Deep learning-based model detects atrial septal defects from electrocardiography: a cross-sectional multicenter hospital-based study

Kotaro Miura,[a,b,h] Ryuichiro Yagi,[c,d,h] Hiroshi Miyama,[a] Mai Kimura,[a] Hideaki Kanazawa,[a] Masahiro Hashimoto,[e] Sayuki Kobayashi,[f] Shiro Nakahara,[f] Tetsuya Ishikawa,[f] Isao Taguchi,[f] Motoaki Sano,[a] Kazuki Sato,[b] Keiichi Fukuda,[a] Rahul C. Deo,[c,d] Calum A. MacRae,[c,d] Yuji Itabashi,[f] Yoshinori Katsumata,[a,b,i,*] and Shinichi Goto[a,c,d,g,i,**]

[a]Department of Cardiology, Keio University School of Medicine, Tokyo, Japan
[b]Institute for Integrated Sports Medicine, Keio University School of Medicine, Tokyo, Japan
[c]Division of Cardiovascular Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, USA
[d]Harvard Medical School, Boston, MA, USA
[e]Department of Radiology, Keio University School of Medicine, Tokyo, Japan
[f]Department of Cardiology, Dokkyo Medical University Saitama Medical Center, Saitama, Japan
[g]Division of General Internal Medicine & Family Medicine, Department of General and Acute Medicine, Tokai University School of Medicine, Isehara, Japan

eClinicalMedicine
2023;63: 102141

Published Online 17 August 2023
https://doi.org/10.1016/j.eclinm.2023.102141

Summary

Background Atrial septal defect (ASD) increases the risk of adverse cardiovascular outcomes. Despite the potential for risk mitigation through minimally invasive percutaneous closure, ASD remains underdiagnosed due to subtle symptoms and examination findings. To bridge this diagnostic gap, we propose a novel screening strategy aimed at early detection and enhanced diagnosis through the implementation of a convolutional neural network (CNN) to identify ASD from 12-lead electrocardiography (ECG).

Methods ECGs were collected from patients with at least one recorded echocardiogram at 3 hospitals from 2 continents (Keio University Hospital from July 2011 to December 2020, Brigham and Women's Hospital from January 2015 to December 2020, and Dokkyo Medical University Saitama Medical Center from January 2010 and December 2021). ECGs from patients with a diagnosis of ASD were labeled as positive cases while the remainder were labeled as negative. ECGs after the closure of ASD were excluded. After randomly splitting the ECGs into 3 datasets (50% derivation, 20% validation, and 30% test) with no patient overlap, a CNN-based model was trained using the derivation datasets from 2 hospitals and was tested on held-out datasets along with an external validation on the 3rd hospital. All eligible ECGs were used for derivation and validation whereas the earliest ECG for each patient was used for the test and external validation. The discrimination of ASD was assessed by the area under the receiver operating characteristic curve (AUROC). Multiple subgroups were examined to identify any heterogeneity.

Findings A total of 671,201 ECGs from 80,947 patients were collected from the 3 institutions. The AUROC for detecting ASD was 0.85–0.90 across the 3 hospitals. The subgroup analysis showed excellent performance across various characteristics Screening simulation using the model greatly increased sensitivity from 80.6% to 93.7% at specificity 33.6% when compared to using overt ECG abnormalities.

Interpretation A CNN-based model using 12-lead ECG successfully identified the presence of ASD with excellent generalizability across institutions from 2 separate continents.

Funding This work was supported by research grants from JST (JPMJPF2101), JSR corporation, Taiju Life Social Welfare Foundation, Kondou Kinen Medical Foundation, Research fund of Mitsukoshi health and welfare foundation, Tokai University School of Medicine Project Research and Internal Medicine Project Research, Secom Science and Technology Foundation, and Grants from AMED (JP23hma922012 and JP23ym0126813). This work was partially supported by One Brave Idea, co-funded by the American Heart Association and Verily with significant support from AstraZeneca and pillar support from Quest Diagnostics.

*Corresponding author. Department of Cardiology, Keio University School of Medicine, 35 Shinanomachi Shinjuku-ku, Tokyo 160-8582, Japan.
**Corresponding author. Division of Cardiovascular Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, USA.
E-mail addresses: goodcentury21@keio.jp (Y. Katsumata), sgoto1@bwh.harvard.edu, sgoto2@tsc.u-tokai.ac.jp (S. Goto).
[h]These authors contributed equally to this work.
[i]These authors shared senior authorship.

Keywords: Atrial septal defect; Artificial intelligence; Deep learning; Electrocardiography

Research in context

Evidence before this study

We searched PubMed using the terms ("deep learning" OR "artificial intelligence") AND "atrial septal defect" AND ("electrocardiogram" OR "electrocardiography") from database inception up to May 8, 2023, with no language restrictions. Only two studies about the deep learning-based prediction of atrial septal defect (ASD) using electrocardiography (ECG) were published. However, both studies had various limitations, including a small sample size, single-center design, and less modern model construction approaches.

Added value of this study

The purpose of our study is to develop a deep learning-based algorithm for predicting ASD using 12-lead ECGs. Our model was constructed using a large multi-center ECG dataset obtained from medical centers across 2 continents. The approach allowed the model to be validated in a multi-country setting including external validation.

Implications of all the available evidence

Our study findings demonstrate that the deep learning-based model developed for detecting ASD using only a single recording of 12-lead ECG shows excellent performance compared to known ECG abnormalities associated with ASD. Our model enhances the screening process, enabling early detection and intervention for ASD closure, and is expected to improve patient outcomes.

Introduction

Atrial septal defect (ASD) is one of the most common adult congenital heart diseases (ACHD). If untreated, ASD increases the risk of irreversible complications such as atrial fibrillation (AF), stroke, heart failure, and pulmonary hypertension.[1–3] While minimally invasive percutaneous closure can be performed in most cases, the disease is largely underdiagnosed due to absent or mild clinical symptoms and subtle findings on physical exam. The diagnosis is usually incidental during medical screening or detected relatively later in life with the onset of symptoms. Early detection is important since early-stage closure has been shown to improve life expectancy and long-term medical follow-up is essential to determine the timing of intervention, if not immediately indicated.[4,5] An effective screening strategy needs to be developed to reduce the complications arising from the disease. Echocardiography is a highly sensitive and accurate imaging modality for detecting ASD. However, the modality is time and labor intense, although non-invasive. Thus, it is not feasible to perform echocardiography on a large population without symptoms.

ECG is another modality used to detect cardiac abnormality. In contrast to echocardiograms, ECGs can be performed within a very short time (approximately 1 min/test) and thus could be performed in large populations including those without symptoms. For the detection of ASD, ECGs are usually used as a screening test to select those who should undergo echocardiograms. While several well-known ECG changes have been observed in patients with ASD, the sensitivity and specificity of ECG abnormalities are limited.[1,6,7] Therefore, a large portion of ASD patients are missed with the current screening approaches using ECG criteria.

The advancement of high-performance computer and deep learning (DL) technologies has enabled the construction of models that detect diseases, predict outcomes, and automate measurements using raw ECG voltage data.[8,9] Several DL models have shown the ability to perform tasks beyond that of expert ECG operators.[10–12] By leveraging the power of DL in detecting subtle ECG changes, we aimed to enhance the screening process for ASD by building an accurate and scalable model to identify patients who should undergo further diagnostic tests with echocardiograms.

