ESC
European Society
of Cardiology

European Heart Journal - Digital Health (2022) **3**, 20–28
https://doi.org/10.1093/ehjdh/ztab102

ORIGINAL ARTICLE
Digital health/other

Artificial intelligence-based detection of aortic stenosis from chest radiographs

Daiju Ueda [1]*, Akira Yamamoto[1], Shoichi Ehara[2], Shinichi Iwata[2], Koji Abo[3], Shannon L. Walston[1], Toshimasa Matsumoto[1], Akitoshi Shimazaki[1], Minoru Yoshiyama[2], and Yukio Miki[1]

[1]Department of Diagnostic and Interventional Radiology, Graduate School of Medicine, Osaka City University, 1-4-3 Asahi-machi, Abeno-ku, Osaka 545-8585, Japan; [2]Department of Cardiovascular Medicine, Graduate School of Medicine, Osaka City University, 1-4-3 Asahi-machi, Abeno-ku, Osaka 545-8585, Japan; and [3]Central Clinical Laboratory, Osaka City University Hospital, 1-4-3 Asahi-machi, Abeno-ku, Osaka 545-8585, Japan

Received 13 August 2021; revised 16 October 2021; accepted 30 November 2021; online publish-ahead-of-print 7 December 2021

Aims

We aimed to develop models to detect aortic stenosis (AS) from chest radiographs—one of the most basic imaging tests—with artificial intelligence.

Methods and results

We used 10 433 retrospectively collected digital chest radiographs from 5638 patients to train, validate, and test three deep learning models. Chest radiographs were collected from patients who had also undergone echocardiography at a single institution between July 2016 and May 2019. These were labelled from the corresponding echocardiography assessments as AS-positive or AS-negative. The radiographs were separated on a patient basis into training [8327 images from 4512 patients, mean age 65 ± (standard deviation) 15 years], validation (1041 images from 563 patients, mean age 65 ± 14 years), and test (1065 images from 563 patients, mean age 65 ± 14 years) datasets. The soft voting-based ensemble of the three developed models had the best overall performance for predicting AS with an area under the receiver operating characteristic curve, sensitivity, specificity, accuracy, positive predictive value, and negative predictive value of 0.83 (95% confidence interval 0.77–0.88), 0.78 (0.67–0.86), 0.71 (0.68–0.73), 0.71 (0.68–0.74), 0.18 (0.14–0.23), and 0.97 (0.96–0.98), respectively, in the validation dataset and 0.83 (0.78–0.88), 0.83 (0.74–0.90), 0.69 (0.66–0.72), 0.71 (0.68–0.73), 0.23 (0.19–0.28), and 0.97 (0.96–0.98), respectively, in the test dataset.

Conclusion

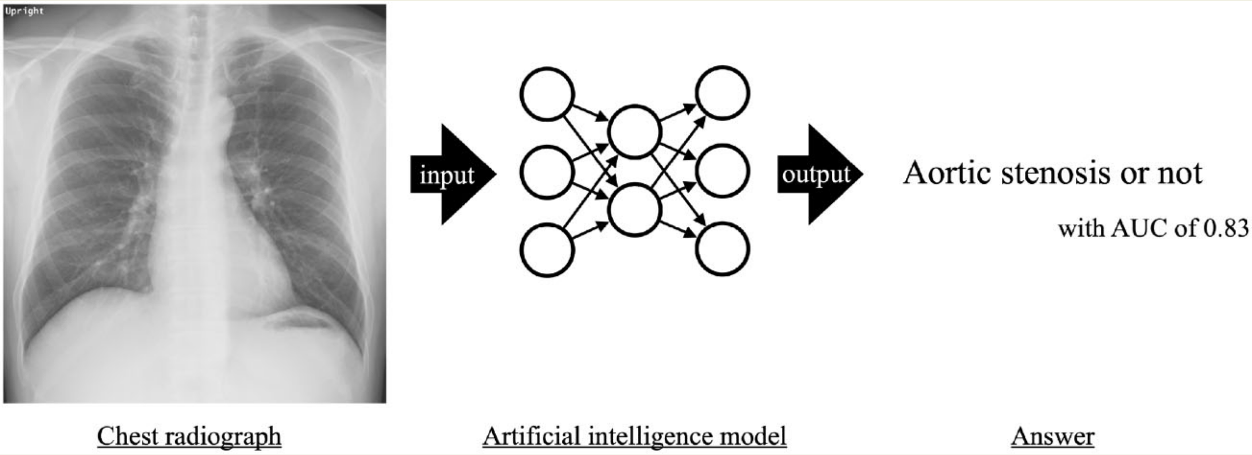
Deep learning models using chest radiographs have the potential to differentiate between radiographs of patients with and without AS.

Lay Summary

We created artificial intelligence (AI) models using deep learning to identify aortic stenosis (AS) from chest radiographs. Three AI models were developed and evaluated with 10 433 retrospectively collected radiographs and labelled from echocardiography reports. The ensemble AI model could detect AS in a test dataset with an area under the receiver operating characteristic curve of 0.83 (95% confidence interval 0.78–0.88). Since chest radiography is a cost-effective and widely available imaging test, our model can provide an additive resource for the detection of AS.

