





# An ensemble learning model for detection of pulmonary hypertension using electrocardiogram, chest X-ray, and brain natriuretic peptide

Risa Kishikawa <sup>1</sup>, Satoshi Kodera<sup>1,\*</sup>, Naoto Setoguchi <sup>1</sup>, Kengo Tanabe<sup>2</sup>, Shunichi Kushida<sup>3</sup>, Mamoru Nanasato <sup>4</sup>, Hisataka Maki<sup>5</sup>, Hideo Fujita<sup>5</sup>, Nahoko Kato<sup>6</sup>, Hiroyuki Watanabe<sup>6</sup>, Masao Takahashi<sup>7</sup>, Naoko Sawada<sup>8</sup>, Jiro Ando<sup>8</sup>, Masataka Sato<sup>1</sup>, Shinnosuke Sawano<sup>1</sup>, Hiroki Shinohara<sup>1</sup>, Koki Nakanishi <sup>1</sup>, Shun Minatsuki<sup>1</sup>, Junichi Ishida<sup>1</sup>, Katsuhito Fujii<sup>1,9</sup>, Hiroshi Akazawa<sup>1</sup>, Hiroyuki Morita<sup>1</sup>, and Norihiko Takeda<sup>1</sup>

<sup>1</sup>Department of Cardiovascular Medicine, The University of Tokyo Hospital, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8655, Japan; <sup>2</sup>Division of Cardiology, Mitsui Memorial Hospital, Tokyo, Japan; <sup>3</sup>Department of Cardiovascular Medicine, Asahi General Hospital, Chiba, Japan; <sup>4</sup>Department of Cardiology, Sakakibara Heart Institute, Tokyo, Japan; <sup>5</sup>Division of Cardiovascular Medicine, Saitama Medical Center, Jichi Medical University, Omiya, Japan; <sup>6</sup>Department of Cardiology, Tokyo Bay Urayasu Ichikawa Medical Center, Urayasu, Japan; <sup>7</sup>Department of Cardiology, JR General Hospital, Tokyo, Japan; <sup>8</sup>Department of Cardiology, NTT Medical Center Tokyo, Tokyo, Japan; and <sup>9</sup>Department of Advanced Cardiology, The University of Tokyo, Tokyo, Japan

Received 9 August 2024; revised 5 October 2024; accepted 31 October 2024; online publish-ahead-of-print 16 January 2025

## Aims

Delayed diagnosis of pulmonary hypertension (PH) is a known cause of poor patient prognosis. We aimed to develop an artificial intelligence (AI) model, using ensemble learning method to detect PH using electrocardiography (ECG), chest X-ray (CXR), and brain natriuretic peptide (BNP), facilitating accurate detection and prompting further examinations.

## Methods and results

We developed a convolutional neural network model using ECG data to predict PH, labelled by ECG from seven institutions. Logistic regression was used for the BNP prediction model. We referenced a CXR deep learning model using ResNet18. Outputs from each of the three models were integrated into a three-layer fully connected multimodal model. Ten cardiologists participated in an interpretation test, detecting PH from patients' ECG, CXR, and BNP data both with and without the ensemble learning model. The area under the receiver operating characteristic curves of the ECG, CXR, BNP, and ensemble learning model were 0.818 [95% confidence interval (CI), 0.808–0.828], 0.823 (95% CI, 0.780–0.866), 0.724 (95% CI, 0.668–0.780), and 0.872 (95% CI, 0.829–0.915). Cardiologists' average accuracy rates were 65.0 ± 4.7% for test without AI model and 74.0 ± 2.7% for test with AI model, a statistically significant improvement ( $P < 0.01$ ).