Methods

Data sources and study population

This study involved data from three hospitals in Japan and the United States: Keio University Hospital (KUH), which is a tertiary hospital in an urban area in Japan, Brigham and Women's Hospital (BWH), which is a tertiary hospital in an urban area in the United States, and Dokkyo Medical University Saitama Medical Center (SMC), which is a primary/secondary care center in a rural area in Japan. We used the data from KUH and BWH to develop and validate a DL-based algorithm to detect ASD, and the algorithm was externally validated using the data from SMC. Patients ≥18 years, who underwent both ECG and transthoracic echocardiography at KUH (from July 2011 to December 2020) and BWH (from January 2015 to December 2020) were recruited.

The presence or absence of ASD was confirmed using echocardiography. If the ASD was closed during the study period, the date of ASD closure was identified through chart review, and only the ECG data prior to ASD closure were extracted to prevent the model from learning ECG features caused by the closure. Patients from KUH and BWH were randomly assigned to the derivation, validation, and test cohorts in a 5:2:3 ratio using the random package in Python 3.7.3. Since the randomization was performed at a patient level, ECG data from the same patient were not allocated to more than one cohort. In cases where multiple eligible ECGs were available for a single patient, all ECGs were used in the training process to maximize the data. For the test datasets, the earliest ECG for each patient was selected since the more recent ECGs would be likely to be affected by left-right shunt over time which could lead to an overestimation of the model performance compared to the deployment aiming for early detection. The model was trained using data from the derivation dataset. The validation cohort was used for hyperparameter tuning and model selection. The only hyperparameter tuned was the selection of optimizers (Adams vs RMSprop). The final model was tested once using the test datasets from KUH and BWH to evaluate the model's performance. Additionally, to verify the accuracy of the algorithm in a different dataset, external validation was performed using the SMC dataset which was collected using a similar procedure as KUH and BWH, including patients ≥18 years who underwent ECG and echocardiography between January 2010 and December 2021 at SMC. Similar to the test dataset, the earliest ECG for each patient was selected for this dataset.

Data processing

The 12-lead ECG data were stored as measurements of time-series voltage. All the ECGs from KUH were recorded using a Nihon Kohden ECG machine (Nihon Kohden, Tokyo, Japan) with a sampling rate of 1000 Hz whereas those from BWH were recorded with an ECG machine by GE (GE HealthCare, Chicago, USA) at multiple sampling rates (250 Hz and 500 Hz). Conversion of ECG data to matrices was done using our previously published method with slight modifications.[13,14] Briefly, the data were converted to 250 Hz by subsampling to allow the model to utilize all available data. The first 10 s were used if the recording was longer as this was the standard length of 12-lead ECG recording in all the institutions included in the study. Only the data from the standard 12-leads (I, II, III, aVR, aVF, aVL, V1, V2, V3, V4, V5, V6) were used. ECGs with incomplete recordings (e.g., those that did not have data for all 12 leads) were excluded. Finally, the data were saved as a 2D matrix of shape 12 × 2500 in a binary format designed to hold a multidimensional matrix (i.e., NumPy array) (Supplementary Fig. S1).

Model training

For time series data where the current observation depends on previous observations, traditional fully connected networks cannot efficiently learn the data due to the lack of mechanisms to retain past information. Recurrent neural networks (RNN) introduced the concept of memory to neural networks by including the dependency between data points and are suitable for learning time-series data. However, RNNs require sequential processing and thus are limited in the capability of parallelizing the calculations. A convolutional neural network (CNN) is another variant of a neural network that detects local patterns from a multi-dimensional space. In contrast to RNNs, CNNs only require data from the local area and thus could be effectively parallelized. By considering the time axis as one of the axes of the multi-dimensional space, CNNs could detect local patterns across time, which effectively is learning the patterns within the time series. Thus, the current study utilized a 2D-CNN-based model as previously described to achieve optimal training performance on a massively parallel environment using graphics processing units (Fig. 1).[12] The architecture utilized the concept of Inception-Net (multiple CNN with different shapes) and ResNet (residual connection) but was designed to fit the unique shape of the ECG data. The model was trained to minimize the binary cross-entropy against the label (ASD yes/no encoded as 1/0) for 150 epochs (an epoch indicates one pass over the entire training data set analyzed by the algorithm) using the RMSprop optimizer with an initial learning rate of 0.0001, which provided better result compared to Adams optimizer.

To eliminate the need to centralize the data, the model was trained by alternating between the 2 datasets by transferring the weights within institutions (cycle learning, see Supplementary Methods for detail). At the end of each epoch, the performance of the model was evaluated using the validation dataset from KUH and BWH. The model with the best performance, according to the mean of area under the receiver operating characteristics curve (AUROC) KUH and BWH validation datasets, across the 150 epochs was chosen as the final model (Fig. 2). The performance of the final model was evaluated using the test datasets from KUH and BWH along with an external dataset from SMC. To explore the features that the model utilized for prediction, an enhanced class activation mapping technique called GRAD-CAM was applied to the outputs of the model.[15]

Statistical analyses

Baseline characteristics such as age, sex, and body mass index were compared between the ASD and non-ASD groups. No attempts of imputation were made and the statistics were calculated based on complete cases for missing data. Additionally, we conducted comparisons of patient characteristics between the three randomly

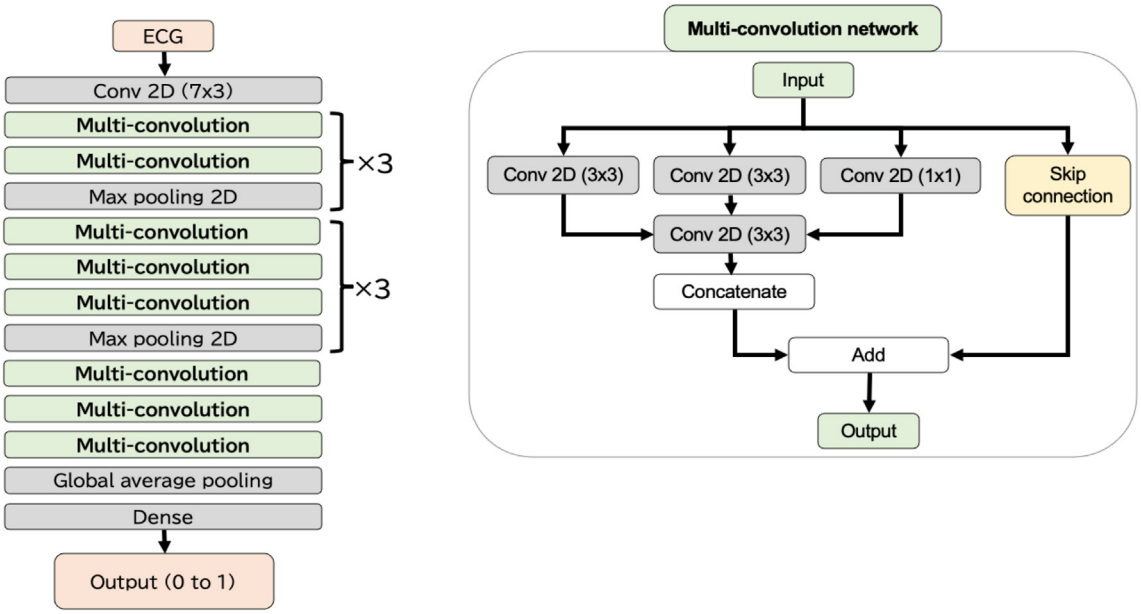

***Fig. 1:* Structure of the neural network of our deep learning model.** Schematic illustration of the neural network model. The details of the multi-convolution cell in the network are shown in the right panel, and the overall network structure is shown in the left panel. The input value is the 2D matrix data of a 12-lead ECG, and the output value takes a continuous value between 0 and 1 and represents the likelihood of the outcome; closer to 1 means a higher probability and closer to 0 means a lower probability. ECG, electrocardiography; Conv 2D, two-dimensional convolution; max pooling 2D, two-dimensional max pooling.