* Corresponding author. Tel: +81 6 6645 3831, Fax: +81 6 6646 6655, Email: ai.labo.ocu@gmail.com

Graphical Abstract



Keywords Aortic stenosis • Artificial intelligence • Chest radiography • Deep learning

Introduction

Aortic stenosis (AS), a stenosis of the left ventricular outflow tract and valve, causes a chronic increase in pressure in the left ventricle. In high-income countries, AS is predominantly associated with ageing; severe AS is estimated to occur in <1% of people under 70 years of age but is estimated to occur in ∼7% of people over 80 years of age.[1–3] Once heart failure, syncope, and symptoms such as chest pain appear, a patient with AS has a remaining life expectancy of 2–3 years.[4] Since AS is a progressive disease, and an increase in haemodynamic severity is inevitable once mild AS is observed, American College of Cardiology/American Heart Association (ACC/AHA) guidelines recommend regular monitoring with echocardiography at intervals appropriate for severity, ventricular size, and ventricular function, even in asymptomatic patients with known AS.[5]

Aortic stenosis is usually diagnosed when cardiac auscultation reveals a systolic murmur or upon review of echocardiography imaging that was requested for other indications. On the one hand, cardiac auscultation, being non-invasive, is clinically useful; however, it is subjective, and the accuracy of the examiner varies.[6–8] The estimated sensitivity and specificity of general practitioners for detection of significant valvular heart disease by auscultation are 44% and 69%, respectively.[9] On the other hand, transthoracic echocardiography—the diagnostic procedure recommended by the ACC/AHA—is not feasible as routine screening because of the technical, time, and cost requirements. While echocardiography is important for clinical phenotyping, human interpretation of echocardiogram images varies and can impact clinical care.[10,11] Therefore, we sought a more robust method to identify AS patients. Chest radiography has the advantage of being highly reproducible, as well as being less time-consuming and less costly. Although chest radiographs of patients with AS show findings such as left ventricular hypertrophy as a result of pressure overload, pulmonary venous dilation, and aortic valve calcifications,[12,13] diagnostic accuracies for these findings have not been reported.

In recent years, deep learning-based[14] artificial intelligence (AI) models have attracted attention because they are capable of automatically extracting features from data. Unlike conventional machine learning methods, deep learning does not require *a priori* manual feature definition because the model extracts relevant features from the data. Therefore, these models are advantageous for classification and quantification of objects with complicated features, and particularly, those with unknown features. Worldwide, chest radiographs are widely available and are cost-effective; therefore, a deep learning model capable of AS detection using chest radiographs can contribute to the improved diagnosis of AS.

Methods

Study design

We conducted training, validation, and testing on three deep learning AI models to detect AS using digital chest radiographs. Chest radiographs were collected from patients who had also undergone echocardiography at a single institution; radiographs were labelled based on echocardiography examination findings. The ethics board of our institution reviewed and approved the protocol of the present study. Since the images had been acquired during daily clinical practice, the need for informed consent was waived.

Study patients, examination, and image acquisition

Echocardiography was consecutively collected between July 2016 and May 2019. Comprehensive two-dimensional transthoracic echocardiography was performed to evaluate AS using an iE33 (Philips Medical Systems, Andover, MA, USA), Vivid E9 (GE Healthcare, Milwaukee, WI, USA), or Aplio 500/Aplio 80/Artida (Canon Medical Systems Corporation, Otawara, Tochigi, Japan) with a high-frequency transducer. Echocardiography was performed by operators with 3–20 years of echocardiography experience. If a patient had undergone echocardiography more than once in the period, all examinations were included.

Chest radiographs (posteroanterior view in the standing position) of the same patients taken within 30 days of the collected echocardiography examinations were retrospectively collected. The radiographs were taken by DR CALNEO C 1417 Wireless SQ (Fujifilm Medical, Tokyo, Japan), DR AeroDR1717 (Konica Minolta, Tokyo, Japan), or DigitalDiagnost VR (Philips Medical Systems). All eligible radiographs were collected.

Ground truth labelling

The severity of AS was classified as mild, moderate, or severe according to American Society of Echocardiography recommendations.[15] Chest radiographs corresponding to findings of AS from echocardiography (all severities from mild to severe) were defined as AS-positive, while chest radiographs corresponding to echocardiography examination reports with no findings of AS were defined as AS-negative. We also extracted data regarding valve type (tricuspid or bicuspid) and left ventricular ejection fraction (LVEF) from the echocardiography reports.

Data partitioning

All labelled chest radiographs were divided into training, validation, and test datasets in an 8:1:1 ratio. The definition of each dataset is shown in Supplementary material online, *Appendix pp 2*. The training dataset is used for training the model, the validation dataset is used for tuning the model, and the test dataset is used for evaluating the model. This test dataset was prepared to verify that there was no overfitting in our model.[16] Partitioning was performed with multiple radiographs from a single patient taken into account so that there was no overlap of images or patients among the respective datasets.

Model development

The models were developed based on three deep learning models: InceptionV3,[17] ResNet50,[18] and DenseNet121.[19] A fully connected layer in the model was connected to a sigmoid activation function with binary cross-entropy as the loss function to classify images as with or without AS. The deep learning-based models were trained from scratch with the training dataset and tuned with the validation dataset. Using the validation dataset, the model when the value of the loss function was the smallest within 100 epochs was adopted as the best-performing model. The three best-performing models were combined to create an ensemble model.[20] We used soft voting to create the ensemble model, which sums the weighted means of the probability scores of the three models. The models were built using Python 3.5 in the TensorFlow 1.15 framework. All images were augmented using random rotation from -0.1 radians to 0.1 radians, with a random shift of 10%, a brightness range of 10%, and reflected horizontally. An outline of the models is shown in the Supplementary material online, *Figure S1*, detailed hyperparameter tuning is available in Supplementary material online, *Appendix pp 2*. The trained model is available with Apache License 3.0 from https://github.com/xp-as.

