ESC European Society of Cardiology

European Heart Journal - Digital Health (2022) **3**, 141–152
https://doi.org/10.1093/ehjdh/ztac029

ORIGINAL ARTICLE

Efficient screening for severe aortic valve stenosis using understandable artificial intelligence: a prospective diagnostic accuracy study

Hisaki Makimoto [1,*,†], Takeru Shiraga [2,†], Benita Kohlmann [1], Christofori-Eleni Magnisali [1], Shqipe Gerguri [1], Nobuaki Motoyama [2], Lukas Clasen [1], Alexandru Bejinariu [1], Kathrin Klein [1], Asuka Makimoto [1], Christian Jung [1], Ralf Westenfeld [1], Tobias Zeus [1], and Malte Kelm [1,3]

[1]Division of Cardiology, Pulmonology and Vascular Medicine, Medical Faculty, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany; [2]Mitsubishi Electric Inc., Kamakura, Japan; and [3]CARID - Cardiovascular Research Institute Düsseldorf, Düsseldorf, Germany

Received 1 February 2022; revised 8 May 2022; online publish-ahead-of-print 16 May 2022

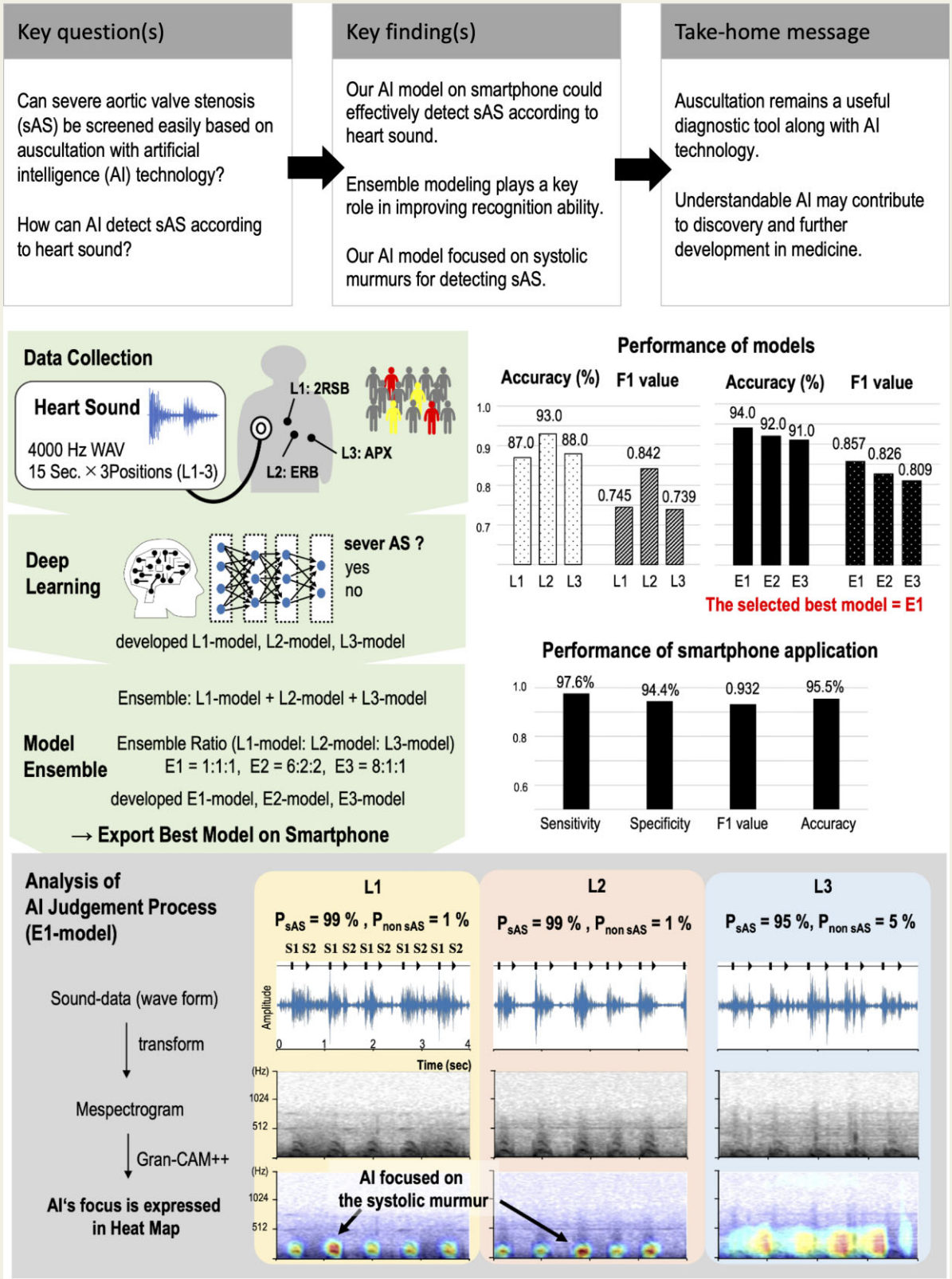
Aims The medical need for screening of aortic valve stenosis (AS), which leads to timely and appropriate medical intervention, is rapidly increasing because of the high prevalence of AS in elderly population. This study aimed to establish a screening method using understandable artificial intelligence (AI) to detect severe AS based on heart sounds and to package the built AI into a smartphone application.

Methods and results In this diagnostic accuracy study, we developed multiple convolutional neural networks (CNNs) using a modified stratified five-fold cross-validation to detect severe AS in electronic heart sound data recorded at three auscultation locations. Clinical validation was performed with the developed smartphone application in an independent cohort (model establishment: $n = 556$, clinical validation: $n = 132$). Our ensemble technique integrating the heart sounds from multiple auscultation locations increased the detection accuracy of CNN model by compensating detection errors. The established smartphone application achieved a sensitivity, specificity, accuracy, and F1 value of 97.6% (41/42), 94.4% (85/90), 95.7% (126/132), and 0.93, respectively, which were higher compared with the consensus of cardiologists (81.0%, 93.3%, 89.4%, and 0.829, respectively), implying a good utility for severe AS screening. The Gradient-based Class Activation Map demonstrated that the built AIs could focus on specific heart sounds to differentiate the severity of AS.

Conclusions Our CNN model combining multiple auscultation locations and exported on smartphone application could efficiently identify severe AS based on heart sounds. The visual explanation of AI decisions for heart sounds was interpretable. These technologies may support medical training and remote consultations.

* Corresponding author. Tel: +49 211 81 18800, Fax: +49 211 81 19520, E-mail: h1sak1mak1m0t0@gmail.com
†H.M. and T.S. contributed equally to this work.

Graphical Abstract



Keywords Auscultation • Artificial intelligence • Aortic valve stenosis • Screening • Understandable AI

Introduction

Aortic valve stenosis (AS) is one of the most common valvular heart diseases (VHDs) in developed countries. Recent guidelines on VHD management have updated treatment recommendations for asymptomatic severe AS.[1] Further, the recent OxVALVE study observed a substantial number of undiagnosed patients with VHD among the elderly population.[2] Therefore, an early and easily accessible screening method is necessary to unmask severe AS in asymptomatic patients to improve prognosis and to prevent sudden cardiac death.

Currently, echocardiography is the gold standard for diagnosis of patients with VHD.[1,3] Auscultation is a simple, cost-effective, and a basic diagnostic examination procedure that has occupied an important place in medical education and training. However, Gardezi *et al.* reported that using cardiac auscultation alone for VHD detection is a poor screening tool in primary care and recommended to have an easier access to echocardiography.[4–7]

Artificial intelligence (AI) technologies are developing drastically and have extended its use in the medical field. The potential recognition capability of AI is shown to be superior than that of human capability.[8,9] Chorba *et al.*[10] developed and evaluated an AI system to detect VHD based on heart sounds in patients with single valvular disease. It still remains unclear if AI can distinguish the heart sounds in presence of systolic murmurs caused by other common VHD. The current AI architecture has often been criticized because it is difficult to explain how the output (decision) is determined. Therefore, explainability and interpretability are key requisites in introducing AI technology in medical practice.