assigned groups in KUH and BWH, as well as among three facilities. For the test cohort in KUH, the most recent hemodynamic parameters including catheter results on Qp/Qs and mean pulmonary artery pressure (PAP), and ASD diameter measured by transesophageal echocardiography in the unclosed state were collected through chart review. For the test cohort in BWH, chart review allowed us to collect up-to-date parameters such as Qp/Qs by catheterization, magnetic resonance imaging (MRI), transthoracic echocardiography, mean PAP by catheterization, and unclosed ASD diameter measured by transesophageal echocardiography, MRI, and transthoracic echocardiography. Continuous values are presented as mean ± standard deviation (SD) and categorical values are presented as numbers and percentages. All probability values were 2-tailed, and P-values < 0.05 were considered to be statistically significant. All the models were trained using TensorFlow 2.4.1. The performance of the model was evaluated by assessing the area under the curve of the receiver operating characteristic (AUROC) with two-sided 95% confidence intervals (CI), accuracy, sensitivity (or recall), specificity, positive predictive value (PPV, or precision) and F1 score (harmonic mean of the precision and recall). The receiver operating characteristic (ROC) curves and precision-recall curves were also plotted. From the model output results, Youden index was calculated, and the cut-offs were calculated in 5% increments of PPV on KUH test datasets. The cut-offs were then assigned to the BWH and SMC datasets, and accuracy, sensitivity, specificity, PPV, and F1 scores were calculated. In addition, AUROC were calculated for groups stratified by age, sex, BMI, and rhythm (i.e., AF or not), and the differences were examined using the DeLong test. The AUROC at KUH were also calculated when ASD group is stratified by hemodynamics, ASD size, and the existence of other congenital heart diseases (see Supplementary Methods). Finally, sensitivity analyses regarding closure indication (Qp/Qs ≥ 1.5) and BMI classification were performed using KUH cohort. All statistical analyses were performed using scikit-learn 0.23.2 and SciPy 1.5.2 packages on Python 3.7.3.

Ethics statement

The study protocol was approved by the Institutional Review Board (IRB) of Keio University School of Medicine (approval number: 20200030) and Mass General Brigham (2019P002651), and conducted in accordance with the Declaration of Helsinki. Due to the retrospective nature of the study and minimal risk of harm, the IRB waived the need for informed consent.

Role of the funding source

The funders had no role in the study design, data collection and analysis, or preparation of the manuscript. They also had no access to datasets and were not involved in any decisions for the submission or publication.

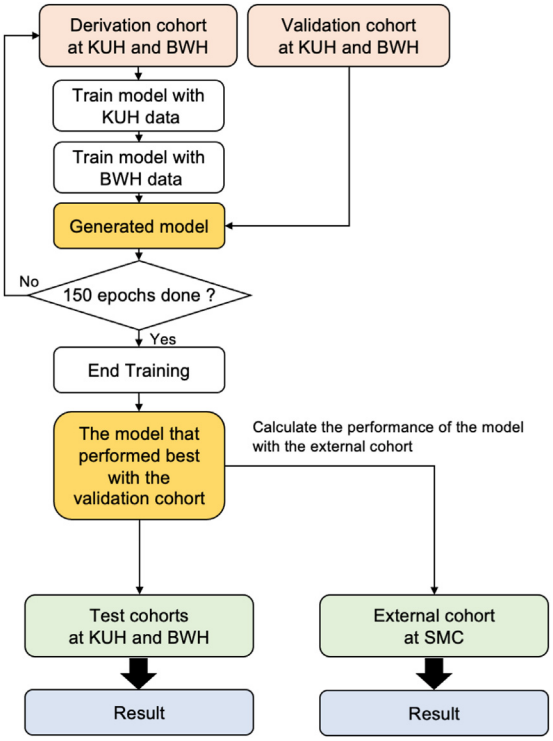

Fig. 2: **Training and testing of the model.** Schematic illustration of the training and testing processes. The model was trained using data from the derivation cohort at KUH and BWH, and the performance of each model was calculated using the validation dataset from KUH and BWH at the end of each epoch. The model that performed the best in the validation cohort during the 150 epochs was chosen as the final model. The performance of the final model was calculated only once using data from the testing dataset at KUH and BWH as internal validation. Furthermore, the performance of the model was verified using the external cohort at SMC. KUH; Keio University Hospital, BWH; Brigham and Women's Hospital; SMC; Dokkyo Medical University Saitama Medical Center.

Results

Data selection

ECG data were collected from 33,431 patients diagnosed with or without ASD by transthoracic echocardiography at KUH (Supplementary Fig. S2). Of these, 857 patients were diagnosed with ASD. ECGs performed after ASD closure were excluded from the analysis since distinct ECG characteristic changes can be expected due to hemodynamic improvement or the closure procedure itself. Of the 666 patients who underwent ASD closure, 532 had at least one ECG prior to the closure procedure. The ECG data of 160 patients without ASD closure were combined, and 692 patients were finally defined as the ASD group. After excluding patients with missing or invalid ECG data, 30,234 patients were allocated to the non-ASD group from the remaining 32,574 patients. A total of 399,056 ECGs were collected for the 30,926 patients and 15,415 (199,922 ECGs), 6164 (77,495 ECGs), and 9347 (9347 ECGs) patients were allocated to derivation, validation, and test cohorts, respectively. Similarly, of 25,067 patients who had transthoracic echocardiography in BWH, 187 had a diagnosis of ASD (Supplementary Fig. S3). After excluding those who did not have available ECG data in the study period or before closure, the cohort consisted of 297 ECGs from 72 patients with a diagnosis of ASD and 246,670 ECGs from 24,771 without a diagnosis of ASD. From the SMC, 25,850 patients underwent both ECG and echocardiography. Among them, 154 had a diagnosis of ASD without closure, and the remaining patients were not diagnosed with ASD. After excluding patients with invalid ECG data, 25,178 patients were included in the external cohort (Supplementary Fig. S4). The prevalence of ASD was 2.24% (692/30,926 patients) at KUH, 0.29% (72/24,843 patients) at BWH, and 0.27% (69/25,178 patients) at SMC (Supplementary Table S1).

Patient characteristics

Patients with ASD were younger (54.2 ± 18.5 vs 63.5 ± 16.0, 41.4 ± 21.0 vs 62.2 ± 16.0, and 57.2 ± 19.6 vs 64.1 ± 15.7 for with-ASD vs non-ASD group in KUH, BWH, and SMC, respectively; all $P < 0.05$) and had lower body mass index (22.0 ± 4.2 vs 22.9 ± 4.2, 22.7 ± 4.7 vs 28.9 ± 8.7, and 22.3 ± 3.8 vs 23.6 ± 10.2 for with-ASD vs non-ASD group in KUH, BWH, and SMC, respectively; all $P < 0.05$) compared to those without ASD (Table 1). Higher prevalence of AF (10.2% vs 5.7% with and without ASD, respectively) and other ACHD (5.3% vs 1.7% with and without ASD, respectively) were observed in KUH, while the prevalence of AF was not different in BWH and SMC. The number of patients with other ACHD was too small for statistical analysis in BWH and was not available in SMC. Similar trends were observed in the derivation and validation datasets from KUH and BWH (Supplementary Tables S2–S5). The baseline characteristics were not different between derivation, validation, and test dataset in KUH and BWH (Supplementary Tables S6 and S7).

