European Heart Journal - Imaging Methods and Practice (2025) **3**, qyaf051
https://doi.org/10.1093/ehjimp/qyaf051

ORIGINAL ARTICLE

Machine learning-enabled screening for aortic stenosis with handheld ultrasound

Samuel Karmiy[1,†], Zhe Huang[2,†], Divya Velury[1], Eileen Mai[3], Jing Li[3], Monica M. Dehn[3], Dikran R. Balian[4], Davinder Ramsingh[5], John Martin[5], Jacob Kantrowitz[1], Ayan R. Patel[3], Michael C. Hughes[2], and Benjamin S. Wessler[3,6,*]

[1]Department of Medicine, Tufts Medical Center, Boston, MA, USA
[2]Department of Computer Science, Tufts University, Medford, MA, USA
[3]The Cardiovascular Center, Tufts Medical Center, Boston, MA, USA
[4]Tufts University School of Medicine, Boston, MA, USA
[5]Butterfly Inc., Burlington, MA, USA
[6]Predictive Analytics and Comparative Effectiveness Center (PACE), Tufts Medical Center, Boston, MA, USA

Received 22 December 2024; accepted after revision 27 March 2025; online publish-ahead-of-print 8 May 2025

Abstract

Aims

Neural network classifiers can detect aortic stenosis (AS) using limited cardiac ultrasound images. While networks perform very well using cart-based imaging, they have never been tested or fine-tuned for use with focused cardiac ultrasound (FoCUS) acquisitions obtained on handheld ultrasound devices.

Methods and results

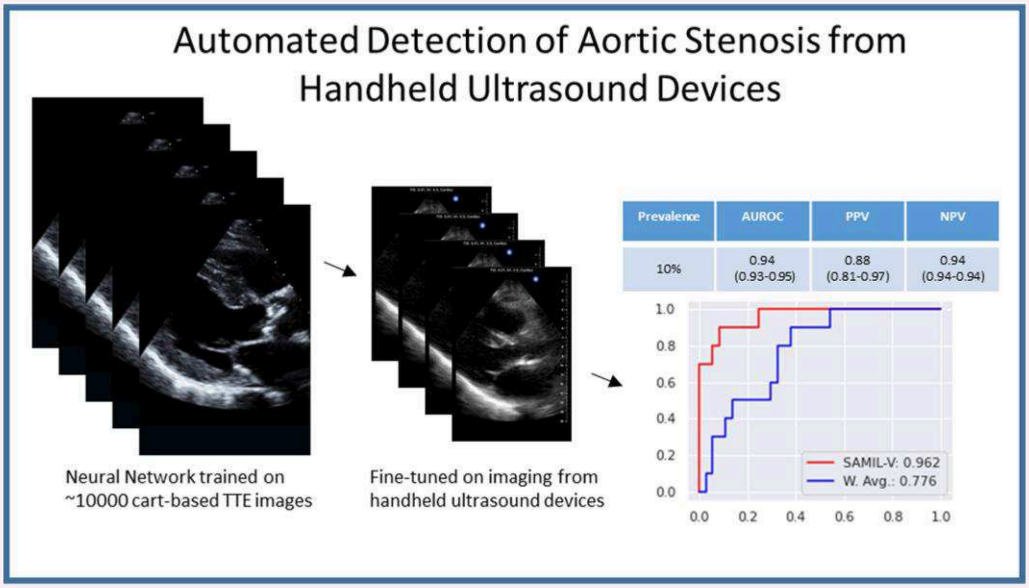
Prospective study performed at Tufts Medical Center. All patients ≥65 years of age referred for clinically indicated transthoracic echocardigraphy (TTE) were eligible for inclusion. Parasternal long axis and parasternal short axis imaging was acquired using a commercially available handheld ultrasound device. Our cart-based AS classifier (trained on ~10 000 images) was tested on FoCUS imaging from 160 patients. The median age was 74 (inter-quartile range 69–80) years, 50% of patients were women. Thirty patients (18.8%) had some degree of AS. The area under the received operator curve (AUROC) of the cart-based model for detecting AS was 0.87 (95% CI 0.75–0.99) on the FoCUS test set. Last-layer fine-tuning on handheld data established a classifier with AUROC of 0.94 (0.91–0.97). AUROC during temporal external validation was 0.97 (95% CI 0.89–1.0). When performance of the fine-tuned AS classifier was modelled on potential screening environments (2 and 10% AS prevalence), the positive predictive value ranged from 0.72 (0.69–0.76) to 0.88 (0.81–0.97) and negative predictive value ranged from 0.94 (0.94–0.94) to 0.99 (0.99–0.99) respectively.