Model test

Diagnostic performance of the models was assessed on the validation and test datasets using the same thresholds as those used to validate each of the three best-performing models (InceptionV3, ResNet50, and DenseNet121) and the ensemble model.

Additionally, a heat map was generated for each chest radiograph to visualize the focus of the best-performing deep learning model as it classified radiographs as with or without AS. A classification activation map applied global average pooling on the last convolutional layer in the trained deep learning model.[21] The trained weights for each output from the global average pooling layer indicated the importance/relevance of each feature map from the last convolutional layer. The weights were then applied on the corresponding feature maps, which were superimposed on the original chest radiographs, thereby creating class-discriminative visualization. A detailed explanation of the heat map generation model is shown in the Supplementary material online, *Figure S2*, and the source code is available online (https://github.com/xp-as/).

Statistical analysis

Sensitivity, specificity, accuracy, positive predictive value, negative predictive value, and the area under the receiver operating characteristic curve (AUC) were assessed for the best-performing models. Sensitivity was assessed by the severity of AS. Both sensitivity and specificity were assessed by the valve type (tricuspid or bicuspid) and grading of the LVEF. We calculated the positive predictive value and negative predictive value for all possible cohort prevalence values and illustrated the relationship between these for the best-performing AI models. All analyses were performed using R, version 3.6.0. All statistical tests were two-sided (5% significance level). The 95% confidence intervals for sensitivity, specificity, accuracy, positive predictive value, and negative predictive value were calculated using the Clopper–Pearson method.[22]

Role of the funding source

The funder had no role in the study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all data in the study and final responsibility for the decision to submit the report for publication.

Results

Datasets

In total, 10 433 radiographs with 7555 corresponding echocardiography examination reports from 5638 patients were used in the study. For echocardiography, inter-operator variability was 3.31% for peak aortic jet velocity, 6.52% for aortic valve area, and 7.21% for mean pressure gradient. The training dataset included 8327 images [4512 patients; age: range 14–98 years, (mean) 65 ± (standard deviation) 15 years]. The validation dataset included 1041 images (563 patients; age 14–93 years, 65 ± 14 years). The test dataset included 1065 images (563 patients; age 15–99 years, 65 ± 14 years). The flowchart of dataset criteria is shown in *Figure 1*. Dataset information is shown in *Table 1*. Dataset information for echocardiography and chest radiography equipment is shown in Supplementary material online, *Table S1*.

Model development

The models were each independently developed using the training dataset applied for 100 training epochs, then the loss value on a separate validation dataset determined the performance of the model. The final hyperparameters for all models were the Adagrad optimizer, image size = 320 pixels, three channels, global average pooling. The best batch sizes are 16 for ResNet50, and 32 for both InceptionV3 and DenseNet121. During this training period, the

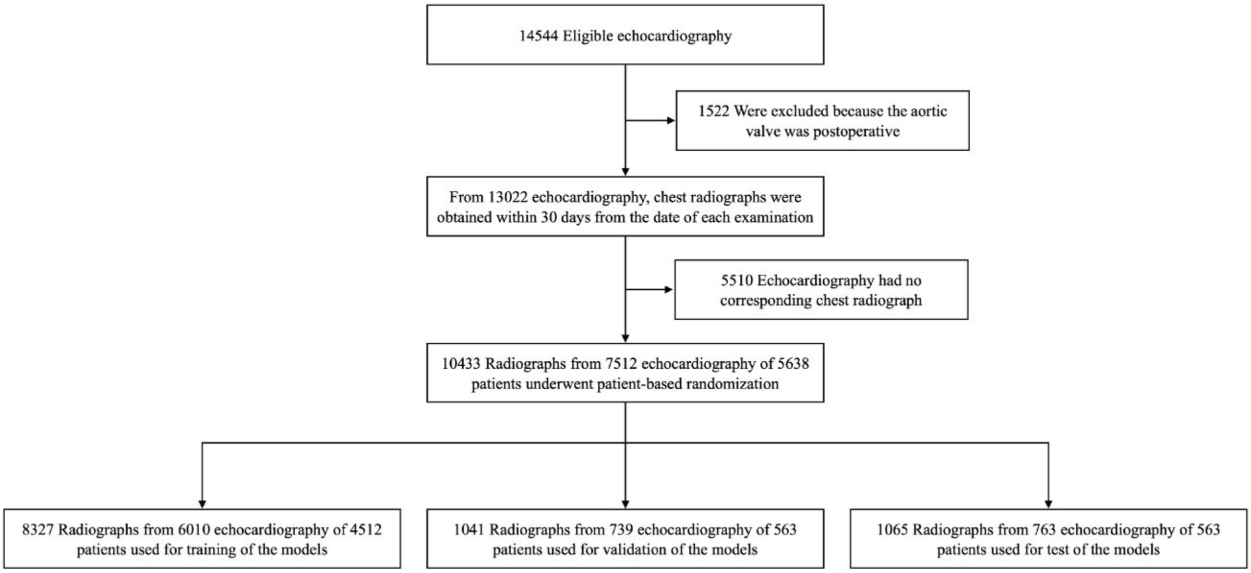

14544 Eligible echocardiography
1522 Were excluded because the aortic valve was postoperative
From 13022 echocardiography, chest radiographs were obtained within 30 days from the date of each examination
5510 Echocardiography had no corresponding chest radiograph
10433 Radiographs from 7512 echocardiography of 5638 patients underwent patient-based randomization
8327 Radiographs from 6010 echocardiography of 4512 patients used for training of the models
1041 Radiographs from 739 echocardiography of 563 patients used for validation of the models
1065 Radiographs from 763 echocardiography of 563 patients used for test of the models

Figure 1 Review and allocation flowchart.

