

Abstract

Background: Transthoracic echocardiography (TTE) is the primary modality for diagnosing aortic valve stenosis (AVS), yet it requires skilled operators and can be resource-intensive. We developed and validated an artificial intelligence (AI)-based system for evaluating AVS that is effective in both resource-limited and advanced settings.

Methods: We created a dual-pathway AI system for AVS evaluation using a nationwide echocardiographic dataset (developmental dataset, n=8,427): 1) a deep learning (DL)-based AVS continuum assessment algorithm using limited 2D TTE videos, and 2) automating conventional AVS evaluation. We performed internal (internal test dataset [ITDS], n=841) and external validation (distinct hospital dataset [DHDS], n=1,696; temporally distinct dataset [TDDS], n=772) for diagnostic value across various stages of AVS and prognostic value for composite endpoints (cardiovascular death, heart failure, and aortic valve replacement)

Findings: The DL index for the AVS continuum (DLi-AVSc, range 0-100) increases with worsening AVS severity and demonstrated excellent discrimination for any AVS (AUC 0.91–0.99), significant AVS (0.95–0.98), and severe AVS (0.97–0.99). DLi-AVSc was independent predictor for composite endpoint (adjusted hazard ratios 2.19, 1.64, and 1.61 per 10-point increase in ITDS, DHDS, and TDDS, respectively). Automatic measurement of conventional AVS parameters demonstrated excellent correlation with manual measurement, resulting in high accuracy for AVS staging (98.2% for ITDS, 81.0% for DHDS, and 96.8% for TDDS) and comparable prognostic value to manually-derived parameters.

Interpretation: The AI-based system provides accurate and prognostically valuable AVS assessment, suitable for various clinical settings. Further validation studies are planned to confirm its effectiveness across diverse environments.

Funding: This work was supported by a grant from the Institute of Information & communications Technology Planning & Evaluation (IITP) funded by the Korea government (Ministry of Science and ICT) (No.2022000972, Development of a Flexible Mobile Healthcare Software Platform Using 5G MEC); and the Medical AI Clinic Program through the NIPA funded by the MSIT. (Grant No.: H0904-24-1002). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Keywords: Aortic valve stenosis; artificial intelligence; echocardiography; diagnostic accuracy; prognostic value.

Research in Context

Evidence before this study

We screened all English-based research articles in PubMed up to December 2023 using the keywords “artificial intelligence," "echocardiography," and "aortic valve stenosis." While some studies have used artificial intelligence (AI) to evaluate aortic valve stenosis (AVS) in echocardiography, these efforts were typically focused on either predicting significant AVS or automating conventional measurements, not both. For instance, Holste G. et al. trained a deep learning model on 5,257 studies and validated it using two external datasets (4,226 and 3,072 studies), achieving high accuracy in detecting severe AVS (area under the receiver operating characteristic curve (AUC): 0.942–0.952). However, their model was limited to the parasternal long-axis view and did not provide conventional quantitative analysis. In contrast, Krishna H. et al. automated conventional AVS evaluation, demonstrating that AI could accurately measure AVS parameters like aortic valve maximal velocity, mean pressure gradient, and aortic valve area in 256 patients, comparable to human measurements, but did not perform qualitative assessment of AVS. Furthermore, no studies investigated the prognostic value of AI-based AVS assessment.

Added value of this study

In this study, we developed a comprehensive AI-based system to evaluate AVS through a dual pathway: 1) assessing AVS presence and severity by deriving a DL index for the AVS continuum (DLi-AVSc) from parasternal long and/or short axis videos only, and 2)

automatically measuring AVS parameters and providing conventional quantitative AVS evaluation if additional images are available. The system was validated internally and in two independent external datasets, where DLi-AVSc increased with AVS severity and demonstrated excellent discrimination for any AVS (AUC 0.91–0.99), significant AVS (0.95–0.98), and severe AVS (0.97–0.99). Additionally, DLi-AVSc independently predicted adverse cardiovascular events. The automatic measurement of conventional AVS parameters showed a strong correlation with manual measurement, resulting in high accuracy for AVS staging (98.2% for internal test set, 81.0%, and 96.8% for external test sets) and offered prognostic value comparable to manually-derived parameters.

Implications of all the available evidence

AI-enhanced echocardiographic evaluation of AVS allows for accurate diagnosis of significant AVS and prediction of severity using only parasternal long or short axis views, typically obtained in the first step of echocardiographic evaluation. This capability can enhance AVS assessment in resource-limited settings and provide novices with guidance on when quantitative analysis is necessary. If additional views are acquired, the system automatically analyses them, enabling conventional quantitative evaluation, thereby saving time and effort while ensuring accurate assessment. Our study findings support the clinical implementation of AI-enhanced echocardiographic analysis.

1. Introduction

Medical advancements have significantly increased life expectancy, with about 10% of the global population over 60, projected to double by 2050.[1] This aging demographic notably increased the incidence of degenerative diseases like aortic valve stenosis (AVS). Studies revealed that 12·4% of individuals aged 75 and older have some degree of AVS, with severe cases at 3·4%.[2] Untreated AVS can cause irreversible myocardial damage, characterized by left ventricular hypertrophy, fibrosis, and functional impairment, leading to increased morbidity, mortality, and socioeconomic burden.[3] Therefore, timely detection and management of AVS are essential to mitigate its severe consequences.

Transthoracic echocardiography (TTE) is the primary imaging modality for assessing AVS. Accurate identification and staging of AVS via TTE require advanced expertise in scanning and interpretation, often unavailable in a general community healthcare setting. Even in tertiary care centres, the process is time-consuming and labour-intensive, involving multiple measurements, calculations, and precise interpretation. These complexities highlight the need for innovative solutions that simplify AVS assessment. Such solutions would be particularly beneficial in settings with limited resources by using fewer TTE videos and in more advanced settings by automating the measurement and interpretation processes.

To meet these clinical needs and advance beyond existing research,[4–6] we developed a comprehensive artificial intelligence (AI)-based system to evaluate AVS, suitable for both resource-limited and advanced settings. This system uses deep learning (DL) to diagnose and assess AVS from limited 2-dimensional (2D) TTE videos. Importantly, it does not merely classify the AVS severity but is designed to reflect the disease's progressive continuum.

Simultaneously, the system automatically measures a broad spectrum of structural and hemodynamic parameters, facilitating the conventional calculation of the aortic valve area (AVA) and providing a quantitative assessment of AVS. This paper describes the development process of our AI-based system and evaluates its diagnostic and prognostic potential in assessing AVS.