## Conclusion

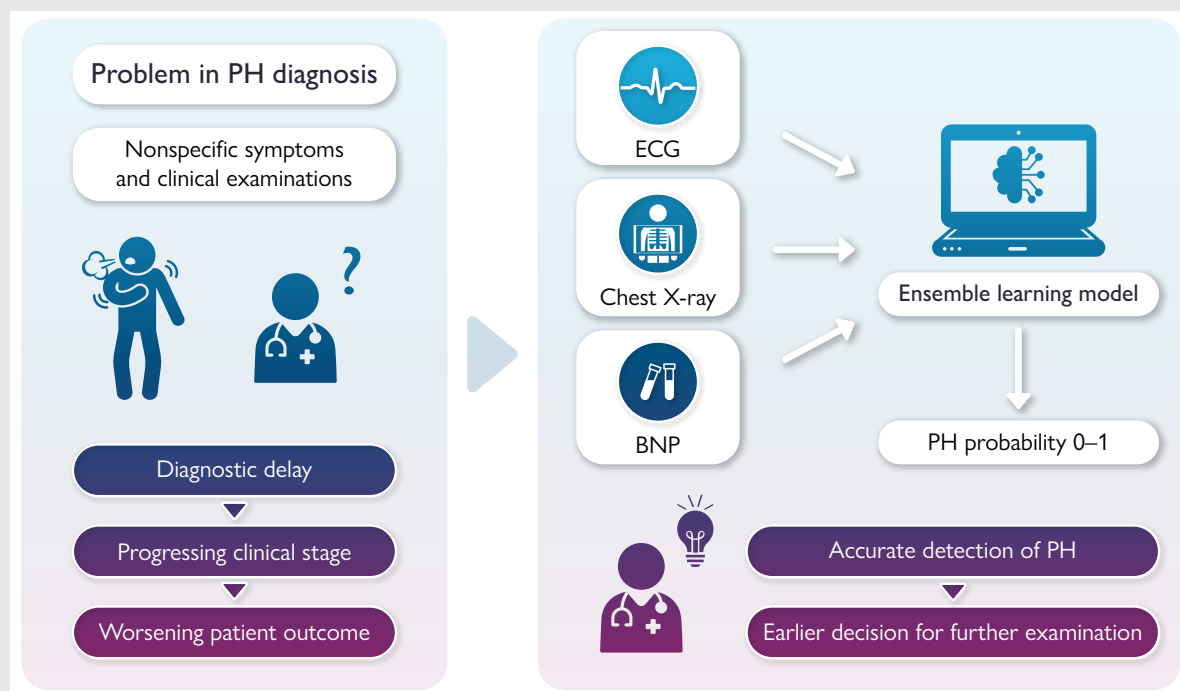
Our ensemble learning model improved doctors' accuracy in detecting PH from ECG, CXR, and BNP examinations. This suggests that earlier and more accurate PH diagnosis is possible, potentially improving patient prognosis.

\* Corresponding author. Tel: +81 3 3815 5411, Email: [koderasatoshi@gmail.com](mailto:koderasatoshi@gmail.com)

© The Author(s) 2025. Published by Oxford University Press on behalf of the European Society of Cardiology.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact [reprints@oup.com](mailto:reprints@oup.com) for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com).

## Graphical Abstract



This figure has been designed using images from Flaticon.com.

## Keywords

Ensemble learning • Multimodality artificial intelligence • Pulmonary hypertension • Diagnosis improvement

## Introduction

Pulmonary hypertension (PH) often begins with nonspecific symptoms such as shortness of breath, oedema, and fatigue.<sup>1</sup> At the 6th World Symposium on Pulmonary Hypertension<sup>2</sup> held in Nice in 2018, PH was defined by a resting mean pulmonary artery pressure of >20 mmHg and a pulmonary vascular resistance of >3 Wood units according to examination by right heart catheterization.<sup>2</sup> In the diagnostic algorithm, patients with unexplained exertional dyspnoea and/or suspected PH first undergo blood tests including brain natriuretic peptide (BNP) measurements and electrocardiography (ECG) to determine whether further testing, including chest X-ray (CXR) and echocardiography, is required. Electrocardiography,<sup>3</sup> CXR,<sup>4,5</sup> BNP measurement,<sup>6</sup> and echocardiography<sup>7</sup> are commonly used for PH and included in diagnostic algorithms.