Table 1 Dataset demographics

	Training dataset	Validation dataset	Test dataset
Total no. of radiographs	8327	1041	1065
Total no. of echocardiography	6010	739	763
Total no. of patients	4512	563	563
Male	2634	340	336
Female	1878	223	227
Mean age (years ± SD)	65 ± 15	65 ± 14	65 ± 14
Mean period between examinations (days ± SD)	5 ± 8	5 ± 8	5 ± 8
Severity of AS			
AS-negative	7407	960	959
Mild	182	19	26
Moderate	152	11	13
Severe	586	51	67
Type of valve			
Tricuspid valve	8277	1027	1060
Bicuspid valve	50	14	5
Left ventricular ejection fraction			
≥50%	6650	853	859
40–50%	672	74	80
<40%	1005	114	126

Data are *n* radiographs unless otherwise noted.
AS, aortic stenosis; SD, standard deviation.

lowest total loss value occurred at 15 epochs when the loss value was 0.24 in the validation dataset in InceptionV3, at 83 epochs when the loss value was 0.25 in the validation dataset in ResNet50, and at 22 epochs when the loss value was 0.25 in the validation dataset in DenseNet121. Learning curves for each AI model are shown in Supplementary material online, *Figure S3*. The model parameters from these timepoints were then applied to the test dataset to evaluate the model.

Model evaluation

The output classified radiograph images as AS-positive or AS-negative and was compared to the ground truth for performance calculations (*Table 2*). Among the three models and their ensemble model, the ensemble model showed the highest overall performance. The ensemble model's AUC, sensitivity, specificity, accuracy, positive predictive value, and negative predictive value were 0.83 (0.77–0.88), 0.78 (0.67–0.86), 0.71 (0.68–0.73), 0.71 (0.68–0.74), 0.18 (0.14–0.23), and 0.97 (0.96–0.98), respectively, in the validation dataset and 0.83 (0.78–0.88), 0.83 (0.74–0.90), 0.69 (0.66–0.72), 0.71 (0.68–0.73), 0.23 (0.19–0.28), and 0.97 (0.96–0.98), respectively, in the test dataset. Receiver operating characteristic curves are shown in *Figure 2*. Confusion matrices are shown in Supplementary material online, *Figure S4*. Supplementary material online, *Figure S5* shows the positive predictive value and negative predictive value of the ensemble model according to the AS prevalence in the cohort tested. Additional results are included in the Supplementary material online, *Appendix* (model results by equipment in Supplementary material online, *Table S3*; non-duplication model results in Supplementary material online, *Figure S6* and *Table S4*; and confusion matrices of the non-duplication models in Supplementary material online, *Figure S7*). Saliency maps of the best-performing model (Inception V3) show hot spots in the aortic valve region and in the left ventricle and left atrium, suggesting calcification of the aortic valve and pressure overload on the left cardiac system (*Figure 3*).

Discussion

We describe the development of models for detecting AS from chest radiographs with deep learning. The best-performing model had an AUC of 0.83 in both the validation and test datasets. These results suggest that chest radiographs have the potential to diagnose AS. To our knowledge, this is the first study to create a diagnostic model for AS from chest radiographs, showing that chest radiographs have intrinsic features valuable to help diagnose AS. We used a visualization technique to view the areas on the chest radiographs indicating AS with heat maps[21] and found they agreed with expected changes during AS progression. Heat maps focused on the aortic valve and the left ventricle. The maps included regions of calcification on the aortic valve as signs of AS. These visual findings were consistent with reported findings;[12,13] however, it is difficult for physicians to detect AS exclusively with these findings.

Previous studies have evaluated the relationship between aortic valve calcification visible on chest radiographs and AS progression. In one study,[23] all cases were from a population with aortic valve disease. The sensitivity and specificity of calcification of the aortic valve region on chest radiographs of patients with AS were 0.66 (84/128) and 0.90 (18/20), respectively. Another study,[24] in which all cases were from an AS-positive population, showed that the distinction between severe and moderate to mild AS was determined by calcification in the aortic valve region on chest radiographs with a sensitivity of 0.43 and specificity of 0.88. The authors stated that severe AS can be detected only when calcification is clear in the aortic region. Considering these, our AI model shows higher accuracy than clinician-determined aortic valve calcification on chest radiographs for determining AS. This may be because our model considers the entire image rather than the aortic valve alone. Saliency maps show hot spots not only in the aortic valve region but also in the left ventricular region, especially when the severity of AS increases. As the severity of AS progresses, morphological changes of the heart progress. Thus, our AI models also showed an increase in classification accuracy with increasing severity of AS.

Our ensemble model created with the soft voting technique showed high performance for detecting moderate and severe AS, with sensitivity values that were over 80% in the test. Detection of moderate and severe AS is important because treatments such as surgery are considered. All models performed well, with an AUC of ∼0.8, but the trend in results by severity, valve morphology, and LVEF varied from model to model. For example, sensitivity increased with increasing severity in the Inception model, and conversely, sensitivity decreased with increasing severity in the ResNet model. This may be an effect of the different characteristics learned from the training data by each model. The ensemble model[20] makes use of these differences by blending multiple models to improve the overall predictive performance. Ensemble models are more effective when individual classifiers are not correlated and work by removing the uncorrelated errors of individual classifiers using averaging.