Conclusion

Our cart-based machine-learning model for AS showed a drop in performance when tested on handheld ultrasound imaging collected by sonographers. Fine-tuning the AS classifier improved performance and demonstrates potential as a novel approach to detecting AS through automated interpretation of handheld imaging.

* Corresponding author. E-mail: bwessler@tuftsmedicalcenter.org
† First Author contribution.

Graphical abstract



Automated Detection of Aortic Stenosis from Handheld Ultrasound Devices

Prevalence	AUROC	PPV	NPV
10%	0.94 (0.93-0.95)	0.88 (0.81-0.97)	0.94 (0.94-0.94)

Neural Network trained on ~10000 cart-based TTE images

Fine-tuned on imaging from handheld ultrasound devices

Previously developed machine-learning aortic stenosis classifier was fine-tuned to work with handheld ultrasound imaging. The new optimized classifier can enable AS detection upstream of traditional echocardiography laboratories.

Keywords aortic stenosis • echocardiography • diagnosis • machine learning

Introduction

Aortic stenosis (AS) is a major public health problem that is under-diagnosed and under-treated.[1] Cardiac auscultation, the current approach to initial detection of AS, is limited by poor sensitivity and specificity.[2] The symptoms of AS are often non-specific, mortality risk might increase earlier than previously recognized,[3] and many patients present late in the disease course[4] leading to worse outcomes after treatment.[5] With the widespread availability of effective treatments for AS[6] and emergence of trials of medical therapies to halt progression,[7] better methods of diagnosing AS are needed.

There is substantial research focused on developing novel methods to improve the diagnosis of AS.[8] Our group has previously developed a machine-learning (ML) classifier to detect AS based on limited cardiac ultrasound images.[9] One use case of our AS classifier (as well as others[10]) is to process handheld ultrasound imaging upstream of traditional echocardiography laboratories. While these classifiers excel at classification tasks using cart-based imaging, they have never been studied using imaging from handheld ultrasound devices.

Model performance often degrades with data shifts[11] and model transportability must be rigorously assessed.[12] For the AS classification task, existing neural networks have not been assessed on focused cardiac ultrasound (FoCUS) that is proposed as the intended use case. Here, we study the transportability of an automated AS classifier to FoCUS acquisitions from a commercially available handheld ultrasound device.

Methods

Study design

Prospective study performed from January 2023 to February 2024 at Tufts Medical Center Cardiovascular Imaging and Hemodynamic Laboratory. This study was approved by the institutional review board at Tufts Medical Center. All imaging [comprehensive transthoracic echocardiography (TTE) and FoCUS using a handheld device] was performed by trained, American Registry for Diagnostic Medical Sonography certified cardiac sonographers from 2023 to 2024.

Inclusion criteria

All patients ≥ 65 years of age referred for clinically indicated TTE were eligible for inclusion. Participants were not selected for inclusion based on image quality and represent an unselected population.

Image acquisition

Participants in the study underwent clinically indicated TTE by trained cardiac sonographers as part of routine care. Following the TTE, FoCUS that included parasternal long axis (PLAX) and parasternal short axis (PSAX) imaging was obtained using the Butterfly IQ + (Butterfly Network Inc., Burlington, MA) handheld device. Handheld imaging was also done by a trained sonographer.

Diagnostic labels

The reference AS grade (none, mild, moderate, and severe) for this study was extracted directly from the clinical interpretation of the comprehensive TTE done during the same encounter. TTE imaging was interpreted by experienced echocardiographers in a manner consistent with current guidelines.[13]

Data pre-processing

FoCUS were downloaded from the Butterfly Network Cloud. Following the pre-processing steps, we extracted and standardized the frames of each video to a resolution of 112 × 112 pixels. For this study, the classification task of interest was the binary task of distinguishing between ‘any AS’ (including mild, moderate, or severe AS) vs. ‘no AS’ to provide the most useful output for a screening environment (i.e. to identify patients who should be referred for comprehensive TTE).