Model performance

The model showed excellent discrimination for patients with ASD with AUROCs of 0.90 (95% CI, 0.87–0.93) and 0.88 (95% CI, 0.78–0.99) for the test dataset from KUH and BWH, respectively (Fig. 3a and b) and showed an AUROC of 0.85 (95% CI, 0.80–0.91) in the external dataset at SMC (Fig. 3c) with no statistical differences ($P = 0.52$ for AUROCs between KUH and SMC, and $P = 0.52$ for AUROCs between BWH and SMC). The model performed better compared to the predictions based on overt ECG abnormalities (right bundle branch block, right atrial dilation, and any ECG abnormality) as indicated by the sensitivity/specificity plotted under the ROC curve of the model. Subgroup analyses revealed consistent performance across age, sex, BMI, presence or absence of AF, and presence or absence of any ECG

	KUH							BWH							SMC						
	Total	Missing	ASD	Missing	Non-ASD	Missing	P-value[a]	Total	Missing	ASD	Missing	Non-ASD	Missing	P-value[a]	Total	Missing	ASD	Missing	Non-ASD	Missing	P-value[a]
Number of patients	9347	–	206	–	9141	–	–	7453	–	16	–	7437	–	–	25,178	–	69	–	25,109	–	–
Age, years	63.3 ± 16.4	1204	54.2 ± 18.5	17	63.5 ± 16	1187	P < 0.01	62.1 ± 16.0	0	41.4 ± 21.0	0	62.2 ± 16.0	0	P < 0.01	64.1 ± 15. 7	190	57.2 ± 19.6	0	64.1 ± 15. 7	190	P < 0.01
Male	5250 (56.2%)	0	69 (33.5%)	0	5181 (56.7%)	0	P < 0.01	3823 (52.3%)	0	4 (25.0%)	0	3819 (51.4%)	0	P = 0.06	14,390 (57.6%)	0	30 (43.5%)	0	14,360 (57.6%)	0	P = 0.05
Race									1711		0		1711	P < 0.01							
White								4924 (66.1%)		10 (62.5%)		4914 (66.1%)									
Black								378 (5.1%)		2 (12.5%)		376 (5.1%)									
Asian	9347 (100%)		206 (100%)		9141 (100%)			150 (2.0%)		2 (12.5%)		148 (2.0%)			25,178 (100%)		69 (100%)		25,109 (100%)		
Other								290 (3.9%)		2 (12.5%)		288 (3.9%)									
BMI, kg/m^2	22.9 ± 4.2	1240	22.0 ± 4.2	24	22.9 ± 4.2	1216	P < 0.01	28.9 ± 8.7	1937	22.7 ± 4.7	3	28.9 ± 8.7	1934	P < 0.01	23.6 ± 10.2	681	22.3 ± 3.8	3	23.6 ± 10.2	678	P < 0.01
HR, bpm	72.2 ± 16.6	0	72.3 ± 15.0	0	72.2 ± 16.7	0	P = 0.91	75.8 ± 18.4	0	84.9 ± 19.2	0	75.8 ± 18.4	0	P = 0.08	72.7 ± 17.1	0	73.0 ± 15.9	0	72.7 ± 17.1		P = 0.87
SBP, mmHg								127.3 ± 16.9	1709	116.2 ± 18.6	0	127.3 ± 16.9	1709	P = 0.03							
Normal ECG	3111 (33.3%)	0	40 (19.4%)	0	3071 (33.6%)	0	P < 0.01	1511 (20.3%)	0	3 (18.8%)	0	1508 (20.3%)	0	P = 1.00	7106 (28.2%)	0	16 (23.2%)	0	7096 (28.3%)	0	P = 0.43
AF	543 (5.8%)	0	21 (10.2%)	0	522 (5.7%)	0	P = 0.01	613 (8.2%)	0	1 (6.3%)	0	612 (8.2%)	0	P = 1.00	1552 (6.2%)	0	5 (7.2%)	0	1547 (6.2%)	0	P = 0.90
Other CHD	162 (1.7%)	0	11 (5.3%)	0	151 (1.7%)	0	P < 0.01	–	–	1 (6.3%)	0	–	–	–	–	–	–	–	–	–	–
Qp/Qs	–	–	2.3 ± 0.8	73	–	–	–	–	–	2.0 ± 0.8[b]	5[b]	–	–	–	–	–	–	–	–	–	–
Mean PAP, mmHg	–	–	19.1 ± 6.4	94	–	–	–	–	–	28.0 ± 4.4	13	–	–	–	–	–	–	–	–	–	–
ASD size, mm	–	–	18.2 ± 6.1	65	–	–	–	–	–	18.0 ± 7.9[c]	12[c]	–	–	–	–	–	–	–	–	–	–

Values are presented as mean ± SD or number (percentage). KUH, Keio University Hospital; BWH, Brigham and Women's Hospital; SMC, Dokkyo Medical University Saitama Medical Center; ASD, atrial septal defect; ECG, electrocardiography; CHD, adult congenital heart disease; AF, atrial fibrillation; PAP, pulmonary artery pressure; MRI, magnetic resonance imaging. [a]Difference between ASD and non-ASD groups. [b]Includes data from catheterization, MRI, transesophageal echocardiography, and transthoracic echocardiography. Measurement priority is in this order. [c]Includes data from transesophageal echocardiography, MRI, and transthoracic echocardiography. Measurement priority is in this order.

Table 1: **Baseline characteristics in the test cohorts at KUH, BWH, and SMC in patient level.**