The objectives of this study were as follows: (i) to develop an AI technology to screen and detect severe AS based on heart sounds as efficiently as cardiologists, even in the coexistence of other valvular diseases, (ii) to visualize the built AI's focus to verify its judgement process, and (iii) to package the built AI into a smartphone application to efficiently screen and easily interpret the heart sounds.

Methods

Research participants

Patients who underwent transthoracic echocardiography (TTE) in our institute were enrolled in the study. The trained members of our project team (both coordinators and auscultators) had to be present during the enrolment process. The auscultators were blinded to the cardiovascular status of the participants.

Inclusion criteria were as follows: (i) age $\geq$ 18 years, (ii) a complete set of echocardiography tests, and (iii) written informed consent. Exclusion criteria were as follows: (i) history of any heart valve surgery or transcatheter aortic valve implantation, and (ii) use of a cardiac implantable electronic device, except an implantable loop recorder.

The existing guidelines for echocardiography were followed to establish a diagnosis (see Supplemental Material for detail).[1,11] Based on the diagnosis of TTE, we prospectively enrolled 886 participants (establishment cases) including 114 patients with severe AS and 772 patients without severe AS (no AS, $n = 670$; mild AS, $n = 51$; moderate AS, $n = 51$). The Standards for Reporting Diagnostic accuracy studies (STARD) 2015 flow diagram for the enrolment of the participants is shown in see Supplementary material online, *Figure S1*. No patients were enrolled twice in the present study.

Data collection

All data were collected after written informed consent was obtained from the patient and the participants' data were pseudonymized at the data centre. Further, data were anonymized during all processes of the project.

The auscultators were blinded to the patients' clinical information at the time of recording. A phonocardiogram was recorded using an Eko Duo system Eko Devices Inc., Oakland, CA, USA) on the day of echocardiography ($\pm$1 day). The heart sounds were recorded in the 4000 Hz.wav format from the following three auscultation locations in each patient: second intercostal space along the right sternal border (L1: 2RSB), Erb's area (third intercostal space along the left sternal border: L2: ERB), and apex (fifth intercostal space in the midclavicular line: L3: APX). The duration of each recording was 15-s. The 12-lead electrocardiogram and clinical data including medical and treatment history were collected on the day of phonocardiogram recording.

Dataset preparation and data preprocessing

The detailed processes are documented in the Supplementary material online. In brief, we adopted a modified stratified five-fold cross-validation method to train the models. Each training set contained 352 training cases (severe AS = 74, moderate AS = 33, mild AS = 33, and no AS = 212), 100 development cases (severe AS = 20, moderate AS = 9, mild AS = 9, and no AS = 62), and 100 test cases (severe AS = 20, moderate AS = 9, mild AS = 9, and no AS = 62), and there were no overlaps between the training, development, and test cases. Further, the test cases were common to all the five training sets for a fair evaluation of the performance of the model (see Supplementary material online, *Figure S2* for detail).

Based on our previous experience that there is a possibility of misrecognition of systolic murmurs due to AS and mitral regurgitation, we balanced the number of cases with mitral regurgitation and AS while building the cross-validation sets, 330 of 886 participants were excluded from the establishment dataset (*Table 1*).

For preprocessing, a 128-dimension log-Mel spectrogram was extracted (see Supplementary material online, *Figure S3*). All our models were constructed to accept log-Mel spectrogram using 4 s heart sound data as input data.

Training and development of models

We adopted two different architectures of convolutional neural networks (CNNs) based on our experience (see Supplementary material online, *Figure S4A, S4B*).[9] The output of our models interpreted if the input heart sound were that of severe AS or not. Python 3.7 and TensorFlow 2.3 (Google LLC, Mountain View, CA, USA) were used for this project. We trained the CNN models using the entire 4 s heart sound data from all the three collection locations, and then separately for each location. (see Methods in Supplemental Material and see Supplementary material online, *Figure S7* for details).

In terms of performance metrics, we determined the F1 value [harmonic mean of sensitivity and positive predictive value (PPV)], accuracy, sensitivity, specificity, PPV, negative predictive value (NPV), and area under the curve (AUC).

The best-performing model was selected for each data collection location (2RSB, ERB, and APX) based on the performance metrics in the test dataset (4 s data level) and named as the L1-, L2-, and L3-models, respectively. The best models from each data collection location were then assembled in the following combination of ratios: (i) L1-model:L2-model:L3-model = 1:1:1, (ii) L1-model:L2-model:L3-model = 6:2:2, and (iii) L1-model:L2-model:L3-model = 8:1:1.

Table 1 Basic participants' characteristics of establishment cases and excluded cases

	Training/development/test cases: $N = 556$	Excluded cases: $N = 330$
Age	68.9 ± 14.8	60.5 ± 16.0
Male	305 (54.9%)	177 (53.6%)
Hypertension	386 (69.4%)	174 (52.7%)
Diabetes mellitus	132 (23.7%)	47 (14.2%)
Dyslipidaemia	197 (35.4%)	90 (27.3%)
Atrial fibrillation at enrolment	101 (18.2%)	29 (8.8%)
Heart rate (beats per minute)	72.7 ± 13.9	72.0 ± 13.7
History of stroke/TIA	70 (13.0%)	45 (13.6%)
History of myocardial infarction	71 (12.7%)	29 (8.8%)
History of CABG	37 (6.7%)	5 (1.5%)
LVEF (%)	57.9 ± 10.8	61.5 ± 7.8
Aortic stenosis (no/I/II/III)	340/51/51/114 (61.2%/9.2%/9.2%/20.5%)	330/0/0/0 (100%/0%/0%/0%)
Severe low-flow low-gradient AS	28	0
Aortic regurgitation (no/I/II/III)	361/157/36/2 (64.9%/28.2%/6.5%/0.4%)	270/46/14/0 (81.8%/13.9%/4.2%/0%)
Mitral stenosis (no/I/II/III)	508/39/9/0 (91.4%/7.0%/1.6%/0%)	323/6/0/1 (97.9%/1.8%/0%/0.3%)
Mitral regurgitation (no/I/II/III)	211/195/96/54 (38.0%/35.1%/17.3%/9.7%)	210/116/4/0 (63.6%/35.2%/1.2%/0%)
Tricuspid stenosis (no/I/II/III)	556/0/0/0 (100%/0%/0%/0%)	329/0/0/1 (99.7%/0%/0%/0.3%)
Tricuspid regurgitation (no/I/II/III)	228/243/65/20 (41.0%/43.7%/11.7%/20%)	190/124/12/4 (57.6%/37.6%/3.6%/1.2%)
PQ interval (ms)	171 ± 38	162 ± 32
QRS duration (ms)	100 ± 23	95 ± 19
QT interval (ms)	405 ± 37	398 ± 40
QRS axis (ms)	32 ± 52	40 ± 45

CABG, coronary artery bypass grafting; LVEF, left ventricular ejection fraction; TIA, transient ischaemic attack.

For case level performance testing, the prediction per case was calculated by averaging the predictions of all data fractions in each case. During this process, one heart sound data (15-s) was divided into 11 fractions of 4 s data at regular intervals, and each data fraction was preprocessed into a log-Mel spectrogram. The accuracy of this analysis was estimated by dividing the number of patients with correct prediction by the total number of patients analysed. Each location model was tested for diagnostic accuracy, sensitivity, and specificity for severe AS using the same dataset. The best ensembled model was then selected for the smartphone application.

Building the smartphone application and clinical validation

The selected models were exported to a smartphone application using TensorFlow Lite (tflite). A smartphone application ('AudioClassification') was designed in the framework of Swift using Xcode, and iPhone11 Pro with iOS14.4 (Apple Inc., Cupertino, CA, USA) was adopted as the test device. The application was intended to provide an output interpreting whether the input case had severe AS or not.