There continue to be significant racial, socioeconomic, and geographic disparities in both access to care and disease outcomes. It has been hypothesized that automated image interpretation can enable more accessible and accurate cardiovascular assessments and begin to alleviate some of the disparities in cardiovascular care.[25,26] Systematic screening of patients at risk for AS has been suggested,[27] but it may be difficult for cardiologists to supply the additional time required to perform and evaluate this screening. Therefore, primary care physicians play an important role in identifying patients with AS.[28] One of the current problems is the low sensitivity and specificity of auscultation.[9] These values for internal medicine residents are even lower.[7] Furthermore, a heart murmur is not an exclusive sign of AS; it can be indicative of several conditions. One study found that only 30% of murmurs were diagnosed as AS after evaluation.[29] Taking these conditions into account, our AI model could serve as a tool to assist primary care physicians as a more objective test. It may be particularly useful for patients who have difficulty accessing echocardiography or who cannot lie still for the duration of an echocardiography examination. However, careful consideration of the consequences of both false positives and negatives is important.[30] In this study, the positive predictive value was low due to the low prevalence of AS in the dataset. In practice, positive predictive value should be increased by adapting the model to a cohort with a high prevalence of AS. For example, in a cohort enriched with AS-negative patients, the model should only be used for patients with symptoms of AS, or to use it in conjunction with auscultation and medical interview to determine history of arterial hypertension, dyslipidaemia, diabetes, smoking, and alcohol use. Additionally, if we want to use the model to find only moderate to severe AS, we can raise the threshold of the model. Another advantage of AI models is that they can be processed in a fraction of the time. For example, if a non-cardiologist is in charge of an emergency room and sees a patient showing signs of heart failure, our AI may help them to better understand the condition and make decisions.

Table 2 Model results

	Validation dataset				Test dataset			
	Inception	ResNet	DenseNet	Ensemble	Inception	ResNet	DenseNet	Ensemble
Area under the curve	0.82 (0.77–0.87)	0.80 (0.74–0.86)	0.80 (0.74–0.85)	0.83 (0.77–0.88)	0.82 (0.77–0.87)	0.80 (0.75–0.86)	0.80 (0.75–0.85)	0.83 (0.78–0.88)
Sensitivity	0.70 (0.59–0.80)	0.77 (0.66–0.85)	0.73 (0.62–0.82)	0.78 (0.67–0.86)	0.74 (0.64–0.82)	0.76 (0.67–0.84)	0.75 (0.65–0.82)	0.83 (0.74–0.90)
Specificity	0.74 (0.72–0.77)	0.70 (0.67–0.73)	0.71 (0.68–0.74)	0.71 (0.68–0.73)	0.71 (0.68–0.74)	0.68 (0.65–0.71)	0.72 (0.70–0.75)	0.69 (0.66–0.72)
Accuracy	0.74 (0.71–0.77)	0.70 (0.67–0.73)	0.71 (0.68–0.74)	0.71 (0.68–0.74)	0.72 (0.69–0.74)	0.69 (0.66–0.72)	0.73 (0.70–0.75)	0.71 (0.68–0.73)
Positive predictive value	0.19 (0.15–0.24)	0.18 (0.14–0.22)	0.17 (0.14–0.22)	0.18 (0.14–0.23)	0.22 (0.18–0.27)	0.21 (0.17–0.25)	0.23 (0.19–0.28)	0.23 (0.19–0.28)
Negative predictive value	0.97 (0.95–0.98)	0.97 (0.96–0.98)	0.97 (0.95–0.98)	0.97 (0.96–0.98)	0.96 (0.94–0.97)	0.96 (0.95–0.98)	0.96 (0.95–0.98)	0.97 (0.96–0.98)
Sensitivity by severity								
Mild	0.55 (0.23–0.83)	0.82 (0.48–0.98)	0.55 (0.23–0.83)	0.64 (0.31–0.89)	0.77 (0.46–0.95)	0.85 (0.55–0.98)	0.62 (0.32–0.86)	0.77 (0.46–0.95)
Moderate	0.68 (0.43–0.87)	0.79 (0.54–0.94)	0.68 (0.43–0.87)	0.68 (0.43–0.87)	0.73 (0.52–0.88)	0.88 (0.70–0.98)	0.85 (0.65–0.96)	0.85 (0.65–0.96)
Severe	0.75 (0.60–0.86)	0.75 (0.60–0.86)	0.78 (0.65–0.89)	0.84 (0.71–0.93)	0.73 (0.61–0.83)	0.70 (0.58–0.81)	0.73 (0.61–0.83)	0.84 (0.73–0.92)
Sensitivity by valvular type								
Tricuspid valve	0.70 (0.59–0.80)	0.76 (0.64–0.85)	0.73 (0.61–0.83)	0.76 (0.64–0.85)	0.73 (0.64–0.82)	0.76 (0.67–0.84)	0.75 (0.66–0.83)	0.83 (0.74–0.90)
Bicuspid valve	0.71 (0.29–0.96)	0.86 (0.42–1.00)	0.71 (0.29–0.96)	1.00 (0.59–1.00)	0.80 (0.28–0.99)	0.80 (0.28–0.99)	0.60 (0.15–0.95)	0.80 (0.28–0.99)
Specificity by valvular type								
Tricuspid valve	0.74 (0.71–0.77)	0.70 (0.67–0.72)	0.71 (0.68–0.74)	0.70 (0.67–0.73)	0.71 (0.68–0.74)	0.68 (0.65–0.71)	0.72 (0.70–0.75)	0.69 (0.66–0.72)
Bicuspid valve	1.00 (0.59–1.00)	1.00 (0.59–1.00)	1.00 (0.59–1.00)	1.00 (0.59–1.00)				
Sensitivity by LVEF								
≥50%	0.73 (0.61–0.83)	0.77 (0.66–0.86)	0.70 (0.58–0.80)	0.74 (0.62–0.84)	0.72 (0.61–0.82)	0.75 (0.64–0.84)	0.75 (0.64–0.84)	0.82 (0.72–0.90)
40–50%	0.50 (0.07–0.93)	0.50 (0.07–0.93)	0.75 (0.19–0.99)	1.00 (0.40–1.00)	0.75 (0.35–0.97)	1.00 (0.63–1.00)	0.62 (0.24–0.91)	0.88 (0.47–1.00)
<40%	0.57 (0.18–0.90)	0.86 (0.42–1.00)	1.00 (0.59–1.00)	1.00 (0.59–1.00)	0.80 (0.52–0.96)	0.73 (0.45–0.92)	0.80 (0.52–0.96)	0.87 (0.60–0.98)
Specificity by LVEF								
≥50%	0.74 (0.71–0.77)	0.70 (0.67–0.73)	0.70 (0.67–0.73)	0.69 (0.66–0.73)	0.75 (0.71–0.78)	0.71 (0.68–0.74)	0.73 (0.70–0.76)	0.72 (0.69–0.76)
40–50%	0.80 (0.69–0.89)	0.71 (0.59–0.82)	0.71 (0.59–0.82)	0.74 (0.62–0.84)	0.58 (0.46–0.70)	0.64 (0.52–0.75)	0.64 (0.52–0.75)	0.58 (0.46–0.70)
<40%	0.76 (0.66–0.83)	0.65 (0.56–0.74)	0.78 (0.68–0.85)	0.77 (0.67–0.84)	0.58 (0.48–0.67)	0.50 (0.41–0.60)	0.72 (0.63–0.80)	0.54 (0.44–0.64)