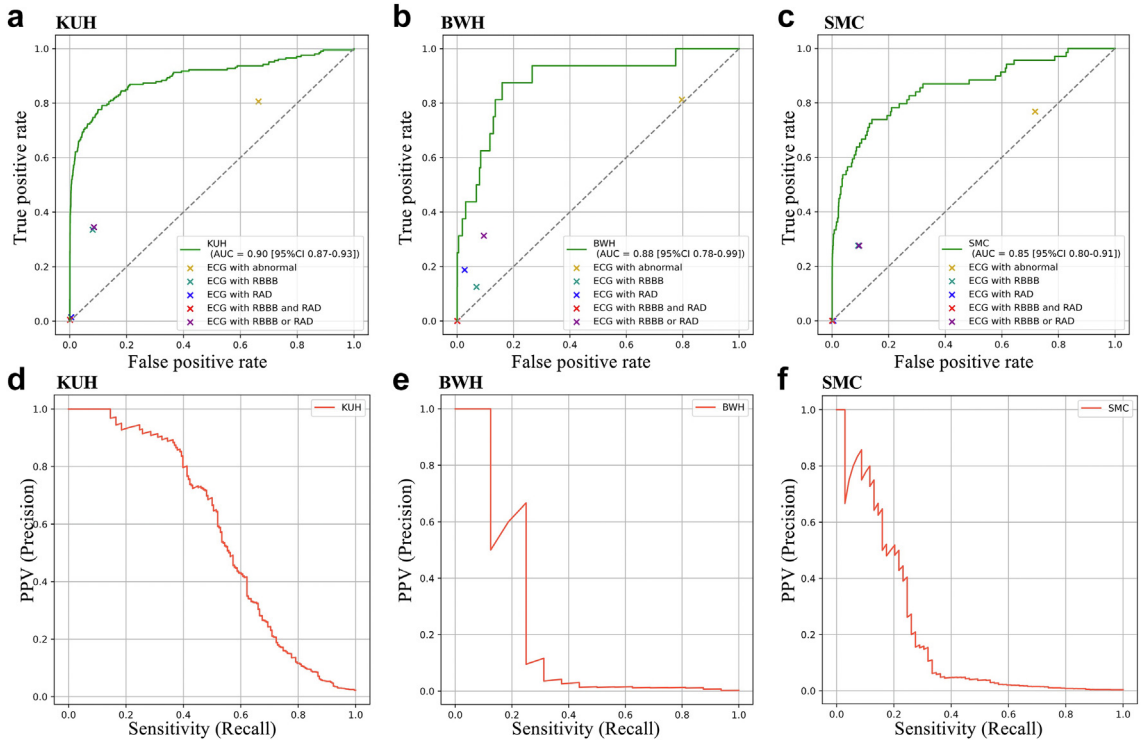

***Fig. 3:* Performance of the final model for the detection of ASD.** (a) ROC curve for the internal test cohort at KUH is shown as a solid green line. Sensitivities/specificities of common ECG abnormalities within the cohort are shown as cross-marks. (b) ROC curve for the internal test cohort at BWH is shown as a solid green line. Sensitivities/specificities of common ECG abnormalities within the cohort are shown as cross-marks. (c) ROC curve for the external validation cohort at SMC is shown as a solid green line. Sensitivities/specificities of common ECG abnormalities within the cohort are shown as cross-marks. (d) The precision-recall curve for the internal test cohort at KUH is shown as a solid red line. (e) The precision-recall curve for the internal test cohort at BWH is shown as a solid red line. (f) The precision-recall curve for the external cohort at SMC is shown as a solid red line. ASD, atrial septal defect; ROC, receiver operating characteristics; KUH, Keio University Hospital, BWH, Brigham and Women's Hospital; SMC, Dokkyo Medical University Saitama Medical Center; ECG, electrocardiography; RBBB, right bundle branch block; RAD, right axis deviation; PPV, positive predictive value; CI, confidence interval.

abnormalities in the three institutions (Fig. 4a–e for KUH, Supplementary Fig. S5a–e for BWH, Supplementary Fig. S6a–e for SMC, and Supplementary Table S8 for statistical analyses between the subgroups). The model performance was also consistent across races in BWH (AUROC: 0.86 [95% CI, 0.71–1.00] for White; 0.91 [95% CI, 0.64–1.00] for Black, 0.87 [95% CI, 0.55–1.00] for Asian, and 0.96 [95% CI, 0.78–1.00] for others; Supplementary Fig. S5f). While the model showed robust performance for mean PAP (AUROC: 0.94 [95% CI, 0.90–0.99] for mean PAP ≥20 mmHg and 0.88 [95% CI, 0.83–0.93] for mean PAP <20 mmHg; Fig. 4f), the model tended to discriminate severe ASD better: Qp/Qs (AUROC: 0.99 [95% CI, 0.96–1.00] for Qp/Qs ≥ 2.5, 0.90 [95% CI, 0.85–0.95] for 1.5 ≤ Qp/Qs < 2.5, and 0.76 [95% CI, 0.63–0.88] for Qp/Qs < 1.5; Fig. 4g), ASD size (AUROC: 0.95 [95% CI, 0.88–1.00] for size ≥25 mm, 0.91 [95% CI, 0.87–0.95] for 10 ≤ size <25 mm, and 0.65 [95% CI, 0.45–0.84] for size <10 mm; Fig. 4h). The model showed robust performance for detecting ASD without other ACHD (AUROC: 0.90 [95% CI, 0.88–0.93] for simple ASD; Fig. 4i). These analyses could not be performed on other institutions due to the small number of patients in the subgroup or the missingness of the data. GRAD-CAM analyses revealed that the model focused primarily on the P wave and QRS complex in the limb leads, regardless of the presence of overt ECG abnormality (Fig. 5).

Sensitivity analyses

The model detected ASD cases with closure indication (defined as Qp/Qs ≥ 1.5) with AUROC 0.91 (95% CI, 0.88–0.94, Supplementary Fig. S7a). The model performance was also consistent across different BMI thresholds (AUROC 0.94 [95% CI, 0.89–1.00], 0.90 [95% CI, 0.86–0.94], and 0.90 [95% CI, 0.90–0.98] for those with BMI < 18.5, 18.5 ≤ BMI < 25, and BMI ≥ 25, respectively, Supplementary Fig. S7b).