We enrolled 132 patients prospectively in the clinical validation cohort to guarantee probative force (*Table 2*; no AS, $n = 34$; mild AS, $n = 28$; moderate AS, $n = 28$; and severe AS, $n = 42$). The detailed enrolment process of participants is shown in Supplementary material online, *Figure S6*. There was no overlap with the individuals in the establishment cases.

Six physicians were tested based on the heart sounds of the clinical validation cohort, which were also used during the clinical validation of the AI we built. All the six physicians were engaged in patient care at our institute for at least 6 years. Four of them were board-certified cardiology consultants, and the other two physicians were board-certified cardiologists. They interpreted whether the case had severe AS or not based on the heart sounds from the three auscultation locations; further, they assigned a score of 0 to non-severe disease and a score of 1 similar to the AI we developed. The 'consensus' of cardiologists was obtained by calculating the average score provided by the 6 physicians. Those cases with a calculated consensus score of less than 0.5 were classified as non-severe AS and those with a score greater than or equal to 0.5 as severe AS.

Visualization of the features identified by deep learning

We generated activation maps (heatmaps) of the final convolutional layer using Gradient-based Class Activation Maps (Grad-CAM++), which illustrated the relative positive activation of the convolutional layer with respect to the network output.[12] This heatmap was overlaid on the grey-scaled Mel spectrogram of the heart sounds. We then quantified the focus of built AI to assess how the AI differentiated the heart sounds between non-severe and severe AS. Further, we also assessed the focus scores according to the phases in a cardiac cycle (see Supplemental Material for detail).

Statistical analyses

Continuous data were expressed as mean ± standard deviation for normally distributed data. Categorical data were presented as numbers and percentages. In cases of non-normal distributed data, these were shown as median values (lower-upper quartile). The chi-square test, the Kruskal–Wallis test, Student's *t*-test, or Fisher's exact test were performed when appropriate. For the global test statistics, we used a significance level of 5%. Analyses were performed using the JMP software

Table 2 Participants characteristics of clinical validation cases

	Clinical validation cases: $N = 132$
Age	75.0 ± 12.5
Male	72 (54.5%)
Hypertension	107 (81.1%)
Diabetes mellitus	33 (25.0%)
Dyslipidaemia	61 (46.2%)
Atrial fibrillation at enrolment)	41 (31.1%)
Heart rate (beats per minute)	74.4 ± 17.3
History of stroke/TIA	16 (12.1%)
History of myocardial infarction	19 (16.5%)
History of CABG	11 (8.3%)
NYHA classification I/II/III/IV	29/37/59/7
LVEF (%)	54.5 ± 13.1
Aortic stenosis (no/I/II/III)	34/28/28/42 (25.8%/21.2%/21.2%/31.8%)
Severe low-flow low-gradient AS	8
Aortic regurgitation (no/I/II/III)	63/46/23/0 (47.7%/34.8%/17.4%/0%)
Mitral stenosis (no/I/II/III)	117/13/2/0 (88.6%/9.8%/1.5%/0%)
Mitral regurgitation (no/I/II/III)	31/56/27/18 (23.5%/42.4%/20.5%/13.6%)
Tricuspid stenosis (no/I/II/III)	132/0/0/0 (100%/0%/0%/0%)
Tricuspid regurgitation (no/I/II/III)	39/77/14/2 (29.5%/58.3%/10.6%/1.5%)
PQ interval (ms)	180 ± 39
QRS duration (ms)	103 ± 26
QT interval (ms)	417 ± 46
QRS axis (ms)	16 ± 49

CABG, coronary artery bypass grafting; LVEF, left ventricular ejection fraction; TIA, transient ischaemic attack.

Version 14 (SAS Institute, Cary, NC, USA) and performed with custom python scripts on macOS computers. We adopted the bootstrapping technique (sampling with replacement, 2000 times) for the statistical analyses of the performance of the built models with the test data and during the clinical validation. To compare the performance metrics under bootstrapping, the Mann–Whitney *U* test with Bonferroni correction was used when appropriate (two-tailed).

IRB approval

The study protocol was approved by the Medical Ethics Commission of the University Hospital Düsseldorf (File Number 2019-763). All participants were informed about the research, and they provided written informed consent in accordance with the tenets of Declaration of Helsinki.

Results

Model development and selection

The performance metrics (accuracy, F1 value, and AUC) of the models for the test data are shown in *Figure 1A*.

Generally, the model performance was significantly better in the 2RSB location (L1) than in the other two locations ($P < 0.01$). The performance metrics of the models for the test data are given in detail in Supplementary material online, *Tables S2A–C*.

For the test data of 2RSB (L1), the 10-layered CNN model trained only with the 2RSB heart sound dataset showed significantly better metrics than the other models, except for sensitivity and NPV.

For the test data of ERB (L2), the 10-layered CNN model trained with the entire dataset from all three locations showed better metrics than the other models (specificity, PPV).

For the test data of APX (L3), the CNN with combination kernels trained with the entire dataset from the three locations showed metrics superior to the other models (AUC, loss).

We selected the best-performing model based on the F1 value for each location. The performance metrics of each selected model are shown in Supplementary material online, *Table S3*.

Recognition by case and ensemble

For the usage of the above-selected models, the results of severe AS recognition of the cases (15-s sound data) are shown in *Figure 1B and 1C*. All three models (L1, L2, and L3) showed a high recognition ability of over 80% for severe and no AS (see Supplementary material online, *Table S4A*). However, their performance varied depending on the severity of AS. The L1-model showed a higher sensitivity (95.0%) for severe AS than the L2-model (80.0%) and L3-model (85.0%), whereas the specificity of the L1-model for severe AS (85.0%) was lower than that of the L2-model (97.5%) and L3-model (88.8%).

To achieve higher recognition ability for the severity of AS, these three location models were combined (ensemble model).

The test results with the ensemble models are also shown in *Figure 1B and 1C* (see also Supplementary material online, *Table S4B*). Among the models, the E1-model achieved the highest accuracy (94/100, 94%) and F1 value (0.857) which was significantly higher as compared to those of the other models ($P < 0.001$).

Clinical validation

Based on the above results, we selected the E1-model for clinical validation on the smartphone with prospectively collected heart sounds of patients ($N = 132$). The 15-s heart sound data (three locations per patient) recorded by an electronic stethoscope were imported into the application. The recognition results after the analysis were shown on the graphical user interface in the application on the smartphone (see Supplementary material online, *Figure S5*).

Our smartphone application achieved an F1 value of 0.932, accuracy of 95.7% (126/132), sensitivity of 97.6% (severe AS, 41/42 cases), and specificity of 94.4% [85/90, no AS, 34/34 (100%); mild AS, 28/28 (100%); and moderate AS, 23/28 (82.1%)]. Only one case with severe AS was misclassified as having non-severe AS.

The recognition performance by physicians as compared to that by our smartphone application is shown in *Figure 2A* and *2B*. The Fleiss' Kappa calculated to assess the interobserver discrepancy of the judgement was 0.32 (95% CI 0.274–0.362) that suggested a fair agreement among the cardiologists.