2. Methods

2.1. Study population and data sources

The AI-based frameworks utilized in this study were developed and validated using the Open AI Dataset Project (AI-Hub) dataset, an initiative supported by the South Korean government's Ministry of Science and ICT.[7] This dataset consists of 30,000 echocardiographic examinations retrospectively collected from five tertiary hospitals between 2012 and 2021, covering a wide range of cardiovascular diseases (***Supplemental Methods 1***). The AI-based frameworks introduced here were all developed using data extracted from the AI-Hub dataset.[8–10] To develop the DL-based AVS continuum assessment algorithm, a key focus of this study, we assembled the Development Dataset (DDS) by deliberately excluding Severance Hospital data among five hospitals. Instead, data from Severance Hospital were used exclusively for external validation (Distinct Hospital Dataset, DHDS). Further external validation was conducted using data collected from Seoul National University Bundang Hospital in 2022 (Temporally Distinct Dataset, TDDS). Detailed methodologies for data utilization in developing and validating the AI-based system are in Supplemental Methods 1. As a result, the DDS comprised TTE images from 8,427 patients, while the DHDS included 1,696 patients, and the TDDS included 772 patients. The study followed the Declaration of Helsinki (as revised in 2013). The institutional review board of each hospital approved this study and waived the requirement for informed consent because of the retrospective and observational nature of the study design. All clinical and echocardiographic data were fully anonymized before data analysis.

2.2. Echocardiogram acquisition and interpretation

All echocardiographic studies were conducted by trained echocardiographers or cardiologists and interpreted by board-certified cardiologists specialized in echocardiography, following recent guidelines[11,12] as part of routine clinical care. To reflect actual clinical practice, parameters were not re-measured for the study; instead, the values from the clinical reports were used as ground truth (GT) labels. AVS presence and severity were determined in the DDS using the standard clinical criteria to ensure appropriate training (***Table 1***).[11] In contrast, for the DHDS and TDDS, the prior clinician's decision regarding AVS severity in the clinical report was used as is, to better reflect actual clinical practice.

2.3. AI-based system

We have developed a fully automated AI-based framework that addresses AVS evaluation through the dual pathway, leveraging innovative and conventional methodologies. (***Central Illustration***) The operational sequence of this system begins by automatically selecting the necessary views, including the parasternal long-axis (PLAX), parasternal short-axis (PSAX) at the aortic valve (AV) level, AV continuous wave (CW) and pulsed wave (PW) Doppler, and left ventricular outflow tract (LVOT) PW Doppler. In the DL-based AVS continuum assessment pathway, the algorithm evaluates AVS using only the PLAX and/or PSAX videos. Concurrently, the DL segmentation network generates masks for each view in the automated conventional AVS assessment pathway. These masks facilitate the measurement of LVOT diameter from the PLAX view and analyse spectral Doppler images to ascertain key indicators such as AV peak velocity (V_{max}), AV velocity time integral (VTI), AV mean pressure gradient (mPG), and LVOT VTI. Then, the system calculates AVA, enabling quantitative evaluation of AVS. This dual

approach (DL-based AVS continuum assessment and automated conventional AVS assessment) can potentially support both resource-limited and advanced settings.

2.3.1. View classification

To assess AVS, we improved our preexisting view classification algorithm.[8] The algorithm could already identify the PLAX view, PSAX at the AV level, AV CW Doppler from apical views, AV PW Doppler, and LVOT PW Doppler. We augmented it to recognize the PLAX-AV zoomed views and the AV CW Doppler obtained from the right parasternal view. Detailed information about this development is in ***Supplemental Method 2***.

2.3.2. DL-based AVS continuum assessment algorithm

Our objective was to develop a network that classifies AVS severity in a way that reflects its continuum nature rather than just discrete categories. We used 3-dimensional (3D) convolutional neural networks (CNNs; r2plus1d18) as a backbone to separate spatial and temporal filters (***Supplemental Methods 3***).[13] This network processes TTE videos - PLAX and/or PSAX at the AV level - to output a score predicting the AVS severity, termed the DL index for the AVS continuum (DLi-AVSc).

To achieve accurate classification reflecting the AVS continuum, we implemented two novel strategies: 1) continuous mapping with ordered labels and 2) multi-task learning with auxiliary tasks that predict numeric parameters indicative of the AVS continuum, such as AV V_{max}, mPG, and AVA. Conventional multi-class classification with cross-entropy loss was

unsuitable for reflecting the AVS continuum as it fails to capture the disease's progressive nature due to equidistance between one-hot encoded severity levels. Instead, the continuous approach assigns each severity level a value between 0 and 1 (e.g., Normal: 0, Sclerosis: 0.25, Mild: 0.5, Moderate: 0.75, and Severe: 1) and trains the model by minimizing negative Bernoulli likelihood $L_{Bernoulli}$. While this method reflects AVS progression, it primarily converts discrete labels into continuous values. To truly capture the continuum and enable nuanced transitions within and between severity levels, we incorporated three auxiliary tasks predicting TTE parameters based solely on 2D TTE videos. These tasks, predicting V_{max}, mPG, and AVA, provide rich information content, allowing the network to learn anatomical features and the motion of the AV. The loss function for each auxiliary task is the mean squared error (MSE) between the predicted and actual TTE parameter values: $L_{MSE}^{V_{max}}, L_{MSE}^{mPG}, L_{MSE}^{AVA}$. Training the network to predict continuous TTE parameters allows it to capture both discrete transitions and subtle variations within each severity category. For instance, it can distinguish between cases classified as "moderate" closer to mild AVS and those nearing severe AVS. The combined loss function integrates the negative Bernoulli likelihood and the MSE losses for the auxiliary tasks $L_{combined} = L_{Bernoulli} + \lambda\left(L_{MSE}^{V_{max}} + L_{MSE}^{mPG} + L_{MSE}^{AVA}\right)$, where λ is a weighting parameter balancing the contributions of the classification and regression tasks. Detailed network configurations and implementation details are in ***Supplementary Methods 3***.

Finally, to determine the patient-level DLi-AVSc, if multiple PLAX or PSAX videos were available for a single patient, the DLi-AVSc was extracted from each video individually. Scores from PLAX and PSAX videos were averaged separately. If only one view type (either PLAX or PSAX) was available, its average score was used directly. If both views were

available, the final DLi-AVSc was calculated by averaging the scores from both PLAX and PSAX views.

2.3.3 Automated conventional AVS assessment algorithm

Our AI-based system also automates the conventional method to calculate AVA and assess AVS severity. Automating conventional AVA assessment in our system involves three key steps: 1) segmentation of anatomical structures and spectral Doppler envelopes, 2) uncertainty quantification to assess the confidence of the predicted segmentation masks, and 3) post-processing algorithms to extract clinical measurements from segmentation masks.

We had previously developed and validated algorithms for analysing spectral Doppler by segmenting the Doppler envelope to capture velocity profiles with essential topological features.[9,10] This approach automatically measures AV V_{max}, AV VTI, and LVOT VTI by segmenting Doppler envelopes in every analysable cycle in all provided images. In this study, to quantify AVA, we further developed a DL network based on the SegFormer transformer architecture to measure the LVOT diameter in the PLAX view.[14] This advanced model can segment all anatomical structures visible in the PLAX view, including the left ventricle (LV), LV septum and posterior wall, left atrium, right ventricle, aorta, and even the mitral valve and AV. Detailed information is provided in ***Supplemental Methods 4*** and ***Videos S1***.