Data in parentheses are 95% CIs.
LVEF, left ventricular ejection fraction.

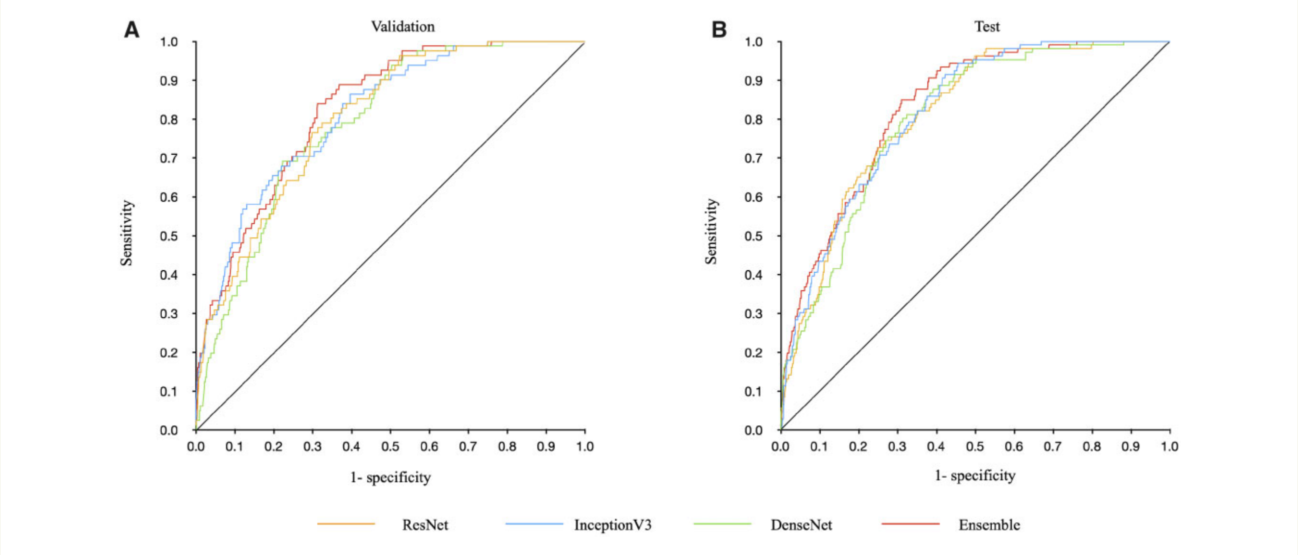


Figure 2 Receiver operating characteristic curves for validation and test datasets of each model. Panels *A* and *B* show the results of the models evaluated in the validation and test datasets, respectively. The orange line shows the results of the ResNet50 model, the blue line shows the results of the InceptionV3 model, the green line shows the results of the DenseNet121 model, and the red line shows the results of the ensemble model. The ensemble model shows the highest area under the receiver operating characteristic curve of 0.83 (0.77–0.88) in the validation dataset and 0.83 (0.78–0.88) in the test dataset.