Baseline classifier

As previously described,[9] we developed a weighted averaging classifier for AS with two components: a view classifier and a diagnosis classifier. This classifier is intended to make study-level AS diagnoses given all 2D images from a routine cart-based patient scan, without needing to filter by view type. Each 2D image is fed into the view classifier and the diagnosis classifier. The view classifier provides a probability that each image shows a relevant view (PLAX or PSAX) of the aortic valve that can be used for assessing AS. For each image the diagnostic classifier outputs a probability distribution over three possible severity levels: 'no AS', 'early AS' (comprising mild and mild-moderate AS), and 'significant AS' (comprising moderate, moderate-severe, and severe AS). Following our prior work, we employ prioritized view weighting to make 'study-level' diagnosis predictions. Specifically, we calculate a weighted average of the AS severity level probability vectors across all images in a study, with weights determined by the relevant view probability from the view classifier.

The view and diagnosis classifiers each use a wide residual network backbone with 28 layers containing 5 931 683 parameters.[14] As in prior work, we obtained three instances of each classifier, each trained on one of the three train/test splits of our original cart-based dataset. For each split, neural networks were trained on data from 338 patients (average $n = 10\,253$ labelled images), validated on 119 patients (average $n = 3505$ labelled images), and performance was assessed on a test set of 120 patients (average $n = 3511$ labelled images).

View classifiers were trained to minimize a five-class cross entropy summed over all view-labelled images in the labelled set. Diagnosis classifiers were trained via multitask training, in which the loss function includes both a primary three-class cross entropy for AS severity level and an auxiliary five-class cross entropy for view. Each model was trained via stochastic gradient descent until the validation balanced accuracy for its primary task did not improve for at least 30 epochs.

Training on the established three data splits yields three model instances (pairs of view and diagnosis classifiers). In this work, we derive the final diagnosis prediction for a new study via an ensemble approach, averaging the study-level diagnosis predictions generated by the three model instances.

Advanced classifier via supervised attention multiple-instance learning

Our group has developed improved neural network architectures for predicting study-level AS severity[15,16] on cart-based imaging. Unlike the simplistic weighted average approach described above, these new methods make use of modern attention-based multiple-instance learning (MIL) to more flexibly combine many 2D images or videos and make one coherent study-level prediction. We call our architecture SAMIL-V (supervised attention MIL),[15] using the -V suffix to indicate use of 32-frame videos rather than single frame images. We pre-trained this SAMIL-V architecture in semi-supervised fashion[16] on the provided three train/test splits of our cart-based data.

Last-layer fine-tuning on handheld imaging

After pre-training on cart-based data, we adapt each SAMIL-V neural net classifier to handheld ultrasound imaging. For simplicity, we treated all video encoders and intermediate layers as frozen. Only the last layer's parameters were updated. That last layer consumes a study-level embedding vector and predicts AS severity levels via a linear-softmax transformation. The weight and bias parameters of this last layer were initialized to the optimal values from cart-based training, then fine-tuned to minimize 3-class cross entropy on the handheld data while applying L2 regularization to last layer weights. The L2 penalty strength hyperparameter was tuned via grid search to maximize an estimate of held-out area under the received operator curve (AUROC) from four-fold cross-validation on available training data.

Statistical analysis

To obtain robust estimates of performance, we repeat all evaluations over five separate train/test splits of the 160 FoCUS scans in our dataset. Each split is drawn randomly and independently (not mutually exclusive). To create each split, we stepped through each AS severity level (none, early, significant) in turn and randomly assigned FoCUS patient-scans with that severity into train and test sets. This class-stratified strategy ensures the overall distribution of AS severity levels remains similar across the train and test sets. Each split's test set contains roughly 50 scans (37-42 'no AS' and 10 'any AS'), with remaining scans in training set. Reported performance metrics represent an average over the test sets from five separate data splits.

The performance of the automated AS networks on FoCUS was assessed by AUROC as the primary metric. We also present area under the precision recall curve (AUPRC). Performance of the model was tested for the binary classification of 'any AS' vs. 'no AS'. To adapt pre-trained models to this task, we aggregated predicted probabilities of finer-scale severity levels when needed.

Confusion matrices were created, and we report sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). To model performance in a screening environment we tested network performance at 2 and 10% AS prevalence rates.

External validation

A temporal external validation was done for our top-performing model on a distinct handheld FoCUS cohort of 41 scans acquired at Tufts Medical Center from April 2024 to August 2024.