Deep segmentation networks are highly effective due to their ability to learn complex patterns and features from large datasets. However, quantifying uncertainty in their predictions is crucial because segmentation errors can impact subsequent post-processing for automatic

measurement. To address this, we used predictive entropy from the segmentation network's probability map, which combines two sources of uncertainty: lack of knowledge in DL (epistemic uncertainty) and poor data quality (aleatoric uncertainty).[15] By evaluating the predictive entropy, cases requiring manual review due to poor image quality or model uncertainty can be identified. Detailed methodologies are provided in ***Supplemental Method 5*** and ***Videos S2***.

In the post-processing stage, the segmented masks were utilized to extract clinical measurements. From the predicted segmentation mask, we identified points where the mitral valve intersects with the aorta and where the septum intersects with the aorta to determine annulus points. Considering the differing opinions on the appropriate location for measuring the LVOT diameter,[16] our algorithm was designed to measure the LVOT diameter at three different locations: at the annulus, 2.5mm, and 5mm away from the annulus towards the LV cavity. In this study, the measurements taken at the annulus were used for analysis as they showed the highest agreement with the GT. For technical details and performance information, please refer to ***Supplemental Method 6*** and ***Video S1***.

For spectral Doppler images, AV V_{max} and VTI were derived from the segmented Doppler envelope of AV CW Doppler. This analysis included AV CW Doppler obtained from both the apical and right parasternal views, selecting the largest envelope across all cycles in all images to obtain AV V_{max} and VTI. The LVOT PW Doppler analysis also spanned all cycles, using the average value of LVOT VTI to avoid overestimating LVOT flow.[12] These measurements were then used to calculate mPG and AVA, which were used to assess the presence and severity of AVS.[11]

2.4 Ascertainment of clinical information and outcome definition

The clinical data were acquired by a dedicated review of the electronic health records at the study institutions. The clinical outcome was defined as a composite endpoint of cardiovascular death, hospitalization for heart failure, and AV replacement via surgical or transcatheter approaches.

2.5 Validation of AI-based AVS evaluation system and statistical analysis

The performance of each stage in our AI-based framework was validated using an internal test dataset (ITDS) and two external datasets (DHDS and TDDS). Additionally, we evaluated the execution time of each module, including the DLi-AVSc computation, PLAX auto-measurement, and Spectral Doppler auto-measurement, across 20 repeated runs. This analysis was conducted under the following conditions: OS Windows 10, CPU Intel i7-8565U @1.80GHz, Memory 16GB, and no GPU.

The view classification algorithm, the shared initial step, was evaluated against human expert labels. Precision, recall, and F1 scores were calculated for each view, with overall accuracy determined by the ratio of correctly classified images to the total number of images.

The performance of the DL-based AVS continuum assessment algorithm was evaluated by examining the distribution of the DLi-AVSc across various stages using violin plots. To verify that DLi-AVSc accurately reflects the continuum of AVS progression, we used Uniform Manifold Approximation and Projection (UMAP) to visualize this relationship,[17]

projecting the data into a 2D space, using 15 nearest neighbours, a minimum distance of 0.1, and the Euclidean distance. To highlight the areas with the greatest influence on the model's prediction, we generated saliency maps using the Gradient-weighted Class Activation Mapping (Grad-CAM).[18] We present representative samples for each severity level in both PLAX and PSAX views.

The conventional AVS assessment algorithm was validated by comparing AI-derived parameters with manual measurements. Since these parameters are not typically measured in normal or AV sclerosis groups, the comparison was limited to the AVS group. Moreover, as manual measurements were not always available for all AVS cases, details on GT measurements availability and the success rate of automatic measurements are provided in ***Supplemental Methods 7***. The association between automated and manual measurements was assessed using the Pearson Correlation Coefficient (PCC) and mean absolute error (MAE). The AVS severity determined from the automatic measurements was also compared to the ground truth label made by the clinician's prior decision.

We also evaluated the discrimination ability of the DLi-AVSc and other AI-derived conventional parameters for various stages of AVS, including mild or greater AVS (any AVS), moderate or greater AVS (significant AVS), and severe AVS. This evaluation was conducted through receiver operating characteristic (ROC) curve analysis, from which we calculated the area under the curve (AUC).

Lastly, we assessed the prognostic capability of AI-derived parameters for composite endpoints. Specifically, we conducted a spline curve analysis for our novel index, the DLi-AVSc, to visualize its predictive power. Additionally, we applied Cox regression analysis to

validate the prognostic relevance of the DLi-AVSc and other AI-derived AVS parameters, with adjustment for clinical risk factors (age, sex, body mass index, hypertension, and diabetes).

2.6 Role of the funders

The study was supported by a grant from the Institute of Information & communications Technology Planning & Evaluation (IITP) funded by the Korea government (Ministry of Science and ICT); and the Medical AI Clinic Program through the NIPA funded by the MSIT. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

3. Results

3.1 Baseline characteristics

The distribution of AVS severity across three datasets is shown in Table 1: ITDS (n=841), DHDS (n=1,696), and TDDS (n=772). ITDS and TDDS exhibited a higher prevalence of mild AVS (28% and 41%, respectively), with fewer moderate and severe cases. Conversely, DHDS displayed a more balanced severity distribution (12% mild, 15% moderate, and 12% severe, respectively). Baseline clinical characteristics are available in ***Supplemental Result 1***.

3.2 View classification

Our view classification algorithm accurately identified the required images for assessing AVS across all datasets. The overall accuracy rates were 99.6% for ITDS, 99.5% for DHDS, and 99.4% for TDDS. Detailed metrics are in ***Supplemental Result 2***.

3.3 Performance of DL-based AVS continuum assessment algorithm

The DLi-AVSc was calculated with an average processing time of less than 2 sec (1.8 ± 0.05 sec). The distribution of the DLi-AVSc, produced by the DL-based AVS continuum assessment algorithm, exhibited a consistent trend of increasing scores with the severity of AVS across all datasets (***Figure 1a***). Interestingly, at the AV sclerosis stage, the DLi-AVSc already significantly increased compared to the normal stage, indicating the algorithm's ability to detect early changes. When discordant cases excluded from the training dataset were included in the

ITDS, mild to moderate and low-flow, low-pressure gradient moderate AVS were distributed between mild and moderate AVS, while moderate to severe and low-flow, low-pressure gradient severe AVS were distributed between moderate and severe AVS (***Supplemental Results 3***). The DLi-AVSc demonstrated an increasing trend as conventional parameters assessing AVS severity, such as AV V_{max}, mPG, and AVA, worsened (***Supplemental Results 4***).

Furthermore, when we utilized UMAP to verify that the DLi-AVSc accurately represents the AVS continuum, the DLi-AVSc, derived from the approach incorporating both ordered labels and multi-task learning, displayed a distinct continuous gradient from normal through AV sclerosis to advancing AVS stages, consistently evident in ITDS and both external datasets (***Figure 1b***). In contrast, a conventional multi-class classification approach using 5-class cross-entropy loss resulted in the stage-based grouping but lacked the continuous progression seen in our approach. The continuous mapping with ordered labels approach, but without additional multi-task learning to predict key TTE parameters, appeared somewhat linear but did not accurately reflect the severity progression (***Supplemental Results 5***).

For each severity level, we present representative samples with Grad-CAM saliency maps overlaid on both PLAX and PSAX views, specifically localizing the AV (***Figure 2*** and ***Video S3***). These results demonstrate that our model accurately identifies the relevant regions for evaluating AVS across all severity levels and views without supervision.