We observed that the smartphone application demonstrated a significantly higher performance in all 5 metrics in comparison

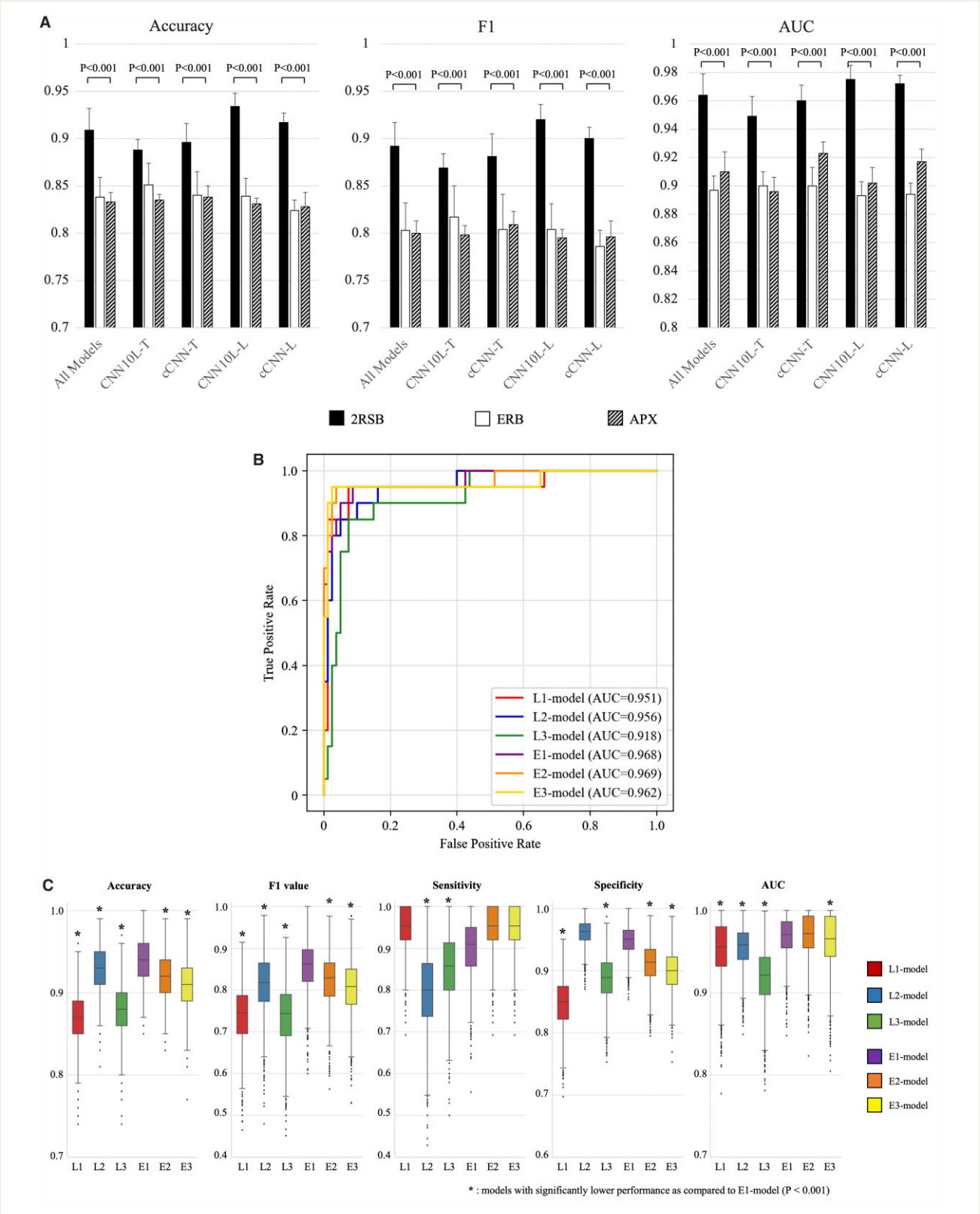
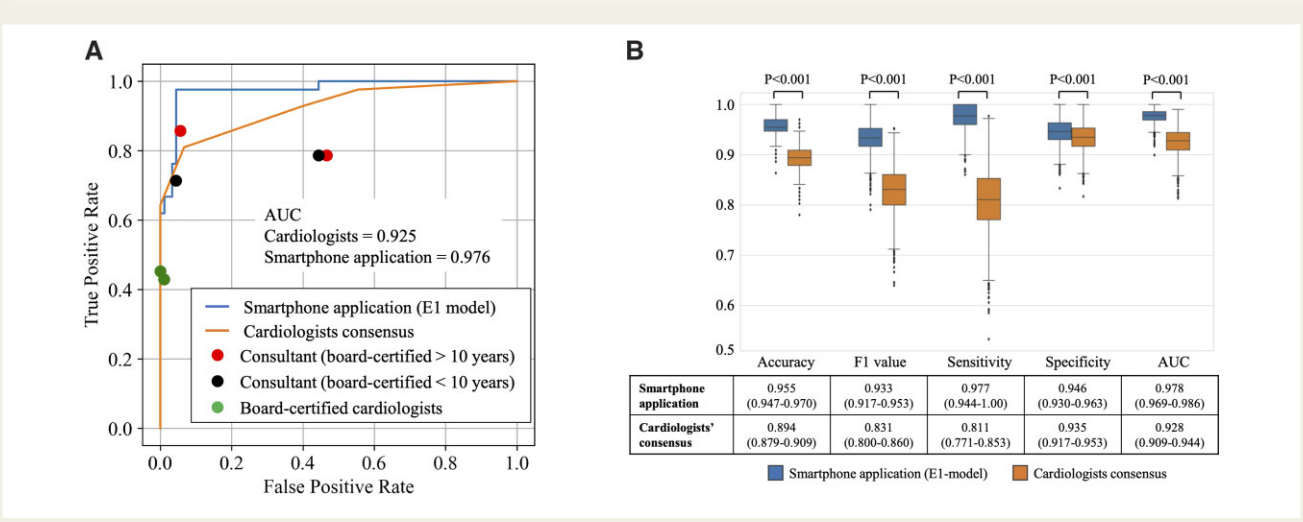


Figure 1 Performances of models in the test dataset. (*A*) The performance metrics according to the locations of the test dataset (4-s data) are shown. The metrics are significantly better in the second intercostal space along the right sternal border location than in the other auscultation locations. (*B*) The receiver-operating-characteristic curves of the developed models based on the 15-s test dataset are shown. All models demonstrated the area under the curve over 0.9. (*C*) Accuracy, F1 value, sensitivity, specificity and area under the curve based on the 15 s test dataset of the selected models (L1, L2, and L3) and the ensemble models (E1, E2, and E3) using bootstrapping are shown. The F1 value and accuracy of E1-model were significantly higher as compared to the other five models ($P < 0.001$). For all parameters the performance of E1-model was not inferior to the other models.



	Accuracy	F1 value	Sensitivity	Specificity	AUC
Smartphone application	0.955 (0.947-0.970)	0.933 (0.917-0.953)	0.977 (0.944-1.00)	0.946 (0.930-0.963)	0.978 (0.969-0.986)
Cardiologists' consensus	0.894 (0.879-0.909)	0.831 (0.800-0.860)	0.811 (0.771-0.853)	0.935 (0.917-0.953)	0.928 (0.909-0.944)

Figure 2 Performance of the smartphone application and cardiologists for clinical validation. (*A*) The receiver-operating-characteristic curves of the smartphone application and the cardiologists' consensus are shown. The performance of each cardiologist is also shown as dots. The cardiologists' performances were visually diverse (Fleiss' Kappa = 0.32). (*B*) The performances of smartphone application and cardiologists' consensus are compared using bootstrapping. In the table below the performance metrics in the clinical validation cohort (lower-upper quartile) are shown. The performances of the smartphone application were significantly higher as compared to those of the cardiologists.

to the performance metrics of cardiologists' consensus (*Figure 2B*).

Analysis of judgement process in the built AI

Figure 3 shows the analyses results of the foci on the heart sound data at each auscultation location. *Figure 3A* shows the focus scores per beat in the correctly classified cases (in detail please also see see Supplementary material online, *Table S7*). These results demonstrated that our AI from each auscultation location paid attention to the different phases of the cardiac cycle.

In the cases correctly classified as severe AS (*Figure 3B*), the L1- and L2-models recognized systolic murmurs (S1-S2), while the L3-model focused on the diastolic phase before the first heart sound (S2-S1). In the cases correctly classified as non-severe AS (*Figure 3C*), the models for L1 and L2 mainly focused on the second heart sound (S2). The L3-models' foci seemed to have a wider range compared to the L1- and L2-models.