Results

Baseline characteristics

One hundred sixty patients were enrolled in the study. Baseline characteristics of the patients are shown in *Table 1*. The median age was 74 (inter-quartile range (IQR) 69–80) years and 50% of the patients were female. One hundred sixteen patients (72.5%) self-identified as

Table 1 Baseline characteristics

Patient characteristic	Value
Age (years)	74 (69–80)
Sex (female)	50%
Race	
White	73%
Black	9%
Latino	1%
Other	18%
Height (cm)	168 (160–175)
Weight (kg)	76 (64–89)
BMI	26.6 (23.7–30.3)
Systolic BP (mmHg)	132 (121–148)
Diastolic BP (mmHg)	73 (65–81)
Other conditions	
Hypertension	84%
Hyperlipidemia	84%
Congestive heart failure	36%
Diabetes	25%
Prior MI	16%
Prior PCI	11%
Prior CABG	8%
Prior CVA	9%
Prior MV replacement	3%

Baseline characteristics of patients scanned with the Butterfly IQ + device. All values reported as median (IQR) unless otherwise specified.

BMI, body mass index; MI, myocardial infarction; PCI, percutaneous coronary intervention; CABG, coronary artery bypass grafting; CVA, cerebrovascular accident; MV, mitral valve.

'white'. Twenty patients (11.1%) had a left ventricular ejection fraction (EF) ≤ 40%. In the entire cohort, 30 patients (18.8%) had some degree of AS. Of those, 6 (3.8%) had 'early AS' and 24 (15%) had 'significant AS'. The aortic valve hemodynamic measurements of patients with AS are shown in *Table 2*. Seventeen patients (10.6%) had a prosthetic aortic valve; these patients were excluded from the analysis.

Baseline model performance on handheld ultrasound imaging

For the task of classifying no AS vs. any degree of AS, the cart-based weighted average classifier[9] had an AUROC of 0.87 (95% CI 0.75–0.99) on handheld imaging, as shown in *Figure 1*. This represents a 20% drop in discrimination compared with that network's performance on cart-based imaging test sets.[9] In the handheld imaging cohort, the sensitivity for detecting any degree of AS was 1.00 (1.00–1.00), the specificity was 0.22 (0.08–0.36), the PPV was 0.25 (0.21–0.29), and NPV was 1.00 (1.00–1.00) (*Table 3*).

Fine-tuned model performance on handheld ultrasound imaging

For the last-layer fine-tuned SAMIL-V classifier, the AUROC for differentiating between no AS and any degree of AS on handheld ultrasound imaging was 0.94 (0.91–0.97) as shown in *Figure 1*. The sensitivity for detecting any degree of AS was 0.42 (0.30–0.54), the specificity was 0.99 (0.98–1.0), the PPV was 0.93 (0.84–1.0), and NPV was 0.87 (0.85–0.89).

Table 2 Aortic valve hemodynamics

Imaging parameter	Early AS (*n* = 6)	Significant AS (*n* = 24)	All AS (*n* = 30)
LVEF (%)	56.0 (43.0–60.0)	62.5 (60.0–67.5)	60 (55.0–65.0)
SV index (mL/m^2)	37.6 (33.8–45.5)	36.9 (31.8–46.1)	38.0 (29.8–46.5)
Valve area (AVA, cm^2)	1.4 (1.3–1.4)	0.9 (0.8–1.2	1.1 (0.8–1.3)
V2 max (m/s)	2.4 (2.1–2.7)	3.4 (3.0–3.8)	2.8 (2.1–3.3)
AV mean gradient (mmHg)	9.4 (7.2–13.3)	27 (19.4–38.0)	18 (11–28)
Dimensionless index (VTI)	0.46 (0.41-0.50)	0.26 (0.22–0.31)	0.36 (0.25–0.50)
Dimensionless index (Vmax)	0.43 (0.43–0.47)	0.28 (0.23–0.32)	0.35 (0.28–0.45)

Hemodynamic parameters for patients with AS imaged with the handheld ultrasound device. Early AS represents progressive AS (mild, mild/mod), Significant AS includes moderate and severe AS.

LVEF, left ventricular ejection fraction; SV, stroke volume; V2, continuous wave Doppler peak velocity; AV, mean gradient is the aortic valve mean gradient; VTI, velocity time integral.