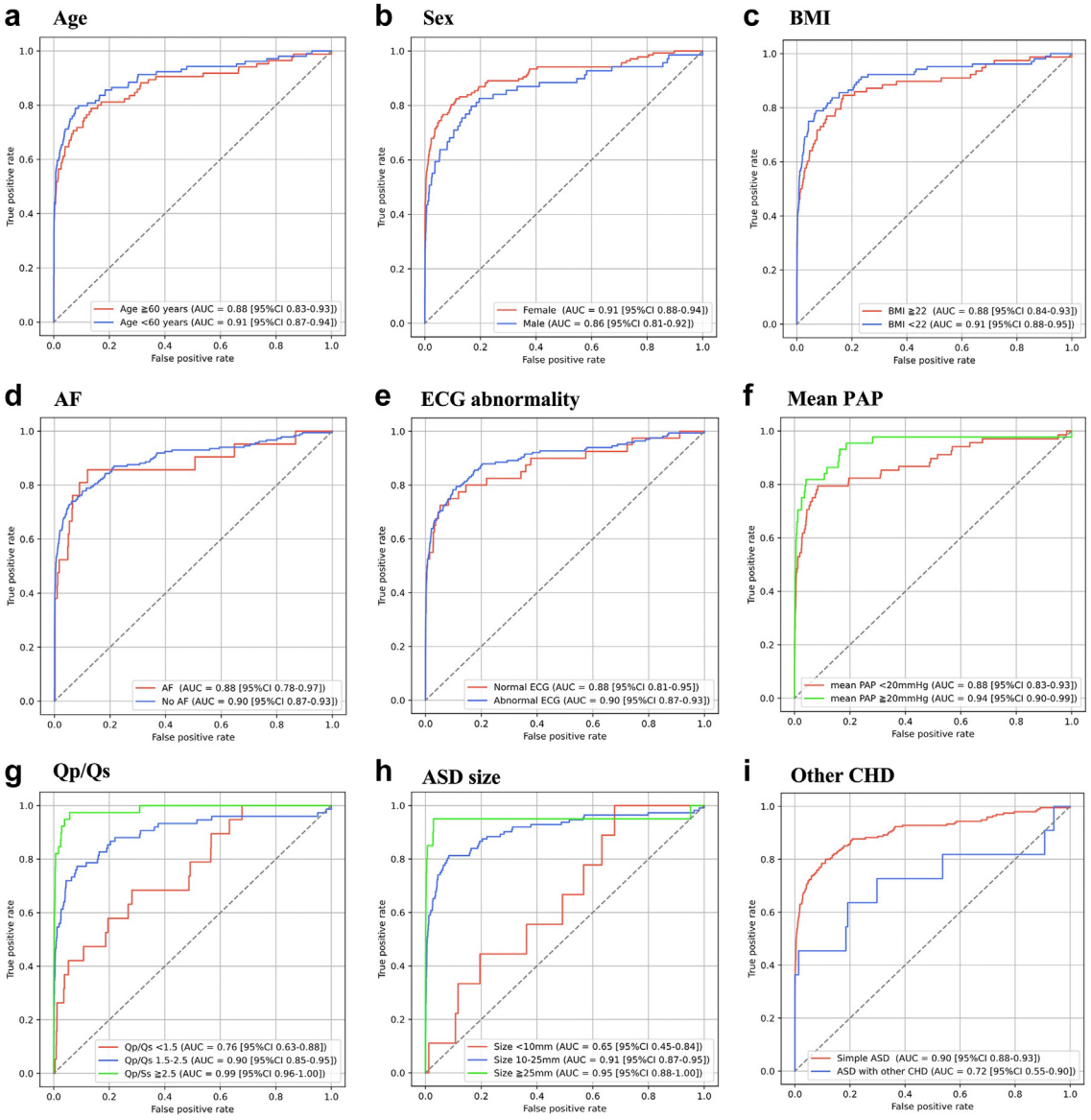

Fig. 4: **Performance of the model for detecting ASD stratified according to subgroups in KUH.** (a) The ROC curve is shown for the internal test cohort stratified according to age. (b) The ROC curve is shown for the internal test cohort stratified according to sex. (c) The ROC curve is shown for the internal test cohort stratified according to BMI. (d) The ROC curve is shown for the internal test cohort stratified according to AF or not. (e) The ROC curve is shown for the internal test cohort stratified according to normal ECG or not. (f) The ROC curve is shown for the internal test cohort stratified according to mean PAP. (g) The ROC curve is shown for the internal test cohort stratified according to Qp/Qs. (f) The ROC curve is shown for the internal test cohort stratified according to ASD size. (i) The ROC curve is shown for the internal test cohort stratified by other CHD. KUH, Keio University Hospital; ASD, atrial septal defect; ROC, receiver operating characteristic; AUC, area under the curve; CI, confidence interval; BMI, body mass index; AF, atrial fibrillation; PAP, Pulmonary artery pressure; CHD, congenital heart disease.

Deployment simulation

When the probability threshold for ASD detection in the KUH dataset was set to the Youden index (0.000199), the accuracy, sensitivity, specificity, and PPV were 88.5%, 79.1%, 88.7%, and 13.6%, respectively. At the cut-off of 0.000082, which demonstrated a 10% PPV at KUH, the sensitivity/specificity was 82.5/83.3% at KUH, 68.8/87.5% at BWH, and 68.1/91.2% at SMC. When adjusting the cut-off to 0.001192 to achieve a 25% PPV at KUH, the sensitivity/specificity remained 69.4/95.3% at KUH, 43.8/96.7% at BWH, and 50.7/96.8% at SMC. These results indicate a relatively consistent performance in the three cohorts, across various cut-off points (Supplementary Tables S9–S11). When

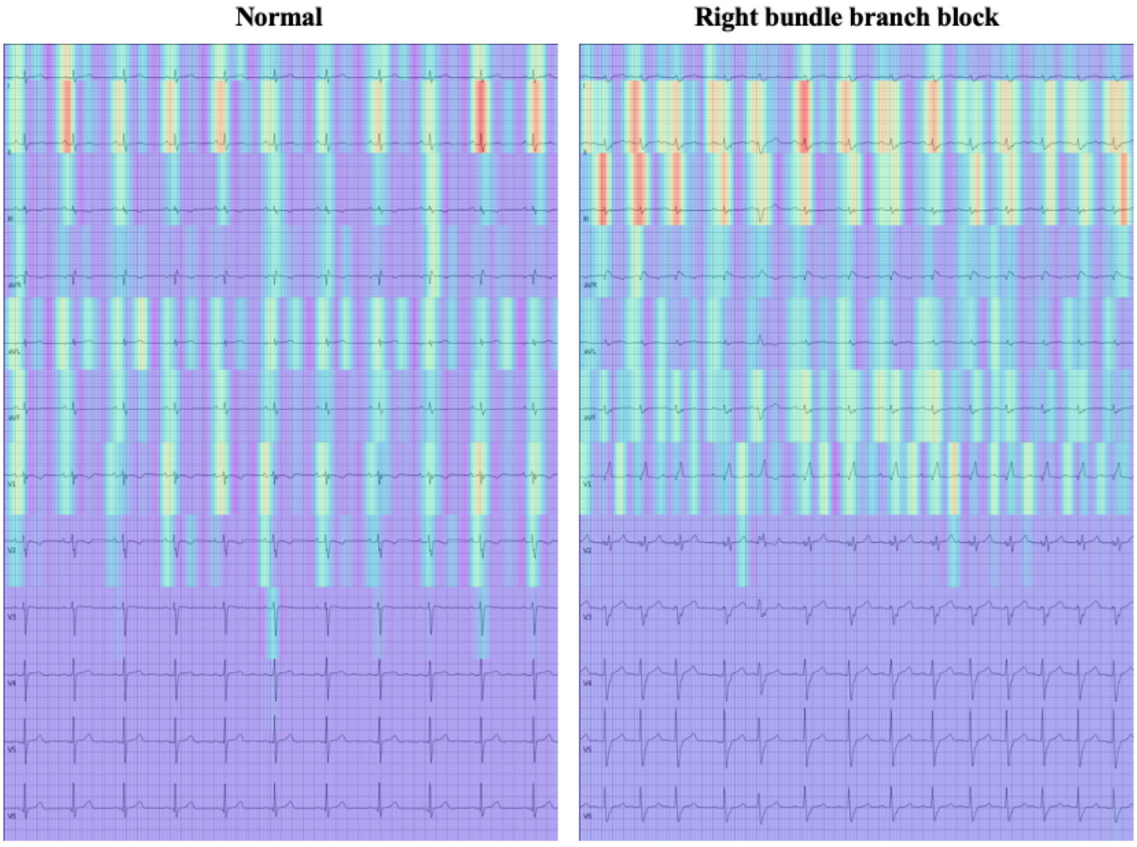

Fig. 5: **Gradient-weighted class activation mapping images.** Gradient-weighted class activation mapping (GRAD-CAM) images for ASD samples with and without overt abnormality. The primary focus of the model is indicated by the colored areas. ASD, atrial septal defect.

compared with the performance of conventional ECG readings, the model showed great improvement in sensitivity. In the KUH dataset, the presence of any ECG abnormality had a sensitivity of 80.6% and a specificity of 33.6%. In contrast, the model demonstrated a much higher sensitivity of 93.7% at the same specificity. Similar improvements in sensitivities were observed for the other institutions as well (Supplementary Table S12). *With respect to detecting ASD cases with closure indication, the model showed a better diagnostic performance at the same sensitivity* (specificities, 90.3% vs 88.7%; PPVs, 14.1% vs 13.6% at a sensitivity of 79.1%) with a higher cut-off of 0.00027.