However, research has pointed out a time lag between the onset of symptoms and PH diagnosis due to their nonspecific nature.<sup>8</sup> The average delay from onset to diagnosis was reported as 47 months, with clinical classification progressing during this period.<sup>9</sup> This delay is divided into patient-related factors (from symptom recognition to seeking medical advice) and physician-related factors (from consultation to diagnosis).<sup>10</sup> Reports indicate that these delays worsen prognosis, highlighting the need for strategies to expedite diagnosis. A method to support earlier diagnosis is required to improve patient outcomes.

In recent years, artificial intelligence (AI) has seen remarkable advancements and can now detect cardiovascular diseases from diagnostic tests, such as ECGs and X-rays. Artificial intelligence models using ECG have demonstrated high predictive accuracy for PH, with area under the curve (AUC) values ranging from 0.87 to 0.90.<sup>11–14</sup> Furthermore, AI models utilizing CXR images have detected PH, with

AUC values ranging from 0.71 to 0.988, surpassing the diagnostic accuracy of experienced physicians.<sup>15,16</sup> Moreover, AI can process multiple data types simultaneously, similar to human doctors. Combining multiple test results, as is usually done in clinical settings, is expected to enhance disease detection accuracy. Multimodality AI models that predict cardiovascular diseases have been reported<sup>17,18</sup>; however, to the best of our knowledge, no multimodality AI models for detecting PH have been reported till date. Using multiple modalities to predict PH is a novel approach that may improve patient outcomes by supporting accurate diagnosis in patients with nonspecific symptoms. Evaluating the clinical usefulness of the developed model requires conducting physician interpretation tests to assess its impact on doctors' decision-making in clinical settings.<sup>19,20</sup> No reports exist of conducting such tests with ensemble learning model for predicting PH, making this study a pioneering effort.

We hypothesized that doctors' accuracy in detecting PH from examination data would improve with an ensemble learning model. In this study, we developed an ensemble learning model that uses three clinical examinations, ECG, CXR, and BNP, to predict the presence or absence of PH and tested whether the AI model supports cardiologists' decisions.

## Methods

### Study sample

For the ECG model, we collected echocardiographic examinations between January 2015 and May 2021 from patients aged 18 years and older at eight institutions [The University of Tokyo Hospital (UTokyo), Asahi General Hospital, Sakakibara Heart Institute, Jichi Medical University Saitama Medical Center, Tokyo Bay Urayasu Ichikawa Medical Center,

Mitsui Memorial Hospital, JR Tokyo General Hospital, and NTT Medical Center Tokyo] as described previously.<sup>21</sup> We excluded data from UTokyo to avoid data leakage. Data were divided into training, validation, and test sets, and we ensured that all examinations from a single patient were allocated to the same set with an allocation ratio of 7:1.5:1.5 based on unique patient IDs to prevent data leakage due to the inclusion of multiple examinations of the same patient in different sets (see [Supplementary material online, Figure S1](#)).

For the CXR, BNP, and ensemble learning model, we created another data set by collecting ECG, CXR, and BNP data. We collected ECG examinations from patients aged 18 and over at The University of Tokyo Hospital Department of Cardiology from 1 January 2015 to 31 December 2018 and gathered data from the most recent chest radiographs and blood tests conducted nearest to the ECG examination date. If the tests were conducted the same number of days before and after the ECG, we selected the test that occurred after the ECG. From the collected data, we excluded patients not evaluated for PH by echocardiography within 1 year from ECG and those without anteroposterior CXR obtained within 1 year after transthoracic echocardiography. Data were divided into training, validation, and test sets for the CXR model, with an allocation ratio of 7:1.5:1.5 based on unique patient IDs to prevent data leakage due to the inclusion of multiple examinations of the same patient in different sets. For the BNP and ensemble learning models, we excluded patients with no plasma BNP data or with plasma BNP data without echocardiographic PH labels within 1 year after BNP measurement, and patients with duplicated BNP data, retaining the oldest data from the training, validation, and test data sets. This data set was used for the BNP and ensemble learning models ([Figure 1](#)).