3.4 Performance of automated conventional assessment algorithm

For conventional AVS evaluation, measurements of spectral Doppler, including AV CW

Doppler and LVOT PW Doppler, as well as LVOT diameter, are required. On average, each of these measurements took less than 0.5 sec (0.2 ± 0.02 sec) and less than 2.0 sec (1.4 ± 0.13 sec), respectively. Among AVS patients with available GT values, our algorithm successfully performed automatic measurements of AV Vmax (100% success rate) and mPG (99.3–100% success rate) (***Supplemental Methods 7***), showing high correlations with the GT values for AV V_{max} (PCC 0.974–0.991; MAE 0.08–0.14 m/s) and mPG (PCC 0.966–0.991; MAE 1.23–2.82 mmHg) (***Figure 3a***). Besides, the algorithm successfully measured the LVOT diameter from PLAX videos, demonstrating robust concordance with manual measurement (PCC 0.669–0.747; MAE 0.10–0.11 cm). The correlation for AVA, calculated from these measurements, was also good (PCC 0.744–0.819; MAE 0.18–0.18 cm^2) but relatively lower than V_{max} and mPG, due to its dependence on multiple measurements. Missing GT values resulted in fewer comparison cases (***Supplemental Methods 7***), and accumulated differences affected the overall accuracy. Nonetheless, the accuracy of AVS severity classification based on these automated measurements remained strong (98.2% for ITDS, 81.0% for DHDS, and 96.8% for TDDS) (***Figure 3b***).

3.5 Comparison of diagnostic performance of two different AI-based approach

The discrimination performance of DLi-AVSc for various stages of AVS was generally excellent: AUC 0.91–0.99 for any AVS, 0.95–0.98 for significant AVS, and 0.97–0.99 for severe AVS (***Figure 4***). When compared to automatically measured conventional parameters, in ITDS, the discrimination performance of DLi-AVSc was lower than that of automatically measured V_{max} and mPG but comparable to AVA. In DHDS, the performance of DLi-AVSc

surpassed AVA in diagnosing all stages of AVS, while in TDDS, it was comparable to AVA in diagnosing all stages of AVS.

3.6 Prognostic value of AI-based AVS assessment

Analysis of spline curves across the ITDS, DHDS, and TDDS showed that an increase in DLi-AVSc correlated with a rising risk of adverse clinical outcomes (***Figure 5***). The multivariable Cox regression analysis affirmed the strong and independent prognostic value of DLi-AVSc. A 10-point increase in DLi-AVSc from limited TTE videos was associated with hazard ratio (95% confidence interval) of 2.19 (1.77–2.71) in ITDS, and 1.64 (1.52–1.78) and 1.61 (1.31–1.99) in DHDS and TDDS, respectively (***Figure 6***). Moreover, the AI-derived parameters, such as V_{max}, mPG, and AVA, demonstrated prognostic values comparable to those of manually-derived parameters (***Figure 6***).

4. Discussion

We have developed and validated a comprehensive AI-based system to evaluate AVS through a dual pathway: 1) assessing the presence and severity of AVS using only the PLAX and/or PSAX videos typically acquired early during TTE, and 2) automatically analysing additional views for conventional quantitative AVS evaluation if obtained. This dual approach enables accurate AVS evaluation across various settings, with internal and external validation demonstrating excellent diagnostic accuracy and strong prognostic capabilities.

While our AI-based system is not the first to evaluate AVS, it stands apart from previous studies in several key aspects. First, our system provides both AVS evaluation using limited 2D TTE videos and automation of conventional measurements. Prior research has typically focused on one of these aspects. For instance, Krishna et al. developed an AI model to automate quantitative AVS evaluation.[6] However, their model did not include the crucial initial visual analysis of the AV from 2D TTE videos, which is essential for initiating conventional quantitative AVS analysis. Several studies used CNNs to extract AVS-related features from 2D TTE videos through end-to-end learning without requiring Doppler information.[5,7,19,20] Although these studies achieved decent performance in classifying AVS severity, they lack conventional evaluation of AVS, compromising trustworthiness, explainability, and interpretation. In contrast, our system is the first to integrate both approaches, identifying the potential for significant AVS using parasternal views typically acquired early in TTE, guiding the acquisition of additional images for conventional AVS evaluation, and providing automated analysis of these views. This approach not only predicts

AVS presence and severity from parasternal views, as human experts do, but also reduces workload by automating the subsequent conventional evaluation.

Another key strength of our study is that, unlike previous research, it reflects the continuous nature of AVS progression. For instance, Wessler et al. trained CNNs to classify AVS severity into three categories (no, early, and significant AVS) using limited 2D images.[7] Similarly, Ahmadi et al. proposed a transformer-based spatiotemporal architecture to classify AVS into four categories (normal, mild, moderate, and severe AVS) by capturing anatomical features and AV motion.[19] Vaseli et al. focused on model explainability in AVS severity classification, incorporating uncertainty estimation and classifying AVS severity into three classes (no, early, and significant AVS).[20] However, these classifiers discretize AVS severity, losing the continuum information of AVS. Recently, Holste et al. proposed a binary classifier based on the 3D-ResNet18 architecture to detect severe AVS, observing that model probabilities generated increase with AVS severity.[5] However, this model focused only on a binary classification task (e.g., non-severe vs. severe), not capturing the full range of AVS severity levels in the training stage. In contrast, our framework employs continuous mapping with ordered labels, providing a more nuanced representation of AVS severity. Importantly, we use multi-task learning with auxiliary tasks to predict continuous AVS-related TTE parameters. This approach not only transitions from discrete labels to continuous values but also captures the underlying continuum of the disease more effectively. In UMAP visualizations, our model demonstrates a clear continuous gradient from normal to severe AVS, unlike other classification models. Additionally, the appropriate distribution of DLi-AVSc in discordant cases further supports the performance of our framework.

The implications of our AI-based system extend beyond the reliable detection of significant AVS by providing guidance to less experienced operators and offering opportunities to correct potential errors they might introduce. For example, since the system can accurately predict the presence and severity of significant AVS using only PLAX or PSAX videos, it can guide operators to acquire additional necessary images, followed by automatic quantitative AVS evaluation. However, despite the accuracy of automated measurements demonstrated in this study, improper image acquisition can still hinder the correct assessment of AVS severity. If AV CW Doppler is not properly acquired, it may lead to AVS underestimation. Additionally, even if AV CW Doppler is properly acquired, inexperienced operators may struggle to accurately interpret low-flow, low-pressure gradient AVS. In such cases, a high DLi-AVSc could prompt a re-evaluation, accounting for potential errors in image acquisition or analysis.

Moreover, we found that the DLi-AVSc increases significantly from normal levels at AV sclerosis and mild AVS stages before significant AVS progression. To our knowledge, this is the first algorithm to achieve such performance. DLi-AVSc is poised to effectively monitor AVS progression from preclinical stages as a score-based tool. We anticipate the clinical utility of our system becoming prominent, especially as new pharmacological treatments are investigated for AVS prevention are explored.[21,22] If such treatments become available, our algorithm's sensitivity in detecting early AVS stages will be highly advantageous. For sure, further studies are needed to confirm whether DLi-AVSc consistently increases in tandem with AVS progression, which will be essential to expanding its clinical application.