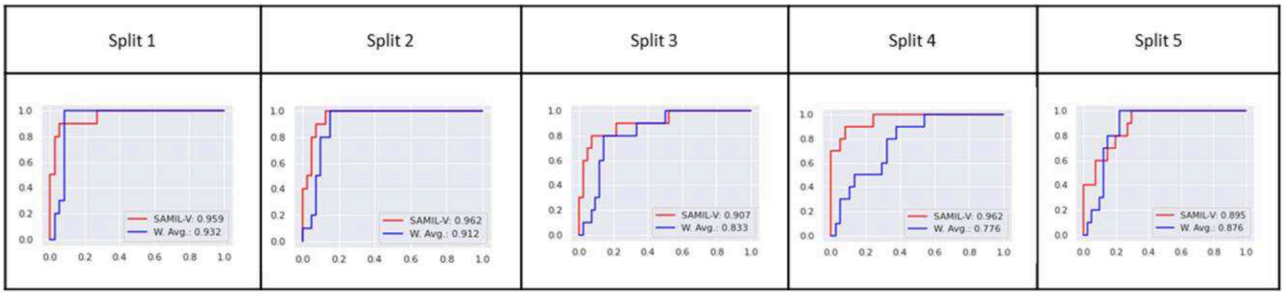


Figure 1 Receiver operator curves differentiating between no AS and any AS. Receiver operator curves in the test set. To obtain robust estimates of performance, we repeat all evaluations over five separate train/test splits of the 160 FoCUS scans in our dataset. Each split is drawn randomly and independently. SAMIL-V is the classifier fine-tuned using handheld imaging. W.Avg is the original cart-based classifier.

Table 3 Validation AUROC on handheld imaging and cart-based imaging

Architecture	Pre-trained	Fine-tuning	AUROC on FOCUS	AUPRC
Baseline (cart-based)	Yes: TMED-2	None	0.87 (0.06)	0.49 (0.09)
SAMIL-V	Yes: TMED-2 with SSL	Last-layer	**0.94** (0.03)	**0.83** (0.07)

Summary performance metrics of the original cart-based network and the model that is fine-tuned for handheld ultrasound imaging. Performance metrics are presented as the mean (SD) values over the tested data splits. The bold values represent the top performing models.

TMED-2, Tufts Medical Echocardiogram Database of cart-based echocardiograms; SSL, semi-supervised learning; AUROC, area under the receiver operator curve; FoCUS, focused cardiac imaging; AUPRC, area under the precision recall curve.

Fine-tuned model performance on held-out temporal validation cohort

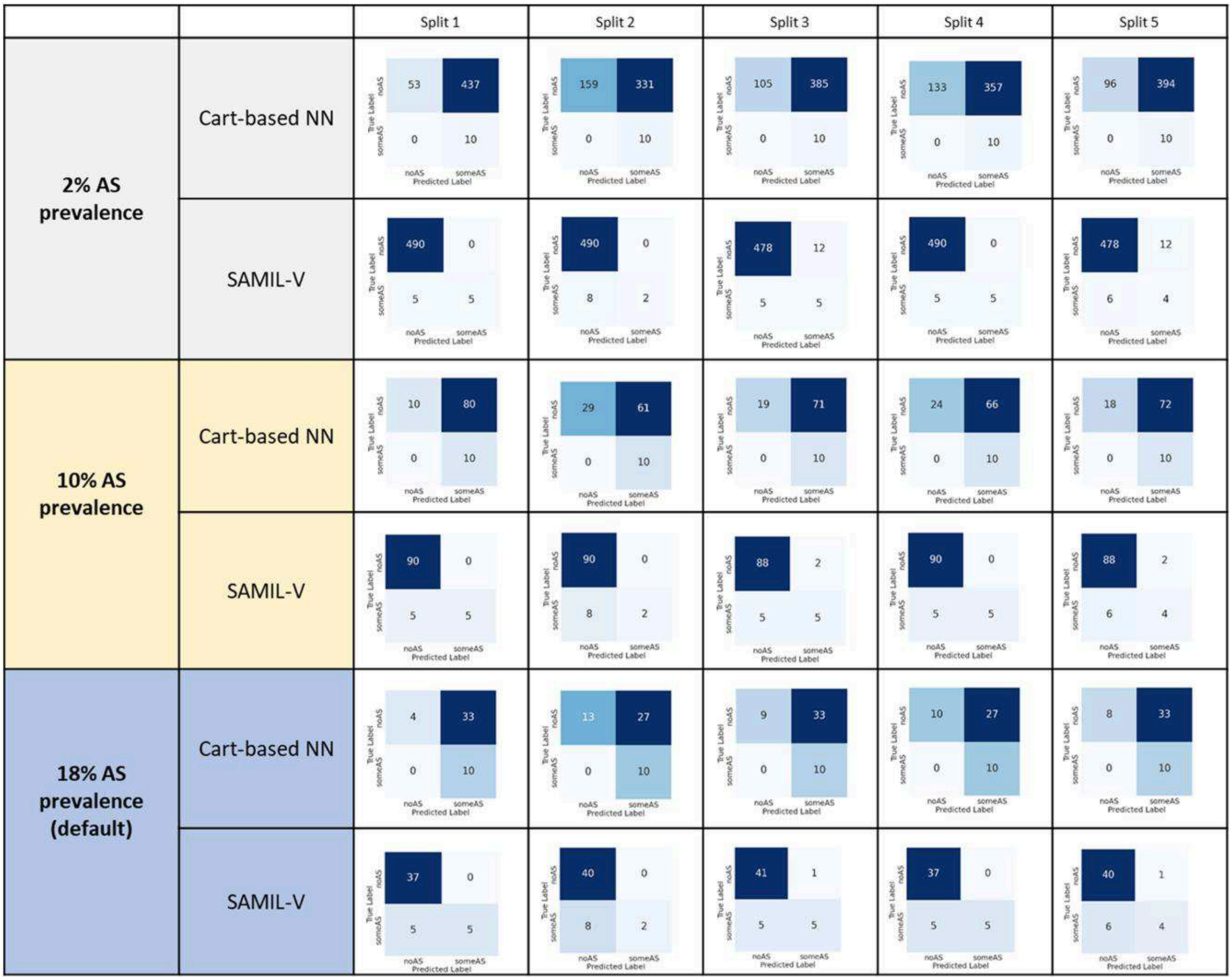
Forty-one patients were prospectively collected as an external temporal validation cohort. In this cohort, there was 6 cases of AS. The fine-tuned SAMIL-V classified has an AUROC of 0.97 (0.89–1.0) for detecting AS, PPV was 1.0 (95% CI 1.0–1.0), NPV was 0.92 (95% CI 0.88–0.97).