We also analysed the cases that were incorrectly classified in the ensemble. In cases with moderate AS (*Figure 3D* and *3E*) the data was misjudged as severe AS when L1- and L2-models were focusing on the systolic phase and the L3-model focused on the late diastolic phase (before S1). Three more patients with moderate AS were similarly classified as severe AS (false positive). In the false negative case (*Figure 3F*), the data was incorrectly classified as non-severe AS when the L2-models were focusing on S2, and the L3-model focused on S1-S2. In the Supplemental Material we have also shown the other cases with their recognition results.

Discussion

We developed an AI model using CNNs, which recognized severe AS based on electronically recorded heart sound data. The major findings of the present study were as follows: (i) the recognition capabilities by CNNs were high and showed dependency on the recorded locations, (ii) the combination of CNNs (ensemble model) from multiple locations increased the total accuracy, (iii) the exported model to a smartphone application showed higher performance metrics as compared to the cardiologists, and (iv) the visualization of AI focus on heart sound data was feasible and comprehensible using Grad-CAM++.

Therefore, our results demonstrated the feasibility of heart sound screening for severe AS cases using a realistic patient cohort including all severity grades of AS and without excluding other valvular diseases.

Although auscultation is a basic examination technique, it requires expertise to accurately diagnose VHD by using auscultation alone.[13] Moreover, the use of auscultation alone in VHD diagnosis is becoming limited after the widespread use of echocardiography.

In the present study, we demonstrated that AI technology can efficiently support the diagnostic process of severe AS. In comparison to the previous studies, the strength of our study is that we enrolled patients with all severities of AS and tested the efficacy of AI technology in recognizing severe AS.[10,14,15] We also enrolled patients with mitral and tricuspid valve regurgitation that cause systolic murmurs leading to a confusion in the diagnosis of VHD. Moreover, in the present study, the heart sound data included certain levels of noise as the auscultation was performed in daily medical practice without a sound shielding room. The recorded data quality was checked by a self-check at bedside. This is a realistic setting that provides a way to screen patients in clinical practice.

Further, this non-invasive and easy-access screening method using smartphone-application will enable earlier medical access for asymptomatic patients with severe AS who may be under-recognized (*Figure 4*). In the present study, 15 patients with severe AS in the clinical validation cohort were classified by the New York Heart Association functional classification as class I ($N = 4$) and class II

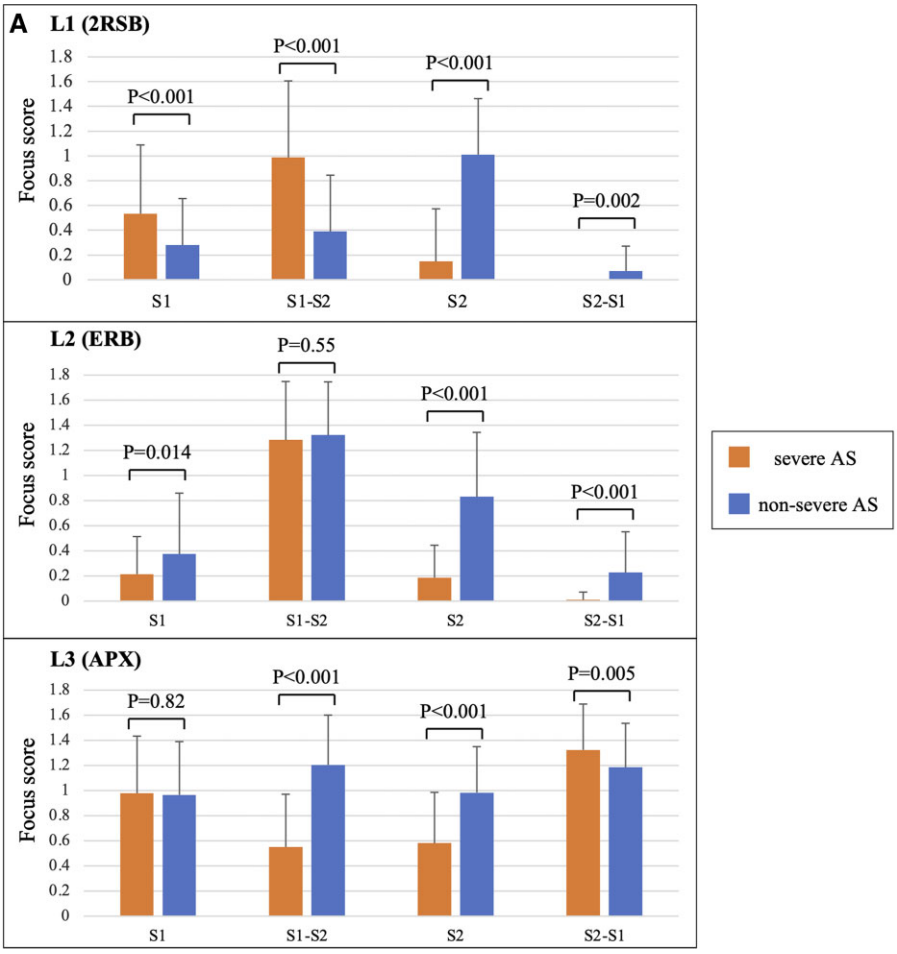


Figure 3 Visualized foci of built models in each auscultation location. (*A*) The Quantified Foci in correctly classified cases in each auscultation location are compared according to the presence of severe aortic stenosis. In the 2RSB, our artificial intelligence focused significantly more on the first heart sound and between first heart sound and second heart sound in the cases of severe aortic stenosis. In contrast, the artificial intelligence focused significantly more on the second heart sound and between second heart sound and first heart sound in the cases of non-severe aortic stenosis (upper panel). In the Erb's area, the artificial intelligence focus located on the second heart sound and between second heart sound and first heart sound in the cases of non-severe aortic stenosis (middle panel). In the apex, interestingly, the artificial intelligence focused significantly more on the phase between first heart sound and second heart sound in the cases of non-severe aortic stenosis compared to the severe aortic stenosis cases. The foci located also on the second heart sound in the non-severe aortic stenosis cases. In the severe aortic stenosis cases, the focus located on the phase between second heart sound and first heart sound (lower panel). (*B*–*F*) The examples of predictions in the clinical validation cases (red zone as strong focus). (*B*) The case correctly classified as severe aortic stenosis is shown. The L1- and L2-model focused on the systolic phase and the L3-model focused visually on the late diastolic phase before first heart sound. (*C*) The case correctly classified as non-severe aortic stenosis is shown. The case had moderate aortic stenosis. The ensembled model compensated for the failure of the L1-model by balancing the L2- and L3-models. The L1-model focused on the systolic phase and classified the data as severe aortic stenosis. In contrast, the L2-model focused on the late systolic phase to the second heart sound and classified the data as non-severe aortic stenosis. The foci of L3-model were wider, including systolic and diastolic phases, and classified the data correctly as non-severe aortic stenosis. (*D*, *E*) The incorrectly classified cases with moderate aortic stenosis are shown. The L1- and L2-models focused on the systolic phase and classified the data as severe aortic stenosis. The L3-model also classified the data as severe aortic stenosis, and the wider foci are indicated from the first heart sound to the systolic phase. (*F*) The incorrectly classified case with severe aortic stenosis is shown. The L1-model focused on the systolic phase and classified the data as severe aortic stenosis (correctly). The L2-model focused on the late systolic phase to the second heart sound, and the L3-model focused mainly on the systolic phase. They classified the data as non-severe aortic stenosis, resulting in the wrong ensembled classification.

($N = 11$). The ensembled model could correctly detect severe AS in all these asymptomatic and mildly symptomatic patients. Notably the ensembled model could recognize all the cases ($N = 8$) even with severe low-flow low-gradient AS correctly in the clinical validation cohort; thus, further supporting the utility of our model.