Limitations

The present study has some limitations. Although we developed and thoroughly validated our AI-based system using data from multiple centres, including internal and external validation, all the data were obtained from tertiary centres in South Korea. This means that skilled operators acquired TTE, and it remains to be seen if the DLi-AVSc will perform well on TTE videos acquired in truly resource-limited and novice settings. However, this also suggests that if PLAX or PSAX videos are adequately acquired, the DLi-AVSc could potentially evaluates AVS as accurately as in advanced settings. Nevertheless, further evaluation is needed to confirm its performance in various clinical environments and among different populations. We are planning additional validation studies in primary clinics and a multi-national study to address these concerns. Additionally, although we designed the DLi-AVSc to reflect the AVS continuum, it needs to be verified whether the DLi-AVSc increases progressively with the natural progression of AVS. This issue will be addressed in future studies.

Conclusions

We developed and validated a comprehensive AI-based system for evaluating AVS. This system operates through a dual pathway: it assesses the presence and severity of AVS using limited TTE videos and simultaneously automates conventional quantitative AVS evaluation. Internal and external validations demonstrated excellent diagnostic accuracy and strong prognostic capabilities. While further validation in diverse clinical settings is necessary, our system is expected to enhance AVS detection and evaluation in resource-limited settings or by novices, while simultaneously reducing workload in advanced settings.

Contributors

All authors contributed equally to this study.

Declaration of Interests

Y.E.Y, J.J., Y.J., Y.H., and S.A.L. are currently affiliated with Ontact Health, Inc. J.J., J.K., and S.A.L are co-inventors on a patent related to this work filed by Ontact Health (Method For Providing Information On Severity Of Aortic Stenosis And Device Using The Same). H.J.C. holds stock in Ontact Health, Inc. The other authors have no conflicts of interest to declare.

Data Sharing Statement

The AI-Hub data may be accessible upon proper request and after approval of a proposal. Data from TDDS cannot be made publicly available due to ethical restrictions set by the IRB of the study institution; i.e., public availability would compromise patient confidentiality and participant privacy. Please contact the corresponding author (yeonyeeyoon@gmail.com) to request the minimal anonymized dataset.

References

1 Sixsmith A. Technology and the Challenge of Aging. In: Sixsmith A, Gutman G, editors. *Technologies for Active Aging*. Boston, MA: Springer US; 2013:7–25.

2 Osnabrugge RL, Mylotte D, Head SJ, et al. Aortic stenosis in the elderly: disease prevalence and number of candidates for transcatheter aortic valve replacement: a meta-analysis and modeling study. *J Am Coll Cardiol*. 2013;62:1002–1012.

3 Iung B, Arangalage D. Community burden of aortic valve disease. *Heart*. 2021;107:1446-1447.

4 Holste G, Oikonomou EK, Mortazavi BJ, et al. Severe aortic stenosis detection by deep learning applied to echocardiography. *Eur Heart J*. 2023;44:4592–4604.

5 Krishna H, Desai K, Slostad B, et al. Fully Automated Artificial Intelligence Assessment of Aortic Stenosis by Echocardiography. *J Am Soc Echocardiogr*. 2023;36:769–777.

6 Wessler BS, Huang Z, Long GM, Jr., et al. Automated Detection of Aortic Stenosis Using Machine Learning. *J Am Soc Echocardiogr*. 2023;36:411–420.

7 National Information Society Agency. Open AI Dataset Project (AI-Hub). https://aihub.or.kr/.

8 Jeon J, Ha S, Yoon Y, et al. Echocardiographic view classification with integrated out-of-distribution detection for enhanced automatic echocardiographic analysis. *arXiv preprint arXiv:2308.16483v1*. 2023.

9 Jeon J, Kim J, Jang Y, et al. A Unified Approach for Comprehensive Analysis of Various Spectral and Tissue Doppler Echocardiography. *arXiv preprint arXiv:2311.08439.*

2023.

10 Park J, Jeon J, Yoon YE, et al. Artificial intelligence-enhanced automation of left ventricular diastolic assessment: a pilot study for feasibility, diagnostic validation, and outcome prediction. *Cardiovasc Diagn Ther*. 2024;14:352–366.

11 Writing Committee M, Otto CM, 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. *J Am Coll Cardiol*. 2021;77:e25–e197.

12 Baumgartner H, Hung J, Bermejo J, et al. Recommendations on the Echocardiographic Assessment of Aortic Valve Stenosis: A Focused Update from the European Association of Cardiovascular Imaging and the American Society of Echocardiography. *J Am Soc Echocardiogr*. 2017;30:372–392.

13 Tran D, Wang H, Torresani L, Ray J, LeCun Y, Paluri M. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2018:6450–6459.

14 Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*. 2021;34:12077–12090.

15 Everett D, Nguyen AT, Richards LE, Raff E. Improving Out-of-Distribution Detection via Epistemic Uncertainty Adversarial Training. *arXiv preprint arXiv:220903148*. 2022.

16 Baumgartner HC, Hung JC-C, Bermejo J, et al. Recommendations on the echocardiographic assessment of aortic valve stenosis: a focused update from the European Association of Cardiovascular Imaging and the American Society of Echocardiography. *Eur*

Heart J Cardiovasc Imaging. 2017;18:254–275.

17 McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:180203426*. 2018.

18 Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*. 2020;128:336–359.

19 Ahmadi N, Tsang MY, Gu AN, Tsang TSM, Abolmaesumi P. Transformer-Based Spatio-Temporal Analysis for Classification of Aortic Stenosis Severity From Echocardiography Cine Series. *IEEE Trans Med Imaging*. 2024;43:366–376.

20 Vaseli H, Gu AN, Ahmadi Amiri SN, et al. ProtoASNet: Dynamic Prototypes for Inherently Interpretable and Uncertainty-Aware Aortic Stenosis Classification in Echocardiography. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer Nature Switzerland, 2023:368–378.

21 Ito S, Oh JK. Aortic Stenosis: New Insights in Diagnosis, Treatment, and Prevention. *Korean Circ J*. 2022;52:721–736.

22 Lindman BR, Sukul D, Dweck MR, et al. Evaluating Medical Therapy for Calcific Aortic Stenosis: JACC State-of-the-Art Review. *J Am Coll Cardiol*. 2021;78:2354–2376.