Modelling performance in a screening environment

The prevalence of any degree of AS in each split's test set was adjusted to 2 and 10% by sampling with replacement the 'no AS' cases (*Figure 2*). We selected 2 and 10% values to simulate expected prevalence rates in potential screening environments. The AUROC for detecting any degree of AS was not significantly affected by decreasing prevalence, while the AUPRC and PPV decrease as expected as prevalence drops. At a prevalence of 2%, the AUROC was 0.94 (0.93–0.94), AUPRC was 0.56 (0.55–0.58), PPV was 0.72 (0.69–0.76), NPV was 0.99 (0.99–0.99). At a prevalence of 10%, the AUROC was 0.94 (0.93–0.95), AUPRC was 0.74 (0.70–0.78), PPV was 0.88 (0.81–0.97), and NPV was 0.94 (0.94–0.94). Confusion matrices of the results are shown in *Figure 2*.

Discussion

The primary finding from this study is a notable drop in discriminatory performance when a cart-based ML model is tested on handheld imaging done by expert imagers. AS classifiers can be successfully adapted to handheld ultrasound imaging via efficient last-layer fine-tuning, yielding useful performance at identifying patients who should be referred for comprehensive echocardiography. These results lay the groundwork for improving detection of AS upstream of traditional echocardiography laboratories.



		Split 1	Split 2	Split 3	Split 4	Split 5
2% AS prevalence	Cart-based NN	53 437 0 10	159 331 0 10	105 385 0 10	133 357 0 10	96 394 0 10
	SAMIL-V	490 0 5 5	490 0 8 2	478 12 5 5	490 0 5 5	478 12 6 4
10% AS prevalence	Cart-based NN	10 80 0 10	29 61 0 10	19 71 0 10	24 66 0 10	18 72 0 10
	SAMIL-V	90 0 5 5	90 0 8 2	88 2 5 5	90 0 5 5	88 2 6 4
18% AS prevalence (default)	Cart-based NN	4 33 0 10	13 27 0 10	9 33 0 10	10 27 0 10	8 33 0 10
	SAMIL-V	37 0 5 5	40 0 8 2	41 1 5 5	37 0 5 5	40 1 6 4

Figure 2 Confusion matrices at different prevalence of AS (2, 10, and 18%). Confusion matrices for the cart-based network and the fine-tuned SAMIL-V networks at varying disease prevalence. To obtain robust estimates of performance, we repeat all evaluations over five separate train/test splits of the 160 FoCUS scans in our dataset. Each split is drawn randomly and independently. The 2 and 10% prevalence are simulated by upsampling with replacement 'no AS' cases from the actual observed test set in each split. NN is neural network.