Discussion

In this study, we showed that a DL algorithm detected ASD with excellent performance using 12-lead ECG data. Due to subtle symptoms and examination findings, ASD is largely underdiagnosed even though it is treatable if diagnosed promptly. The presence of ASD increases the risk of irreversible diseases such as AF, stroke, and heart failure, and early diagnosis is thus essential. The diagnosis of ASD in adults is usually triggered by abnormalities noted on ECGs, chest X-rays, or heart murmur noted on physical examination, as well as by incidental echocardiography, arrhythmias, or stroke.[16] Since ECG is a simple tool for identifying possible ASD, several studies on ECG and ASD have been conducted. Right atrial loading with a right-left shunt can result in increased P-wave amplitude as well as prolonged P-wave duration and increased P-wave dispersion due to delayed atrial conduction,[17,18] and right ventricular loading can produce right bundle branch block.[19] However, these findings are observed in many cardiac diseases other than ASD. Given the sensitivity/specificity associated with these findings, they alone are unlikely to raise suspicion for ASD in a real clinical setting.[20] The crochetage pattern is another characteristic ECG finding of ASD. The specificity of the crochetage pattern was reported as 92–100% when it appeared in the three inductions, II, III, and aVF. However, the sensitivity was limited (28–73%), and it was also reported to appear in patients with ventricular septal defect, pulmonary valve stenosis, and healthy individuals.[6,21] Considering the utility of ECG for screening where false negatives need to be reduced, suspecting ASD based on these ECG findings is unsatisfactory. Consequently, diagnosis of ASD is often delayed, and the condition remains undetected until other

overt symptoms develop in a large number of individuals. In contrast to the conventional ECG abnormalities, our model demonstrated much higher sensitivity when achieving the same specificity. Therefore, our model has a strong potential to reduce missed ASD when deployed, by serving as an effective tool to suggest further investigation with echocardiograms, utilizing the widely available and inexpensive ECG.

In addition to the ability of the model to "rank" the probability of the disease (measured by AUROC), the consistency of sensitivity/specificity for the cutoff points across multiple institutions is an important factor to determine the utility of the model. Even with a high AUROC, if the sensitivity/specificity differs substantially across institutions, the cutoff of the model will need to be re-calculated whenever the model is deployed in a new institution. Since the determination of cutoffs requires a large number of patients, this is not always feasible and thus greatly limits the applicability of the model. Hence, we analyzed the performance of our model across various cutoff points utilizing our multi-institutional data, which revealed similar sensitivities and specificities across all institutions. This finding emphasizes the generalized performance of the model and supports its utility without re-calculating the cutoff for new institutions.

In recent years, due to the rapid growth of available medical data, various DL models have been developed to aid diagnosis and treatment.[22] In the area of cardiology, ECGs have been an attractive target for developing DL models.[23] DL models can extract unrecognized information from ECGs and incorporate minute and bias-free features that are often missed by the human eye, ultimately enabling the construction of more accurate algorithms. Research to improve neural networks has produced various architectures that can be applied to different data structures, for example, models using CNNs[11,24,25] and models using recurrent neural networks (RNN) including long short-term memory (LSTM).[13,14,26–28] In this study, the DL model was constructed using CNNs. While RNNs are suitable for learning time series, they require data from the previous time step, which prevent massive parallelization. In contrast, CNN can process each filter independently allowing the training process to be better parallelized. This characteristic of CNN enabled us to train a larger network compared to our previous work, which detected potential cardiac amyloidosis.[12] We have expanded on this prior work to detect ASD from 12-lead ECG. The DL-enhanced ECG approach could also be applied for the detection of other congenital diseases.[29] With the advancement of artificial intelligence and the efficient collection of longitudinal congenital heart disease data from large registries or real-world databases, it is anticipated that more effective detection of congenital heart diseases will become possible in the near future.

A clear strength of this study was the validation, which was performed in an ethnically heterogeneous population and at an external institution. It has been shown that models tested only at single institutions often display unexpectedly low performances on an external dataset.[30,31] Our model not only showed excellent discrimination on the KUH dataset (AUROC 0.90) and BWH dataset (AUROC 0.88) but also generalized well to a dataset from an external institution (AUROC 0.85 on the SMC dataset), proving the external validity of the algorithm. The prevalence of ASD among the three institutions differs. The particularly high prevalence of ASD at KUH may strongly reflect the characteristics of the KUH facility; KUH is a tertiary care center with multiple specialists in surgical or percutaneous ASD closure and is a high-volume medical center for ASD treatment where ASD patients are referred from a very wide geographic area. On the other hand, BWH and SMC are also advanced medical facilities, but the prevalence of ASD was considerably closer to the general population which is reported to be around 0.085%.[32] The fact that our model performed well on such diverse populations strongly suggests the applicability of the model in real-world practice.

For a thorough understanding of the results, several limitations should be acknowledged. First, KUH is an academic institution that deals with various uncommon diseases, as evidenced by the high prevalence of ASD. Furthermore, since the entire cohort consisted of patients who underwent echocardiography, the results could have been affected by selection bias. To minimize such concerns, we validated the model using data from an academic medical center and a community hospital, through internal and external validation, which demonstrated excellent performance in both settings with different prevalence rates and patient backgrounds. However, further prospective studies in the general populations are needed to validate the applicability of our model to unselected asymptomatic individuals. Second, since this model was trained with ASDs detected by transthoracic echocardiography, it is possible that small ASDs could have been missed and thus misclassified into the non-ASD group. The model also showed a tendency for lower AUROC for detecting less severe ASD. However, since patients with small ASDs usually do not require closure, it is unlikely to cause major clinical problems even if the sensitivity to detect very small ASDs is low. Third, we used all available ECGs rather than matching age or sex between cases and controls. While this approach enabled us to evaluate the model performance on a population reflecting the actual prevalence of ASD at each institution, it could be possible that the model learned the difference in age and sex within the cohort. However, the results of the subgroup analysis showed that there were no differences in performance across different age

or sex groups, which suggests that the influence, if it exists, is small. Fourth, while not statistically significant, the discrimination of ASD was numerically lower for ASD with other ACHDs. However, those with ACHDs likely had an indication for echocardiograms and thus a screening strategy for ASD in this population is less important. Fifth, since some clinical information including medication and comorbidity was not available due to the nature of our datasets, their potential impacts on the model performance could not be assessed. Sixth, the model performance could have been affected by the inter-rater variability for detecting ASD (particularly small ASD). Since we did not have data to identify the individual technicians who performed the examination, we were not able to directly assess the impact of the variability. However, our data showing consistent performance across different institutions support that the impact was minimal. It should also be noted that all the examinations and associated reports were over-read by at least one experienced specialist. Finally, one of the key limitations in DL is the lack of explainability. We attempted to mitigate this limitation by performing a GRAD-CAM analysis, which revealed that the model focused on P waves and QRS complexes. However, the result only provides information on "where" the feature was and does not show "what" the feature was. Thus, the model is not able to fully provide information for humans to understand novel features to look for in ECGs.

In conclusion, this study showed that a neural network-based DL algorithm using 12-lead ECG data can detect ASD excellently with good generalization. The model can be used to improve ASD screening, where symptoms and laboratory findings are subtle.

Contributors

KM and RY contributed equally to this work (co-first author). KM, RY, SG, and YK designed the study. KM, SG, MK, HK, MH, and YK collected, cleaned, and labeled the dataset from KUH; RY, SG, RD and CM collected, cleaned, and labeled the dataset from BWH; KM, SG, YK, and YI collected, cleaned, and labeled the dataset from SMC. RY and SG built the deep learning models. KM, RY, SG, and YK performed the analysis and verified the data. We ensured that all authors had access to all the datasets. KM, RY, SG, and YK wrote the manuscript and verified the data, and HM, SK, SN, TI, IT, MS, KS, RD, CM, and KF provided critical feedback on the manuscript. All authors were responsible for the decision to submit the manuscript.