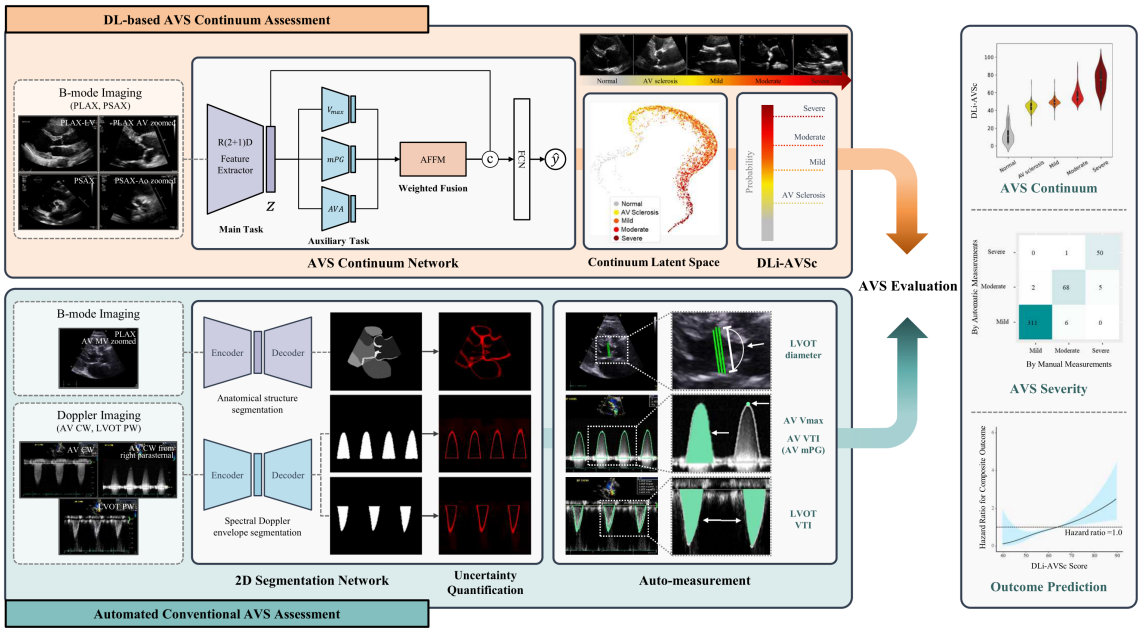
Table 1. Study Population

Classification	Definition	Developmental Dataset (DDS)			Distinct Hospital Dataset (DHDS)	Temporally Distinct Dataset (TDDS)
		Training	Validation	Internal Test Dataset (ITDS)		
Normal	Vmax <2m/s without AV calcification	2,627 (39%)	353 (41%)	310 (37%)	1,037 (61%)	55 (7%)
AV sclerosis	Vmax <2m/s with AV calcification	1,396 (21%)	148 (18%)	203 (24%)	0 (0%)	274 (35%)
Mild AVS	AV Vmax 2-2.9m/s or mPG<20mmHg with AVA >1.5cm2	2,000 (29%)	263 (31%)	237 (28%)	209 (12%)	313 (41%)
Moderate AVS	AV Vmax 3-3.9m/s or mPG 20-39mmHg with 1m2< AVA ≤1.5cm2	416 (6%)	50 (6%)	50 (6%)	251 (15%)	75 (10%)
Severe AVS	AV Vmax ≥4m/s or mPG ≥40mmHg with AVA ≤1cm2	304 (5%)	29 (3%)	41 (5%)	199 (12%)	55 (7%)
		6,743	843	841	1,696	772

Abbreviations: AV, aortic valve; AVA aortic valve area; AVS, aortic valve stenosis; mPG, mean pressure gradient; Vmax, peak aortic valve velocity.

Figure legends

Central Illustration



Central Illustration: AI-Enhanced Echocardiographic Assessment of AVS Continuum

The illustration depicts a dual-pathway AI system for evaluating AVS. The top row illustrates the DL-based assessment of the AVS continuum using limited views, providing a unique DL index for the AVS continuum, termed DLi-AVSc. The bottom row demonstrates the automated AVS assessment, which derives conventional echocardiographic AVS parameters. By integrating both pathways, our AI system enables accurate AVS diagnosis and prognostication, making it broadly applicable in advanced and resource-limited settings.

AVS, aortic valve stenosis; DL, deep-learning; DLi-AVSc, DL index for the AVS continuum.

Figure 1

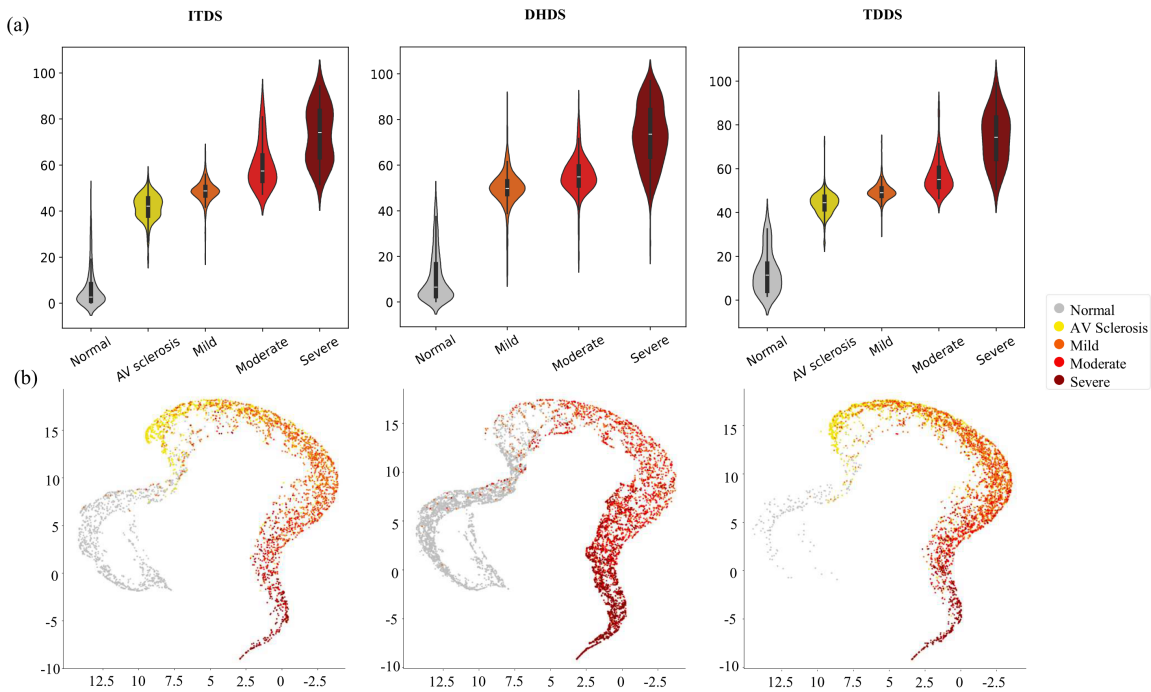


Figure 1. The Distribution of DLi-AVSc According to AVS Severity and UMAP Visualization

(a) The DLi-AVSc, generated by the DL-based AVS continuum algorithm, showed a consistent trend of increasing scores with the progression of AVS severity observed across both internal and external datasets. (b) The UMAP plot demonstrates a continuous nonlinear gradient transition from the normal state (grey) through AV sclerosis (yellow) to advanced AVS stages (red), visually underscoring the DLi-AVSc accurately representing the AVS continuum.

Abbreviations as in Central Illustration: DHDS, distinct hospital dataset; ITDS, internal test dataset; TDDS, temporally distinct dataset; UMAP, uniform manifold approximation and projection.