Our baseline cart-based classifier showed a decrement in discriminatory performance when tested on handheld ultrasound imaging. This is a concern (and potential barrier to use) because handheld datasets more closely approximate the proposed use case for detecting AS upstream of traditional echocardiography laboratories. Additional work is needed to assess our fine-tuned model in the intended use environment upstream of the echocardiogram laboratory with imaging done by non-experts. If the original cart-based model (without fine tuning) is applied to groups with low disease prevalence (i.e. patients in a primary care screening environment), the number of false positives is likely to out-number true positives and screening would lead to low-value follow-up testing. While thresholds for ultimate binary classification can be adjusted to decrease the number of false positives, there would be an expected decrease in sensitivity and cases of AS would be missed. Final threshold optimization should be done in collaboration with the clinicians who will use these predictions.

We present an improved AS classifier that is fine-tuned on FoCUS imaging from handheld ultrasound devices and therefore optimized for the intended clinical use case. This new classifier shows substantially improved performance over the original cart-based baseline for use with FOCUS handheld scans to detect patients with AS who should be referred for comprehensive echocardiography and follow-up cardiology care. These performance gains were achieved by developing comprehensive yet efficient methods that incorporate recent advances in MIL, semi-supervised learning, and (most importantly) transfer learning. Despite relatively small sets of available labelled data, our top-performing classifier demonstrates discriminatory performance that can enable disease screening for high-risk populations.

Care for patients with AS is rapidly changing. There are now effective treatments for severe symptomatic AS that are widely available.[17] These treatments have recently been studied for asymptomatic patients[18] and there are ongoing trials of medical therapies to halt progression of AS.[7] Despite these innovations, a large proportion of patients with AS are undiagnosed[1] and many present late in the disease course once left ventricular dysfunction or significant comorbid conditions are present. Outcomes are worse because of our current approach to diagnosis. An ML-powered automated approach to AS diagnosis that leverages fine-tuning on handheld ultrasound imaging can improve the diagnosis of AS. This improvement could in turn shift case detection upstream of traditional echocardiogram laboratories and improve the yield of comprehensive transthoracic echocardiographic imaging. Our work shows that ML classifiers should be tested (and may need to be fine-tuned for handheld ultrasound) before they can be deployed. Without testing and optimization, these networks may be costly and may not improve care. While these networks automate image interpretation, providers must still be trained to acquire FoCUS, and care pathways for timely comprehensive imaging and follow-up care must be re-defined.

This study has some limitations. While the FoCUS dataset size is modest, to our knowledge this is the largest handheld cardiac imaging database that has been presented. We anticipate the performance will increase as training cohort size increases. The FoCUS imaging in this study was obtained by trained cardiac sonographers; however in order to move case identification upstream of traditional echocardiogram laboratories, non-experts must be trained to acquire imaging. The image quality obtained by non-expert imagers is likely to be lower quality than the images in this study. Additionally, the feasibility of incorporating FoCUS in screening environments (who should acquire imaging and how to identify high-risk patients) requires more study. Model abstention[19] (i.e. when models should withhold low confidence predictions) is an area that needs additional study as well.

Conclusion

A new ML classifier that is fine-tuned for handheld ultrasound imaging acquired by sonographers shows improved performance for detecting AS compared with a cart-based classifier. These methods overcome concerns about transportability to handheld imaging and represent progress towards improving detection of AS upstream of traditional echocardiogram laboratories.

Consent: This study was approved by the institutional review board at Tufts Medical Center and all patient consented to participation.

Conflict of interest: B.S.W. has done consulting work for iCardio.ai unrelated to the present work.

Funding

National Institutes of Health grant K23AG055667. National Center for Advancing Translational Sciences, National Institutes of Health grant UM1TR004398. U.S. National Science Foundation (NSF) grant IIS # 2338962.

Data availability

Cart-based imaging data are hosted at https://tmed.cs.tufts.edu/ and are available upon reasonable request.

Lead author biography

Benjamin S. Wessler is Associate Professor of Medicine at Tufts University School of Medicine, Associate Director of the Predictive Analytics and Comparative Effectiveness Center (PACE) and Director of the Heart Valve Center at Tufts Medical Center.