This study was conducted in accordance with the revised Declaration of Helsinki and approved by the Institutional Review Board of the University of Tokyo [2021132NI-(2)].<sup>18</sup> Informed consent was obtained via the opt-out method from our website.

## Clinical examinations

For ECG examinations, UTokyo and Mitsui Memorial Hospital used equipment from Fukuda Denshi (Tokyo, Japan), while other facilities used equipment from Nihon Kohden (Tokyo, Japan). Electrocardiography data formats were standardized with a sampling rate of 500 Hz and intervals of 10 s. Echocardiographic examinations were conducted by experienced cardiac sonographers and cardiologists, following the guidelines of the American Society of Echocardiography<sup>7</sup> and verified by cardiology specialists. Chest radiography examinations were performed in the radiography room or wards of the University of Tokyo Hospital. Brain natriuretic peptide testing was conducted in the outpatient blood collection room or wards, with measurements taken by the laboratory at the University of Tokyo Hospital.

## Machine learning procedure

The models were developed using Python3.7 on an Nvidia Tesla V-100 32 GB graphics processing unit. For the ECG model, a deep learning model for detecting PH from 123 260 ECG records was developed and employed (see [Supplementary material online, Figure S2](#)), using a convolutional neural network with seven convolutional layers.<sup>21</sup> For the CXR model, a deep learning model for predicting cardiac diseases from CXR images was used, as reported in previous studies.<sup>22,23</sup> The learning parameters were set for a classification task using ResNet18 architecture. For the BNP model, PH presence was predicted from normalized BNP values using logistic regression. Ground-truth labels for each AI model were PH labels obtained by echocardiography.

Data preprocessing information is presented in [Supplementary material online, Data S1](#). Outputs from the ECG, CXR, and BNP models that ranged from 0 to 1 were standardized using StandardScaler before being input into a three-layer fully connected neural network.<sup>24</sup> Because of the minority class of PH, the Synthetic Minority Over-Sampling Technique was used to improve class imbalance by increasing the minority class data volume.<sup>25</sup> We used a sigmoid function as the activation function to produce the model's final output, indicating the presence or absence of PH ([Figure 2](#)). The optimal cut-off value was calculated at the point where the Youden Index, calculated by 'Sensitivity + Specificity – 1,' reached its maximum.<sup>26,27</sup> Further information is presented in the [Supplementary material online, Data S2](#).

## Interpretation test

We randomly selected cases for the interpretation test from each group (with and without PH) from the test data set. In the first block (part A), 30 cases with and without PH were arranged in random order, and doctors determined the presence or absence of PH without AI prediction. In the second block (part B), which consisted of 30 cases with and without PH, doctors determined the presence or absence of PH with AI prediction values and cut-off values at the Youden Index. All data sets (ECG, CXR, BNP, and AI prediction values) in the first and second blocks were different. The test involved 10 cardiologists with at least 4 years of clinical experience, and it was conducted individually. They assessed the presence of PH based on ECG, frontal CXR images, and BNP values displayed on a computer screen and answered with two options: PH or not. In the second block, the AI predictions and Youden Index cut-off values were also displayed ([Figure 3](#)).

## Evaluation criteria

The criterion for PH presence through echocardiography was defined as a right ventricular systolic pressure (RVSP)<sup>28</sup> of 40 mmHg or higher. Right ventricular systolic pressure was calculated by measuring the tricuspid regurgitation velocity (TRV) and applying the simplified Bernoulli equation ( $TRV^2 \times 4$ ) along with the estimated right atrial pressure.<sup>11</sup> To screen PH, the guideline-recommended criterion of  $RVSP > 40$  mmHg<sup>7,29</sup> or  $TRV > 2.8$  m/s<sup>1</sup> has been used to decide whether more invasive testing with right heart catheterization is necessary. Consequently, this study adopted  $RVSP \geq 40$  mmHg as the criterion. The sensitivity for diagnosing PH using systolic pulmonary artery pressure<sup>30</sup> measured with echocardiography was reported to range from 79 to 100%.<sup>31</sup> Additionally, the discrepancy between echocardiography and right heart catheterization was reported to be small.<sup>30</sup>