Figure 2

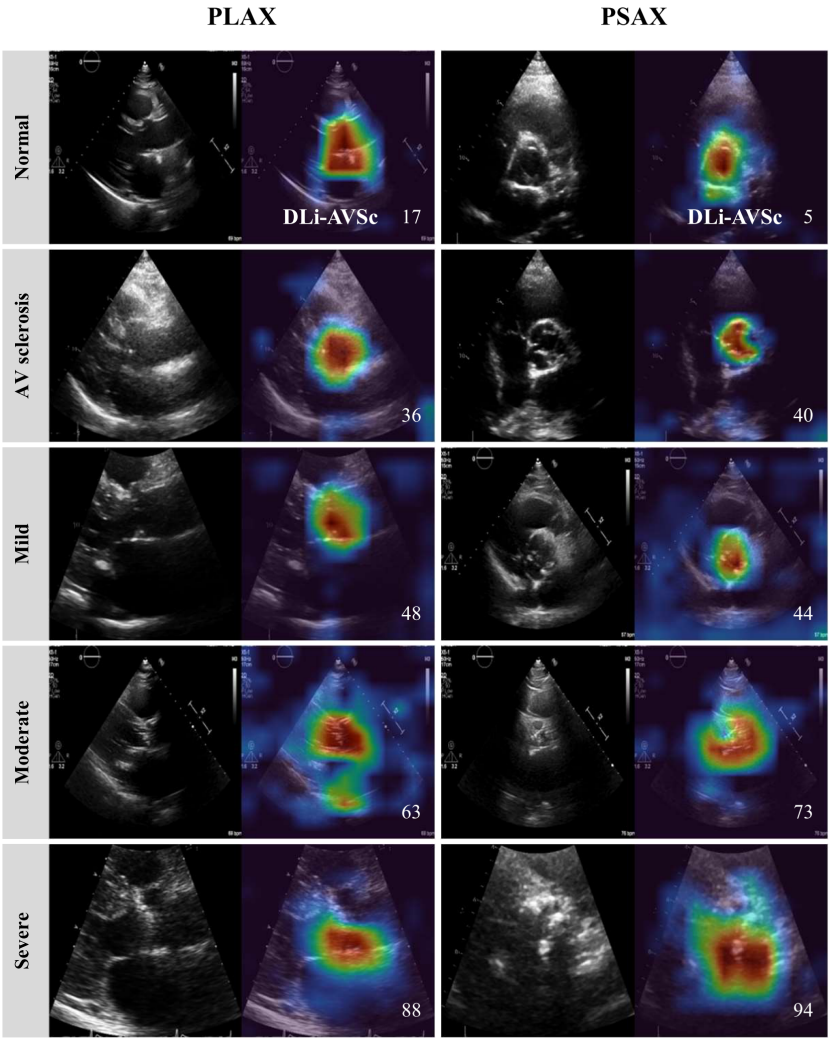


Figure 2. Explainability Analysis using Saliency Map

The figure displays representative PLAX and PSAX views alongside their corresponding Grad-CAM saliency maps during DLi-AVSc calculation. The saliency maps highlight that the DL model accurately focuses on the AV. As AVS severity progresses from normal to severe AVS, the model produces increasingly higher DLI-AVSc scores, corresponding worsening AVS.

Abbreviations as in Central Illustration: AV, aortic valve; PLAX parasternal long-axis view; PSAX, parasternal short-axis view.

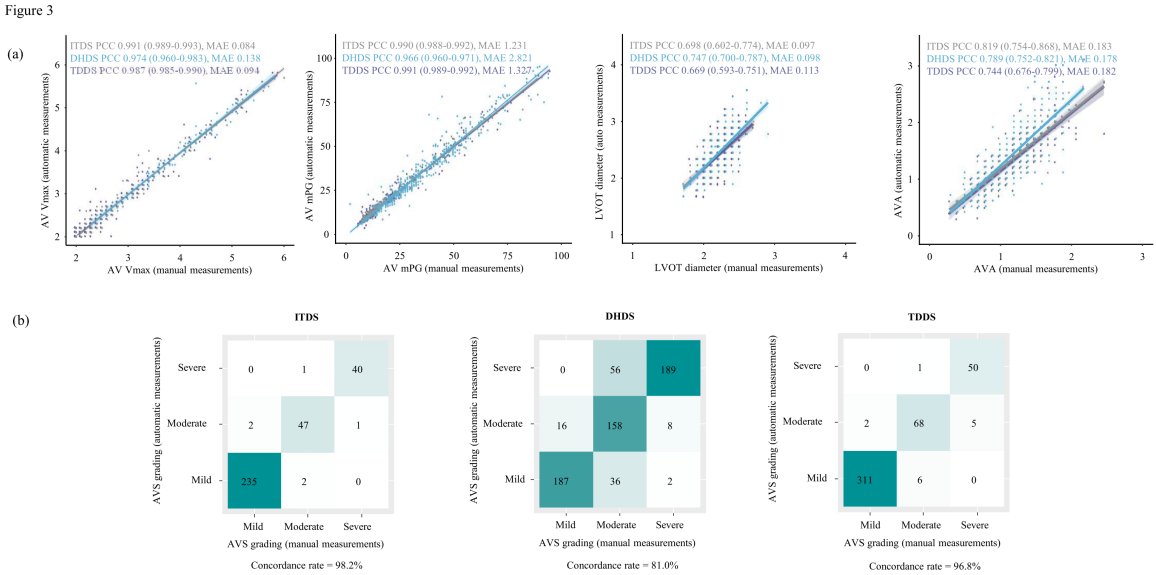


Figure 3. Concordance in AVS Diagnosis Between DL-based Automated Assessment and Conventional Evaluation

(a) Across all datasets, the auto-measured AVS parameters (AV maximal velocity, mean pressure gradient, and valve area) strongly correlated with those obtained from manual measurements. (b) Consequently, AVS gradings from both methods exhibited a high concordance rate, ranging from 81.0% to 96.8%.

Abbreviations as in Figure 1, 2 and 3: AVA, aortic valve area; LVOT, left ventricular outflow tract; MAE, mean absolute error; mPG, mean pressure gradient; PCC, Pearson Correlation Coefficient; V_{max}, maximal velocity.