References

1. d'Arcy JL, Coffey S, Loudon MA, Kennedy A, Pearson-Stuttard J, Birks J *et al.* Large-scale community echocardiographic screening reveals a major burden of undiagnosed valvular heart disease in older people: the OxVALVE population cohort study. *Eur Heart J* 2016;**37**:3515–22.
2. Gardezi SKM, Myerson SG, Chambers J, Coffey S, d'Arcy J, Hobbs FDR *et al.* Cardiac auscultation poorly predicts the presence of valvular heart disease in asymptomatic primary care patients. *Heart* 2018;**104**:1832–5.
3. Strange G, Stewart S, Celermajer D, Prior D, Scalia GM, Marwick T *et al.* Poor long-term survival in patients with moderate aortic stenosis. *J Am Coll Cardiol* 2019;**74**:1851–63.
4. Strange GA, Stewart S, Curzen N, Ray S, Kendall S, Braidley P *et al.* Uncovering the treatable burden of severe aortic stenosis in the UK. *Open Heart* 2022;**9**:e001783.
5. Baron SJ, Arnold SV, Herrmann HC, Holmes DR, Szeto WY, Allen KB *et al.* Impact of ejection fraction and aortic valve gradient on outcomes of transcatheter aortic valve replacement. *J Am Coll Cardiol* 2016;**67**:2349–58.
6. Lancellotti P, Magne J, Dulgheru R, Clavel M-A, Donal E, Vannan MA *et al.* Outcomes of patients with asymptomatic aortic stenosis followed up in heart valve clinics. *JAMA Cardiol* 2018;**3**:1060–8.
7. Lindman BR, Sukul D, Dweck MR, Madhavan MV, Arsenault BJ, Coylewright M *et al.* Evaluating medical therapy for calcific aortic stenosis: JACC state-of-the-art review. *J Am Coll Cardiol* 2021;**78**:2354–76.
8. Cohen-Shelly M, Attia ZI, Friedman PA, Ito S, Essayagh BA, Ko W-Y *et al.* Electrocardiogram screening for aortic valve stenosis using artificial intelligence. *Eur Heart J* 2021;**42**:2885–96.
9. Wessler BS, Huang Z, Long GM, Pacifici S, Prashar N, Karmiy S *et al.* Automated detection of aortic stenosis using machine learning. *J Am Soc Echocardiogr* 2023;**36**(4):411–20.

10. Holste G, Oikonomou EK, Mortazavi BJ, Coppi A, Faridi KF, Miller EJ *et al.* Severe aortic stenosis detection by deep learning applied to echocardiography. *Eur Heart J* 2023;**44**: 4592–604.
11. Pooch EHP, Ballester PL, Barros RC. Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification. https://doi.org/10.48550/arXiv.1909.01940
12. Beery S, van Horn G, Perona P. Recognition in Terra Incognita. 2018. https://doi.org/10.48550/arXiv.1807.04975
13. Baumgartner H, Hung J, Bermejo J, Chambers JB, Edvardsen T, Goldstein S *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**: 372–92.
14. Zagoruyko S, Komodakis N. Proceedings of the British Machine Vision Conference (BMVC). p. 87.1–87.12. 2016.
15. Huang Z, Wessler BS, Hughes MC. Detecting heart disease from multi-view ultrasound images via supervised attention multiple instance learning. *Proc Mach Learn Res* 2023; **219**:285–307.
16. Huang Z, Yu X, Wessler BS, Hughes MC. Semi-supervised multimodal multi-instance learning for aortic stenosis diagnosis. *IEEE Int Symp Biomed Imaging* 2025. https://doi.org/10.48550/arXiv.2403.06024
17. Marquis-Gravel G, Stebbins A, Kosinski AS, Cox ML, Harrison JK, Hughes GC *et al.* Geographic access to transcatheter aortic valve replacement centers in the United States: insights from the society of thoracic surgeons/American College of Cardiology transcatheter valve therapy registry. *JAMA Cardiol* 2020;**5**:1006–10.
18. Généreux P, Schwartz A, Oldemeyer JB, Pibarot P, Cohen DJ, Blanke P *et al.* Transcatheter aortic-valve replacement for asymptomatic severe aortic stenosis. *N Engl J Med* 2024;**392**(3):217–27.
19. Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digit Med* 2021;**4**:4.