Data sharing statement

The data and codes that support the findings of this study are available in a public repository (https://github.com/obi-ml-public/ECG-ASD/tree/main). We provide a web-interface to run our model and generate predictions at http://onebraveideaml.org. De-identified patient-level clinical data for KUH and SMC will be provided upon reasonable request.

Declaration of interests

We declare no competing interests.

Acknowledgements

This work was supported by research grants from JST (JPMJPF2101), JSR corporation, Taiju Life Social Welfare Foundation, Kondou Kinen Medical Foundation, Research fund of Mitsukoshi health and welfare foundation, Tokai University School of Medicine Project Research and Internal Medicine Project Research, Secom Science and Technology Foundation, and Grants from AMED (JP23hma922012 and JP23ym0126813). This work was partially supported by One Brave Idea, co-funded by the American Heart Association and Verily with significant support from AstraZeneca and pillar support from Quest Diagnostics.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2023.102141.

References

1 Webb G, Gatzoulis MA. Atrial septal defect in the adult recent progress and overview. *Circulation*. 2006;114:1645–1653.
2 Geva T, Martins JD, Wald RM. Atrial septal defects. *Lancet*. 2014;383:1921–1932.
3 Udholm S, Nyboe C, Karunanithi Z, et al. Lifelong burden of small unrepaired atrial septal defect: results from the Danish National Patient Registry. *Int J Cardiol*. 2019;283:101–106.
4 Nyboe C, Karunanithi Z, Nielsen-Kudsk JE, Hjortdal VE. Long-term mortality in patients with atrial septal defect: a nationwide cohort-study. *Eur Heart J*. 2018;39:993–998.
5 Brida M, Diller GP, Kempny A, et al. Atrial septal defect closure in adulthood is associated with normal survival in the mid to longer term. *Heart*. 2019;105:1014–1019.
6 Heller J, Hag AA, Besse B, Desngs M, Marie F, Guerot C. "Crochetage" (Notch) on R wave in inferior limb leads: a new independent electrocardiographic sign of atrial septal defect. *J Am Coll Cardiol*. 1996;27:877–882.
7 Izumida N, Asano Y, Wakimoto H, et al. Analysis of T wave changes by activation recovery interval in patients with atrial septal defect. www.elsevier.com/locate/ijcard; 2000.
8 Goto S, McGuire DK, Goto S. The future role of high-performance computing in cardiovascular medicine and science -impact of multi-dimensional data analysis. *J Atherosclerosis Thromb*. 2022;29:559–562.
9 Goto S, Goto S. Application of neural networks to 12-lead electrocardiography — current status and future directions —. *Circ Rep*. 2019;1:481–486.
10 Raghunath S, Ulloa Cerna AE, Jing L, et al. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nat Med*. 2020;26:886–891.
11 Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet*. 2019;394:861–867.
12 Goto S, Mahara K, Beussink-Nelson L, et al. Artificial intelligence-enabled fully automated detection of cardiac amyloidosis using electrocardiograms and echocardiograms. *Nat Commun*. 2021;12:2726. https://doi.org/10.1038/s41467-021-22877-8.
13 Goto S, Kimura M, Katsumata Y, et al. Artificial intelligence to predict needs for urgent revascularization from 12-leads electrocardiography in emergency patients. *PLoS One*. 2019;14:e0210103. https://doi.org/10.1371/journal.pone.0210103.
14 Miura K, Goto S, Katsumata Y, et al. Feasibility of the deep learning method for estimating the ventilatory threshold with electrocardiography data. *NPJ Digit Med*. 2020;3:141. https://doi.org/10.1038/s41746-020-00348-6.
15 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 Computer Vis*. 2016;128:336–359. https://doi.org/10.1007/s11263-019-01228-7. published online Oct 7.
16 Akagi T. Current concept of transcatheter closure of atrial septal defect in adults. *J Cardiol*. 2015;65:17–25.
17 Grignani RT, Tolentino KM, Rajgor DD, Quek SC. Longitudinal evaluation of P-wave dispersion and P-wave maximum in children after transcatheter device closure of secundum atrial septal defect. *Pediatr Cardiol*. 2015;36:1050–1056.
18 Thilén U, Carlson J, Platonov PG, Olsson SB. Atrial myocardial pathoelectrophysiology in adults with a secundum atrial septal defect is unaffected by closure of the defect. A study using high resolution signal-averaged orthogonal P-wave technique. *Int J Cardiol*. 2009;132:364–368.

19 Gaba P, Pedrotty D, Desimone CV, Bonikowske AR, Allison TG, Kapa S. Mortality in patients with right bundle-branch block in the absence of cardiovascular disease. *J Am Heart Assoc*. 2020;9: e017430. https://doi.org/10.1161/JAHA.120.017430.

20 Maheshwari A, Norby FL, Soliman EZ, et al. Relation of prolonged P-wave duration to risk of sudden cardiac death in the general population (from the atherosclerosis risk in communities study). *Am J Cardiol*. 2017;119:1302–1306.

21 Khairy P, Marelli AJ. Clinical use of electrocardiography in adults with congenital heart disease. *Circulation*. 2007;116:2734–2746.

22 Deo RC. Machine learning in medicine. *Circulation*. 2015;132: 1920–1930.

23 Somani S, Russak AJ, Richter F, et al. Deep learning and the electrocardiogram: review of the current state-of-the-art. *Europace*. 2021;23:1179–1191.

24 Kwon JM, Lee SY, Jeon KH, et al. Deep learning–based algorithm for detecting aortic stenosis using electrocardiography. *J Am Heart Assoc*. 2020;9:e014717. https://doi.org/10.1161/JAHA.119.014717.

25 Ko WY, Siontis KC, Attia ZI, et al. Detection of hypertrophic cardiomyopathy using a convolutional neural network-enabled electrocardiogram. *J Am Coll Cardiol*. 2020;75:722–733.

26 Oh SL, Ng EYK, Tan RS, Acharya UR. Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats. *Comput Biol Med*. 2018;102: 278–287.

27 Mori H, Inai K, Sugiyama H, Muragaki Y. Diagnosing atrial septal defect from electrocardiogram with deep learning. *Pediatr Cardiol*. 2021;42:1379–1387.

28 Petmezas G, Haris K, Stefanopoulos L, et al. Automated atrial fibrillation detection using a hybrid CNN-LSTM network on imbalanced ECG datasets. *Biomed Signal Process Control*. 2021;63: 102194. https://doi.org/10.1016/j.bspc.2020.102194.

29 Jone P-N, Gearhart A, Lei H, et al. Artificial intelligence in congenital heart disease. *JACC Adv*. 2022;1:100153.

30 Goto S, Solanki D, John JE, et al. Multinational federated learning approach to train ECG and echocardiogram models for hypertrophic cardiomyopathy detection. *Circulation*. 2022;146:755–769.

31 Yagi R, Goto S, Katsumata Y, MacRae CA, Deo RC. Importance of external validation and subgroup analysis of artificial intelligence in the detection of low ejection fraction from electrocardiograms. *Eur Heart J Digit Health*. 2022;3(4):654–657. https://doi.org/10.1093/ehjdh/ztac065. published online Nov 2.

32 Marelli AJ, Ionescu-Ittu R, Mackie AS, Guo L, Dendukuri N, Kaouache M. Lifetime prevalence of congenital heart disease in the general population from 2000 to 2010. *Circulation*. 2014;130:749–756.