Figure 4

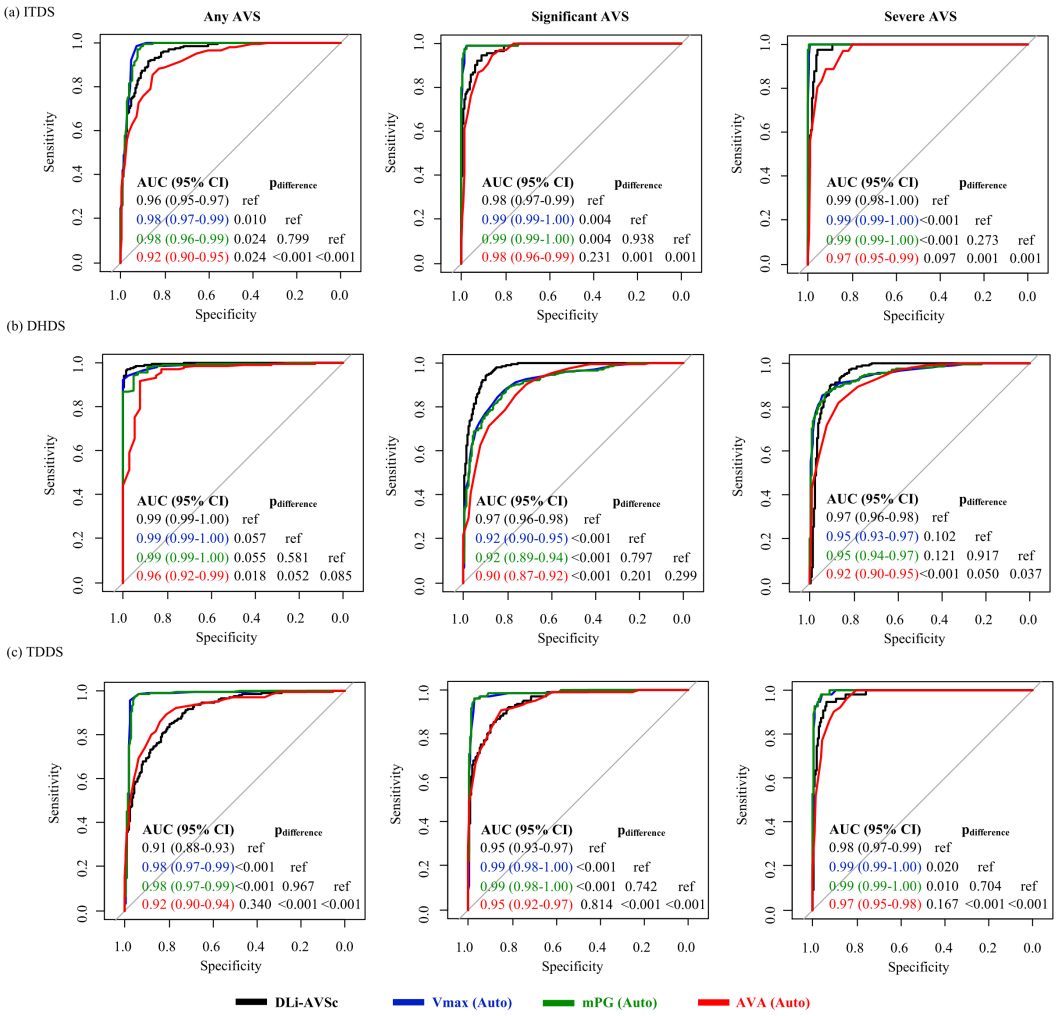


Figure 4. Diagnostic Performances of DLi-AVSc and Other AI-derived Conventional AVS Parameters Across Various Stages

The discriminative ability of DLi-AVSc and other conventional AVS parameters was consistently excellent for diagnosing any AVS, significant AVS (moderate to severe), and severe AVS across all datasets: (a) ITDS, (b) DHDS, and (c) TDDS.

Abbreviations as in Figures 1 and 3: AUC, the area under the curve; DLi-AVSc, DL index for the AVS continuum.

Figure 5

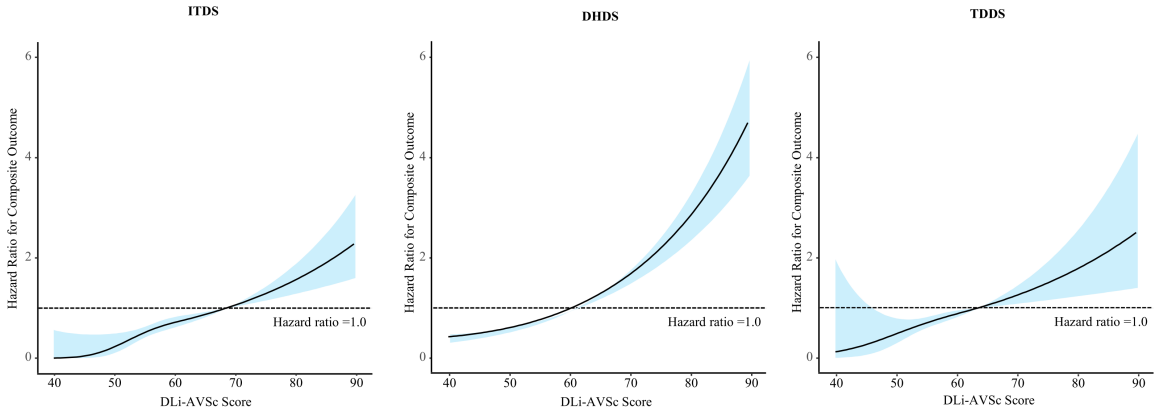


Figure 5. Spline Curves for Composite Outcomes Associated with DLi-AVSc

The risk of composite outcome gradually increased with higher DLi-AVSc across all datasets: (a) ITDS, (b) DHDS, and (c) TDDS. The solid lines represent the hazard ratio, and the blue shaded area represents the 95% confidence interval.

Abbreviations as in Figures 1 and 3.

Figure 6

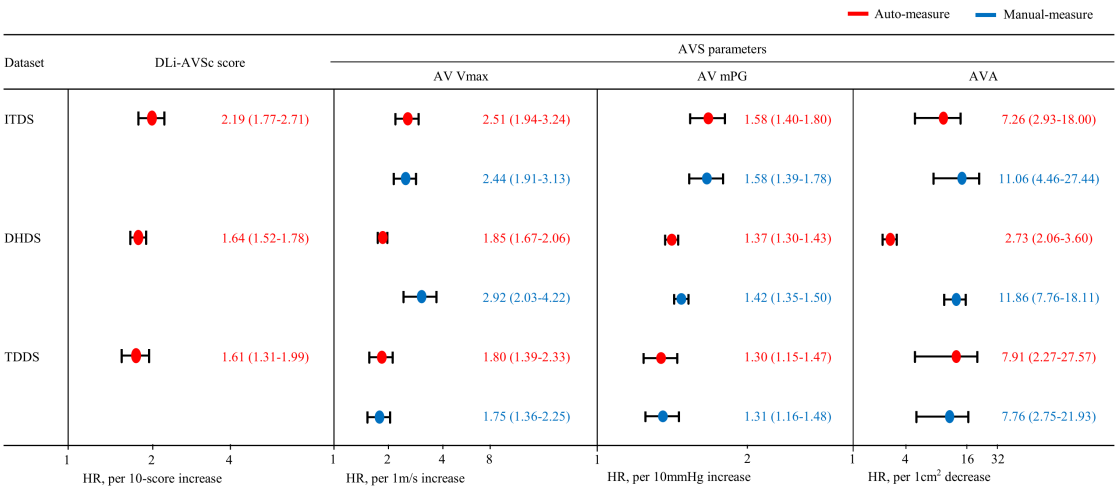


Figure 6. Prognostic value of DLi-AVSc and AVS parameters

The DLi-AVSc showed independent predictive value for composite outcomes. Similarly, other AI-derived AVS parameters were significant predictors for composite outcomes as well as manually-derived AVS parameters.

Abbreviations as in Figures 1, 2 and 3: HR, hazard ratio