We interpreted the heart sounds in images through Mel spectrograms and succeeded in visualizing the AIs focus with heatmaps. The introduction of Grad-CAM++ made our AIs explainable and interpretable. The foci corresponded to the systolic murmur of AS for the detection of severe AS and to the second heart sound for the

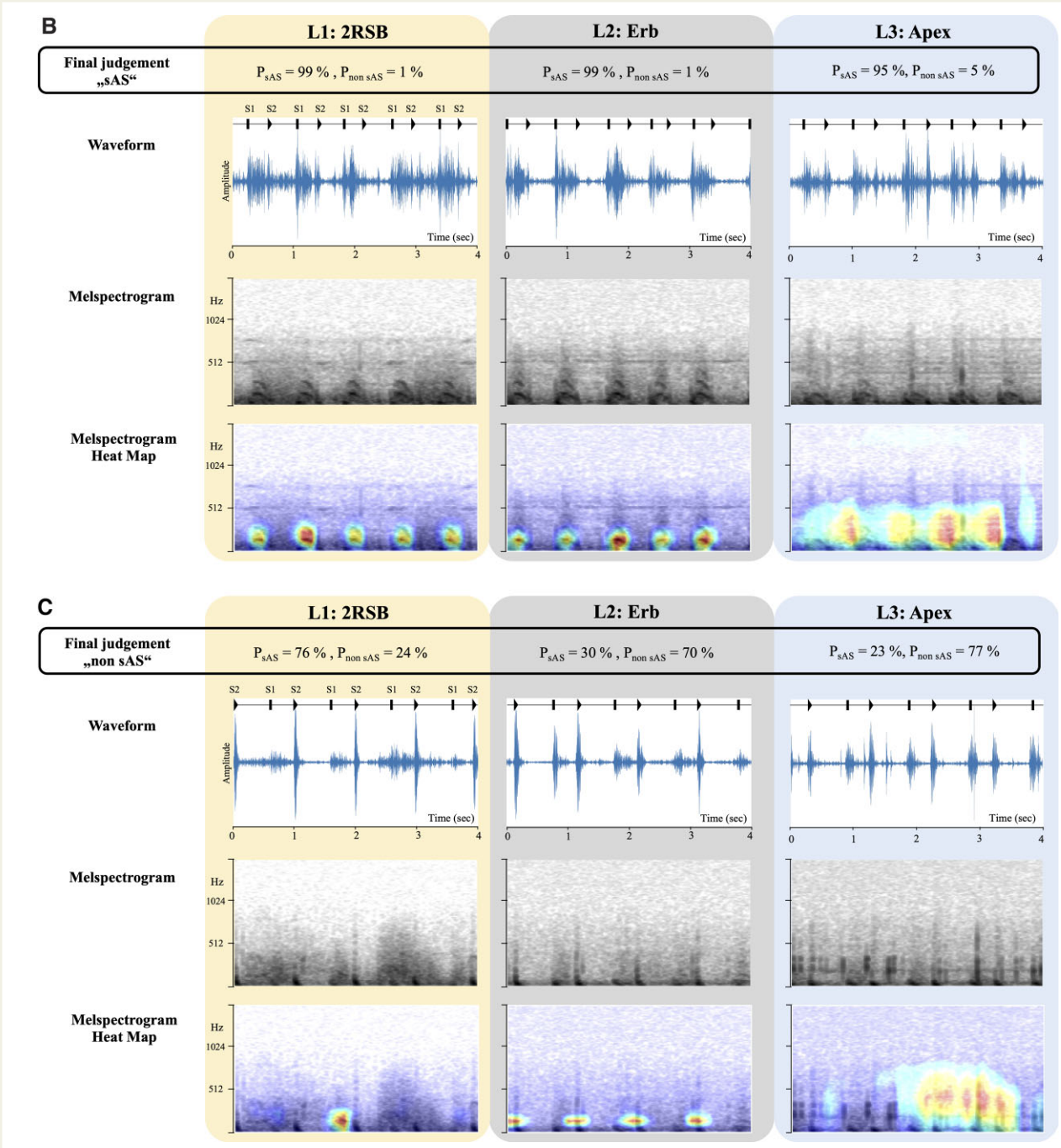


Figure 3 Continued

detection of non-severe AS in the 2RSB and Erb's location. The mid-systolic murmur and attenuation of the second heart sound are known to be typical markers of severe AS, which is consistent with our results. It could be suggested that our AI has automatically learned these traditional tips of auscultation during machine learning. This visualization technique could also bring educational improvement in medical students, young doctors, or nurses and may correct the underestimation of auscultation.

We found that the explainability and interpretability of AI through our visualization technique depended on the CNN architectures. The CNN10L models with small kernels focused on the relatively localized region (mid-systolic or second heart sound), while the CNN with a combination of multiple kernels focused on a wider region to make its decision. Further research is needed to suggest the optimal architecture structure of the neural network including kernel designs.

A recent study demonstrated that a deep learning-based algorithm detects moderate and severe AS based on an electrocardiogram.[16] The combination/ensemble of heart sounds and electrocardiograms may further improve the efficacy of screening cardiovascular diseases, which should be evaluated in the future.