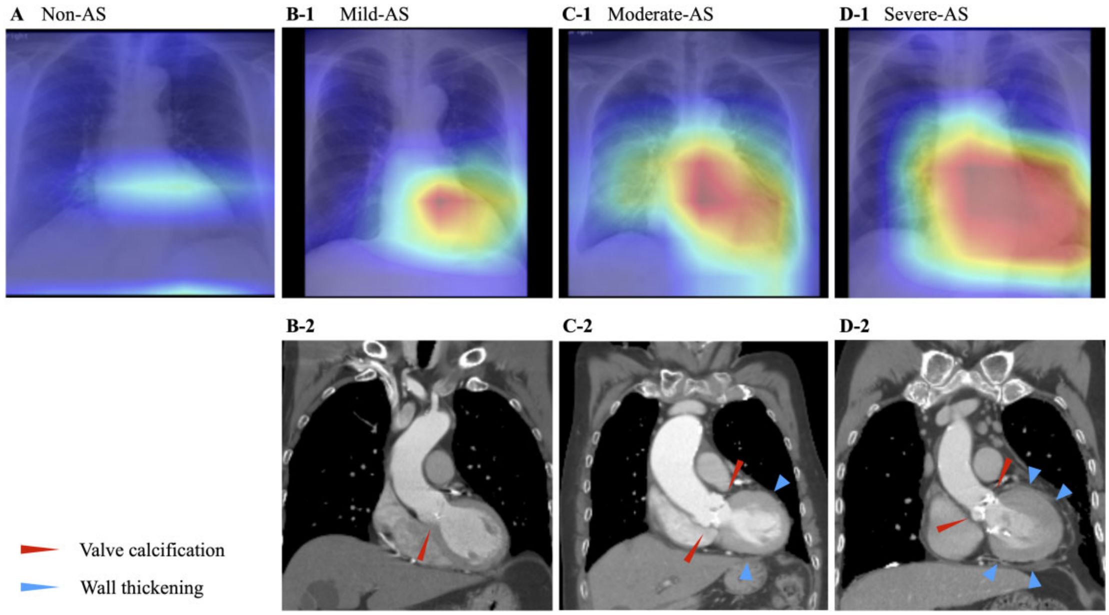


Figure 3 Example saliency maps. These chest radiographs were correctly diagnosed by the Inception V3 model, and the heat maps show the features the model focused on when making the determination of aortic stenosis. Panel *A* shows an 81-year-old asymptomatic man who had echocardiography screening. His aortic valve was normal. Saliency map only shows a very fuzzy hot spot region of interest on the radiograph. Panel *B* shows a 62-year-old man who came to our hospital for further examination of a murmur detected on auscultation. He was diagnosed with mild aortic stenosis. Panel *B-1* shows the saliency map with the radiograph. The hot spot is located on the aortic valve region. Calcifications are visible on the aortic valve (red arrowhead) in the CT image (Panel *B-2*). Panel *C* shows an 85-year-old woman who came to our hospital for further examination regarding exertional dyspnoea. She was diagnosed with moderate aortic stenosis. Panel *C-1* shows the saliency map with the radiograph. The hot spot is located on the aortic valve region and the left ventricular region. Calcifications are visible on the aortic valve (red arrowheads) and left ventricular hypertrophy (blue arrowheads) in the computed tomography image (Panel *C-2*). Panel *D* shows an 86-year-old man who came to our hospital for further examination regarding exertional dyspnoea. He was diagnosed with severe aortic stenosis. Panel *D-1* shows the saliency map with the radiograph. The hot spot covers the aortic valve region and the left ventricular region. Calcifications on the aortic valve (red arrowheads) and left ventricular hypertrophy (blue arrowheads) are visible on the computed tomography image (Panel *D-2*).

Study limitations

There were limitations in this study. Data for this study were collected on equipment from multiple vendors but only from patients at a single centre. Further validation with an external test dataset acquired at another institution would improve the strength of the results. This was also a retrospective study and should be repeated prospectively. Furthermore, the cut-off line between positive or negative model outputs should be set for each intended use and cohort because positive predictive value and negative predictive value vary with the prevalence of AS in a cohort. The main aetiology of AS differs greatly by region and time period; in high-income countries with long life expectancies, age-related degeneration of aortic valve leaflets accounts for the largest proportion of AS. In low- to middle-income countries, rheumatoid AS accounts for a larger proportion of AS. The present study was conducted in a high-income country, and further verification is needed to understand how these differences in aetiology affect the model.

Conclusions

In this study, we developed a deep learning model which has potential to identify patients who have AS using chest radiographs. Our work is open source, and the trained models are available under the Apache 3.0 license. This research can contribute to better health through the detection of AS using chest radiographs.

Supplementary material

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

Acknowledgements

The authors are grateful to Mitsuhiro Inomata and Yoshikazu Hashimoto for technical assistance in regard to developing deep learning.

Funding

This work was supported by the Japan Society for the Promotion of Science (KAKENHI JP20K16769).

Consent: The ethics board of our institution reviewed and approved the protocol of the present study. Since the images had been acquired during daily clinical practice, the need for informed consent was waived.

Conflict of interest: A.S. reports grants from the Japan Society for the Promotion of Science, during the conduct of the study. The other authors have nothing to disclose.

Declaration of Helsinki

This study complies with the declaration of Helsinki. The ethics board of our institution reviewed and approved the protocol of the present study. Since the images had been acquired during daily clinical practice, the need for informed consent was waived.

Data availability

Deidentified participant data upon which the results reported in this article are based (text, tables, and figures) as well as the study protocol will be available. Data will be available beginning 3 months and ending 5 years following article publication. Data will be shared with researchers who provide a methodologically sound proposal. Data will be shared for analyses to achieve the aims in the approved proposal. Proposals should be directed to ai.labo.ocu@gmail.com; to gain access, data requestors will need to sign a data access agreement. The trained model is freely available with Apache License 3.0 (https://github.com/xp-as).