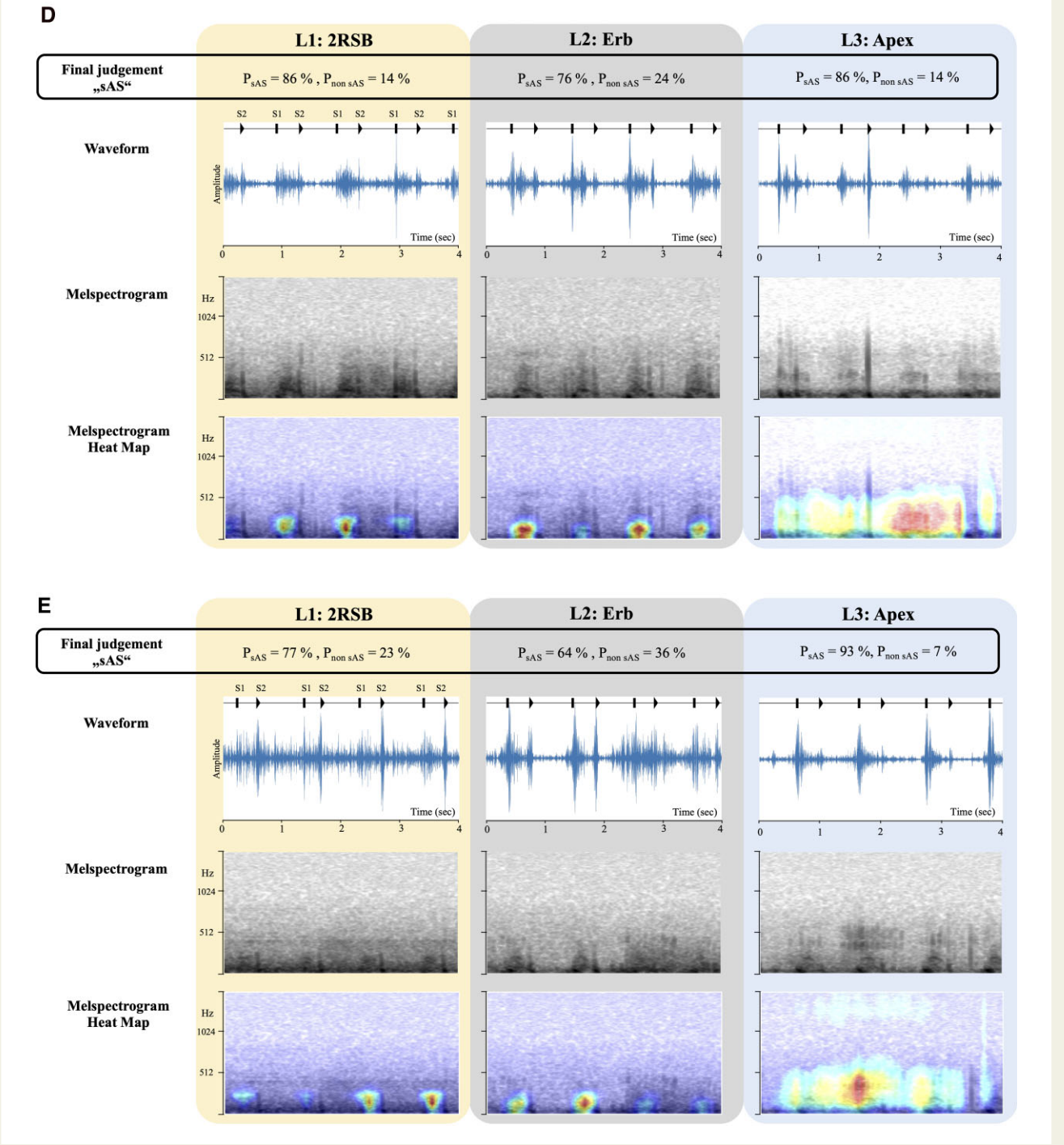


Figure 3 Continued

Study limitations

Due to the characteristics of this project as a feasibility study, this was a single-centre study with a limited number of enrolled patients. Further, the prevalence of severe AS was higher in the clinical validation cohort as compared to the general population. An institutional bias could not be excluded as our institute was transcatheter aortic valve replacement (TAVR) referral centre. We did not perform any assessment of echocardiographic reproducibility in cases having AS with borderline characteristics in this study. Based on our institutional standard as the TAVR referral centre, we believe that the echocardiographic assessments are valid. The enrolment of participants was performed by trained physicians and medical students to maintain the heart sound quality, resulting in a small number of participants. The external validation was not performed due to the training that is necessary for data acquisition and staff requirement. A larger sample size involving multicentre trials and integration of more auscultation sites

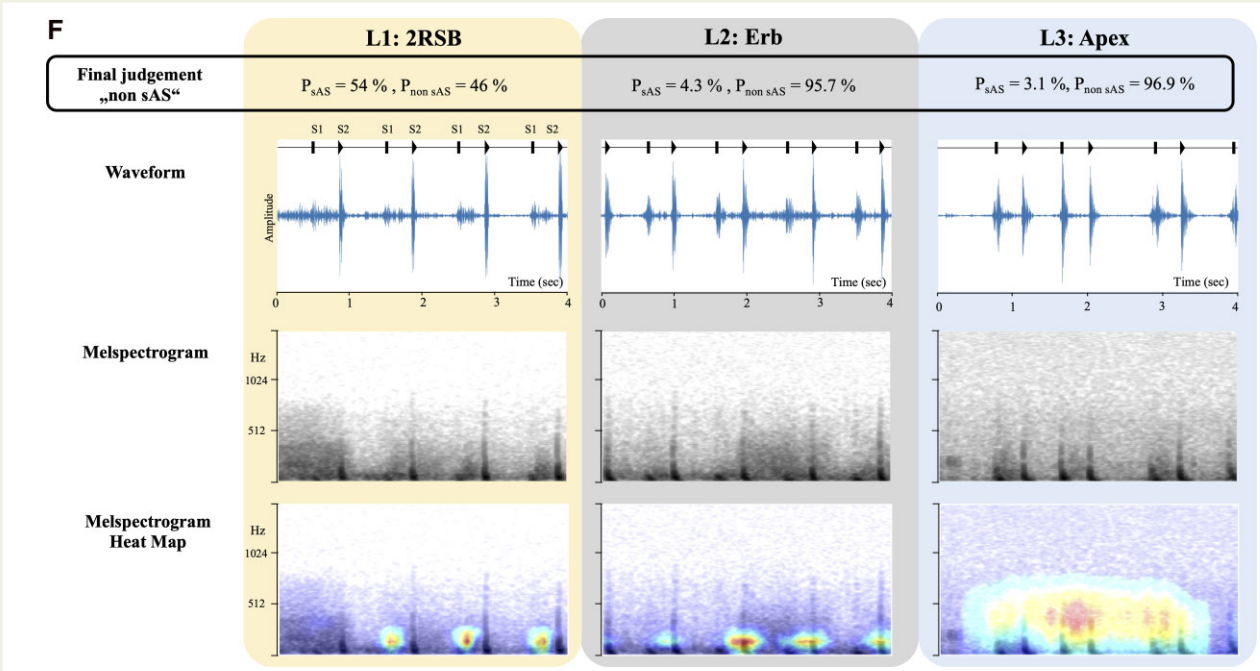


Figure 3 Continued

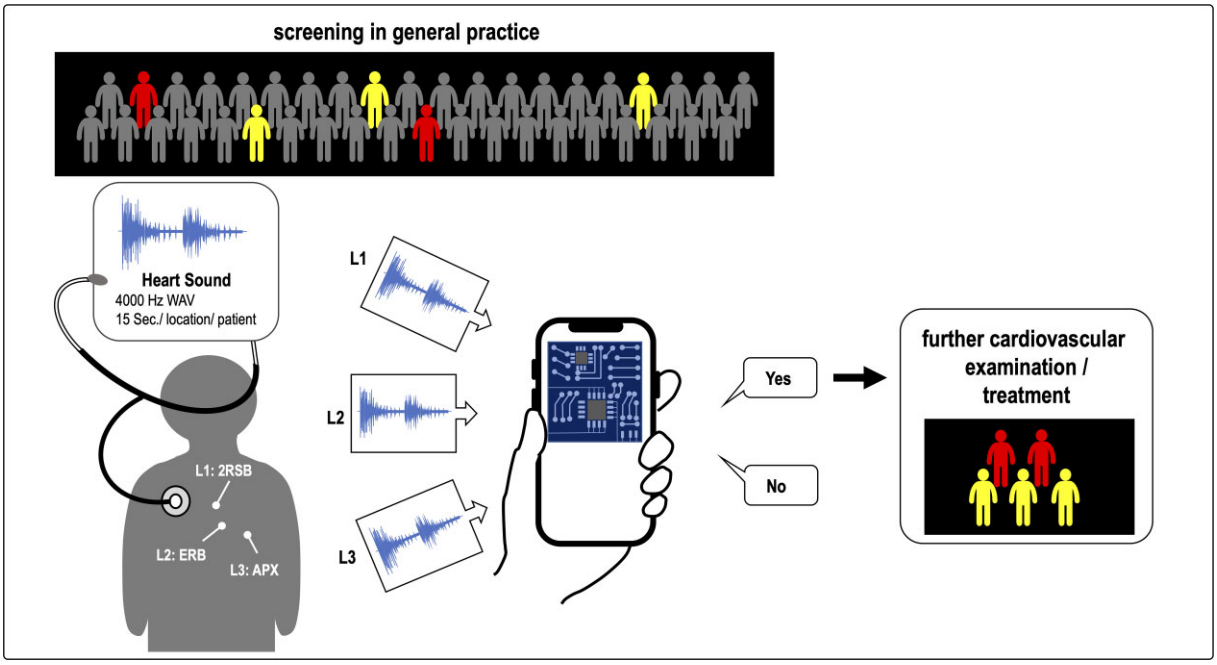


Figure 4 Future screening possibility using smartphone application. The smartphone application enables to detect severe aortic stenosisbased on the heart sounds from three locations. This can be utilized for the screening in the general practice, leading to efficient further cardiovascular examination.

with the ensemble may enhance the quality of models. Furthermore, the device usability, the extent of necessary training for data acquisition and its influence on the data quality affecting the model accuracy should be assessed in future projects, as we observed no recordings of inadequate audio quality for our analyses, in contrast to both prior studies and typical clinical practice.[10,17] We should admit that the

quantification method of visualized heatmaps was semi-quantitative due to the arbitrary threshold of our scoring system.