References

1. Danielsen R, Aspelund T, Harris TB, Gudnason V. The prevalence of aortic stenosis in the elderly in Iceland and predictions for the coming decades: the AGES-Reykjavík study. *Int J Cardiol* 2014;**176**:916–922.
2. Eveborn GW, Schirmer H, Heggelund G, Lunde P, Rasmussen K. The evolving epidemiology of valvular aortic stenosis. The Tromsø study. *Heart* 2013;**99**: 396–400.
3. Nkomo VT, Gardin JM, Skelton TN, Gottdiener JS, Scott CG, Enriquez-Sarano M. Burden of valvular heart diseases: a population-based study. *Lancet* 2006;**368**: 1005–1011.
4. Ross J Jr, Braunwald E. Aortic stenosis. *Circulation* 1968;**38**:61–67.
5. Nishimura RA, Otto CM, Bonow RO, et al.; American College of C, American College of Cardiology/American Heart A, American Heart A. 2014 AHA/ACC guideline for the management of patients with valvular heart disease: a report of the American College of Cardiology/American Heart Association Task Force on practice guidelines. *J Thorac Cardiovasc Surg* 2014;**148**:e1–e132.
6. Etchells E, Bell C, Robb K. Does this patient have an abnormal systolic murmur? *JAMA* 1997;**277**:564–571.
7. Mangione S. Cardiac auscultatory skills of physicians-in-training: a comparison of three English-speaking countries. *Am J Med* 2001;**110**:210–216.
8. Mangione S, Nieman LZ. Cardiac auscultatory skills of internal medicine and family practice trainees. A comparison of diagnostic proficiency. *JAMA* 1997;**278**: 717–722.
9. Gardezi SKM, Myerson SG, Chambers J, et al. Cardiac auscultation poorly predicts the presence of valvular heart disease in asymptomatic primary care patients. *Heart* 2018;**104**:1832–1835.
10. De Geer L, Oscarsson A, Engvall J. Variability in echocardiographic measurements of left ventricular function in septic shock patients. *Cardiovasc Ultrasound* 2015;**13**:19.
11. Wood PW, Choy JB, Nanda NC, Becher H. Left ventricular ejection fraction and volumes: it depends on the imaging method. *Echocardiography* 2014;**31**:87–100.
12. Otto CM, Prendergast B. Aortic-valve stenosis—from patients at risk to severe valve obstruction. *N Engl J Med* 2014;**371**:744–756.
13. Webb WR, Higgins CB. *Thoracic Imaging: Pulmonary and Cardiovascular Radiology*. Philadelphia: Lippincott Williams & Wilkins; 2010.
14. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;**521**:436–444.
15. Zoghbi WA, Enriquez-Sarano M, Foster E, et al.; American Society of Echocardiography. Recommendations for evaluation of the severity of native valvular regurgitation with two-dimensional and Doppler echocardiography. *J Am Soc Echocardiogr* 2003;**16**:777–802.
16. Mongan J, Moy L, Kahn CE. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell* 2020;**2**:e200029.
17. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA: IEEE; 2015, p1–9.
18. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE; 2016, p770–778.
19. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI, USA: IEEE; 2017, p4700–4708.
20. Breiman L. Bagging predictors. *Mach Learn* 1996;**24**:123–140.
21. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vision* 2020;**128**:336–359.
22. Clopper C, Pearson E. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 1934;**26**:404–413.

23. Glancy DL, Freed TA, P. O'Brien K, Epstein SE. Calcium in the aortic valve. *Ann Intern Med* 1969;**71**:245–250.
24. Nitta M, Nakamura T, Hultgren HN, Bilisoly J, Marquess B. Noninvasive evaluation of the severity of aortic stenosis in adults. *Chest* 1987;**91**: 682–687.
25. Cohen MG, Fonarow GC, Peterson ED, et al. Racial and ethnic differences in the treatment of acute myocardial infarction: findings from the Get With the Guidelines-Coronary Artery Disease program. *Circulation* 2010;**121**: 2294–2301.
26. Havranek EP, Mujahid MS, Barr DA, et al.; American Heart Association Council on Quality of Care and Outcomes Research, Council on Epidemiology and Prevention, Council on Cardiovascular and Stroke Nursing, Council on Lifestyle and Cardiometabolic Health, and Stroke Council. Social determinants of risk and outcomes for cardiovascular disease: a scientific statement from the American Heart Association. *Circulation* 2015;**132**:873–898.
27. Thoenes M, Bramlage P, Zamorano P, et al. Patient screening for early detection of aortic stenosis (AS)—review of current practice and future perspectives. *J Thorac Dis* 2018;**10**:5584–5594.
28. Bouma BJ, van der Meulen JHP, van den Brink RBA, et al. Variability in treatment advice for elderly patients with aortic stenosis: a nationwide survey in the Netherlands. *Heart* 2001;**85**:196–201.
29. McBrien ME, Heyburn G, Stevenson M, et al. Previously undiagnosed aortic stenosis revealed by auscultation in the hip fracture population—echocardiographic findings, management and outcome. *Anaesthesia* 2009;**64**:863–870.
30. Salmi LR, Coureau G, Bailhache M, Mathoulin-Pélissier S. To screen or not to screen: reconciling individual and population perspectives on screening. *Mayo Clin Proc* 2016;**91**:1594–1605.