This research does not intend to demonstrate the inferiority of physicians/cardiologists. We recognize that a diagnosis is not purely based on auscultation but on clinical assessment. The enrolment of participants, particularly with severe aortic or tricuspid regurgitation and with implantable cardiac devices should be necessary for further evaluation of the models. The cases with other causes of left ventricular outflow tract obstruction which may mimic heart murmurs of AS were not specifically analysed in the present study; this should be evaluated in a future study.

Conclusions

This study demonstrated the promising possibility of screening for severe AS using an electronic stethoscope and deep learning technology. The visualization of AI foci is feasible and may lead to understandable AI in the medical field, which may further contribute to medical training and redefine the skills of auscultation. In future, these models may be extended to various VHDs.

Supplementary material

Supplementary material is available at *European Heart Journal – Digital Health.*

Funding

This work was funded by GIGA FOR HEALTH: 5G-Medizincampus. NRW, Sponsorship of the State of North Rhine-Westphalia by the Ministry for Economic Affairs, Innovation, Digitalization and Energy, Funding project number 005-2008-0055.

Conflict of interest: CJ: Consulting fees from Janssen Research, Honoraria for lectures from Bristol-Myers Squibb, Edwards Lifesciences, and Daiichi Sankyo, Weltenmacher. TZ: Honoraria from Medtronic , Edwards Lifesciences.

Data availability

The patient data underlying this article cannot be shared publicly to protect the privacy of the individuals who participated in the study. The data regarding the AI models will be shared upon reasonable request to the corresponding author.

Lead author biography

Hisaki Makimoto, MD, PhD is originally from Saitama, Japan. He is the chief electrophysiologist at Heinrich-Heine-University Düsseldorf, Division of Cardiology, Pulmonology and Vascular Medicine. He has 17 years' experience in cardiovascular medicine with special interest in arrhythmias and been a board-certified cardiologist (since 2012 in Japan, since 2013 in Germany) as well as a board-certified electrophysiologist since 2014 in Germany. He is a fellow of ESC (FESC since 2019), EHRA (FEHRA since 2020), and HRS (FHRS since 2021). With his knowledge in medicine and programming skill, he has engaged in medical research using artificial intelligence.

References

1. Vahanian A, Beyersdorf F, Praz F, Milojevic M, Baldus S, Bauersachs J, Capodanno D, Conradi L, De Bonis M, De Paulis R, Delgado V, Freemantle N, Gilard M, Haugaa KH, Jeppsson A, Jüni P, Pierard L, Prendergast BD, Sádaba JR, Tribouilloy C, Wojakowski W, Neumann F-J, Myers P, Abdelhamid M, Achenbach S, Asteggiano R, Barili F, Borger MA, Carrel T, Collet J-P, Foldager D, Habib G, Hassager C, Irs A, Iung B, Jahangiri M, Katus HA, Koskinas KC, Massberg S, Mueller CE, Nielsen JC, Pibarot P, Rakisheva A, Roffi M, Rubboli A, Shlyakhto E, Siepe M, Sitges M, Sondergaard L, Sousa-Uva M, Tarantini G, Zamorano JL, Praz F, Milojevic M, Baldus S, Bauersachs J, Capodanno D, Conradi L, De Bonis M, De Paulis R, Delgado V, Freemantle N, Gilard M, Haugaa KH, Jeppsson A, Jüni P, Pierard L, Prendergast BD, Sádaba JR, Tribouilloy C, Wojakowski W. 2021 ESC/EACTS guidelines for the management of valvular heart disease. *Eur Heart J* 2021:ehab395.
2. d'Arcy JL, Coffey S, Loudon MA, Kennedy A, Pearson-Stuttard J, Birks J, Frangou E, Farmer AJ, Mant D, Wilson J, Myerson SG, Prendergast BD. 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**:3515–3522.
3. Currie PJ, Seward JB, Reeder GS, Vlietstra RE, Bresnahan DR, Bresnahan JF, Smith HC, Hagler DJ, Tajik AJ. Continuous-wave Doppler echocardiographic assessment of severity of calcific aortic stenosis: a simultaneous Doppler-catheter correlative study in 100 adult patients. *Circulation* 1985;**71**:1162–1169.
4. Aronow WS, Schwartz KS, Koenigsberg M. Correlation of aortic cuspal and aortic root disease with aortic systolic ejection murmurs and with mitral anular calcium in persons older than 62 years in a long-term health care facility. *Am J Cardiol* 1986;**58**:651–652.
5. Roberts WC, Perloff JK, Costantino T. Severe valvular aortic stenosis in patients over 65 years of age. A clinicopathologic study. *Am J Cardiol* 1971;**27**:497–506.
6. Willus FA, Keyes TE. *CARDIAC CLASSICS: a collection of classic works on the heart and circulation with comprehensive biographic accounts of the authors.* St. Louis: Mosby; 1941.
7. Gardezi SKM, Myerson SG, Chambers J, Holt J, Kennedy A, Loudon M, Prendergast A, Prothero A, Wilson J, Prendergast BD. Cardiac auscultation poorly predicts the presence of valvular heart disease in asymptomatic primary care patients. *Heart* 2018;**104**:1832–1835.
8. Gonem S, Janssens W, Das N, Topalovic M. Applications of artificial intelligence and machine learning in respiratory medicine. *Thorax* 2020;**75**:695–701.
9. Makimoto H, Höckmann M, Lin T, Glöckner D, Gerguri S, Clasen L, Schmidt J, Assadi-Schmidt A, Bejinariu A, Müller P, Angendohr S, Babady M, Brinkmeyer C, Makimoto A, Kelm M. Performance of a convolutional neural network derived from an ECG database in recognizing myocardial infarction. *Sci Rep* 2020;**10**:8445.
10. Chorba JS, Shapiro AM, Le L, Prince J, Pham S, Kanzawa MM, Barbosa DN, Currie C, Brooks C, White BE, Huskin A, Paek J, Geocaris J, Elnathan D, Ronquillo R, Kim R, Alam ZH, Mahadevan VS, Fuller SG, Stalker GW, Bravo SA, Jean D, Lee JJ, Gjergjindreaj M, Mihos CG, Forman ST, Venkatraman S, McCarthy PM, Thomas JD. Deep learning algorithm for automated cardiac murmur detection via a digital stethoscope platform. *J Am Heart Assoc* 2021;**10**:e019905.
11. Otto CM, Nishimura RA, Bonow RO, Carabello BA, Erwin JP, Gentile F, Jneid H, Krieger EV, Mack M, McLeod C, O'Gara PT, Rigolin VH, Sundt TM, Thompson A, Toly C. 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* 2021;**143**:e72–e227. Erratum in: *Circulation* 2021; **143**:e229.
12. Chattopadhyay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: improved visual explanations for deep convolutional networks. *arXiv*:1710.11063.
13. McGee S. Etiology and diagnosis of systolic murmurs in adults. *Am J Med* 2010;**123**:913–921.e1.
14. Maknickas V, Maknickas A. Recognition of normal–abnormal phonocardiographic signals using deep convolutional neural networks and mel-frequency spectral coefficients. *Physiol Meas* 2017;**38**:1671–1684.
15. Wu JMT, Tsai MH, Huang YZ, Islam SKH, Hassan MM, Alelaiwi A, Fortino G. Applying an ensemble convolutional neural network with Savitzky–Golay filter to construct a phonocardiogram prediction model. *Appl Soft Comput* 2019;**78**:29–40.
16. Kwon JM, Lee SY, Jeon KH, Lee Y, Kim K-H, Park J, Oh B-H, Lee M-M. Deep learning-based algorithm for detecting aortic stenosis using electrocardiography. *J Am Heart Assoc* 2020;**9**:e014717.
17. Jariwala N, Czako S, Brenton L, Doherty A, Singh K, Klapman S, McBride J. Clinically undetectable heart sounds in hospitalized patients undergoing echocardiography. *JAMA Internal Med* 2022;**182**:86.