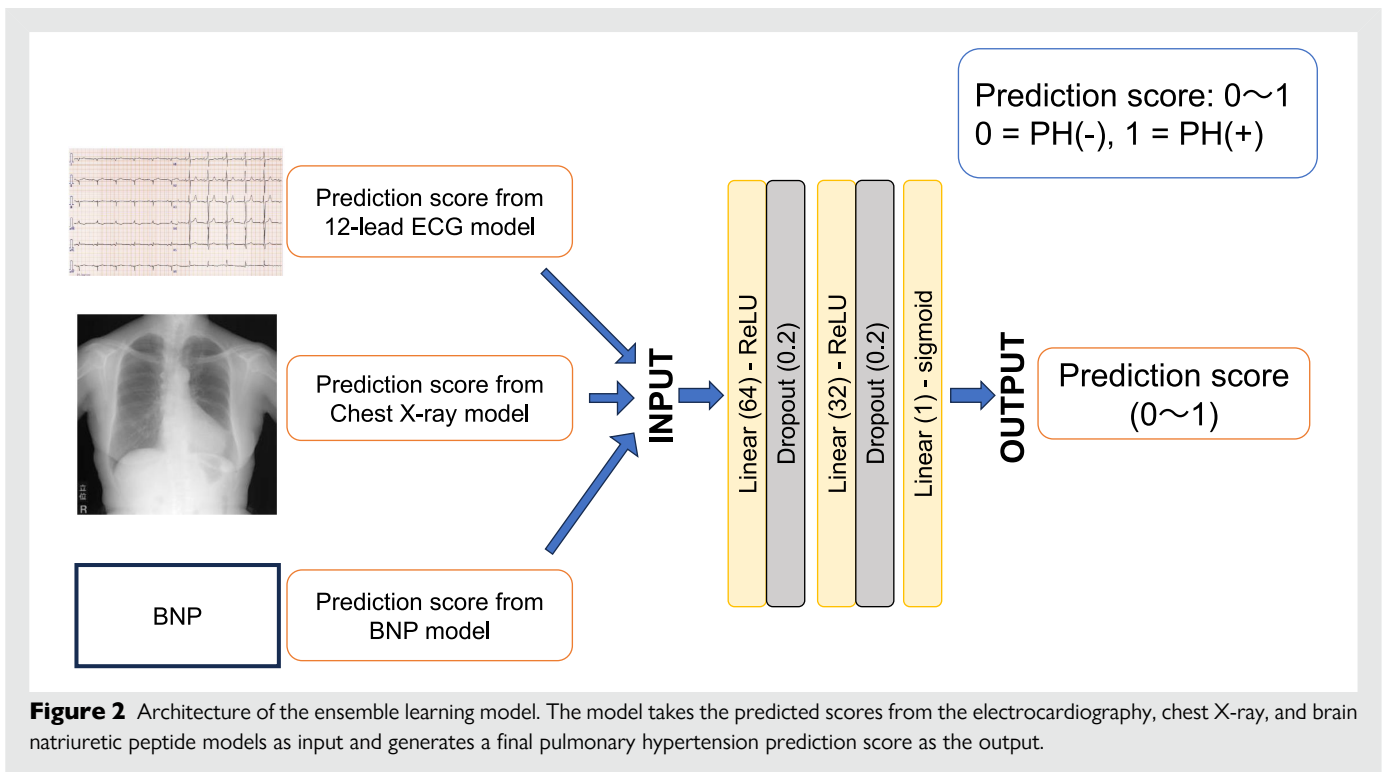
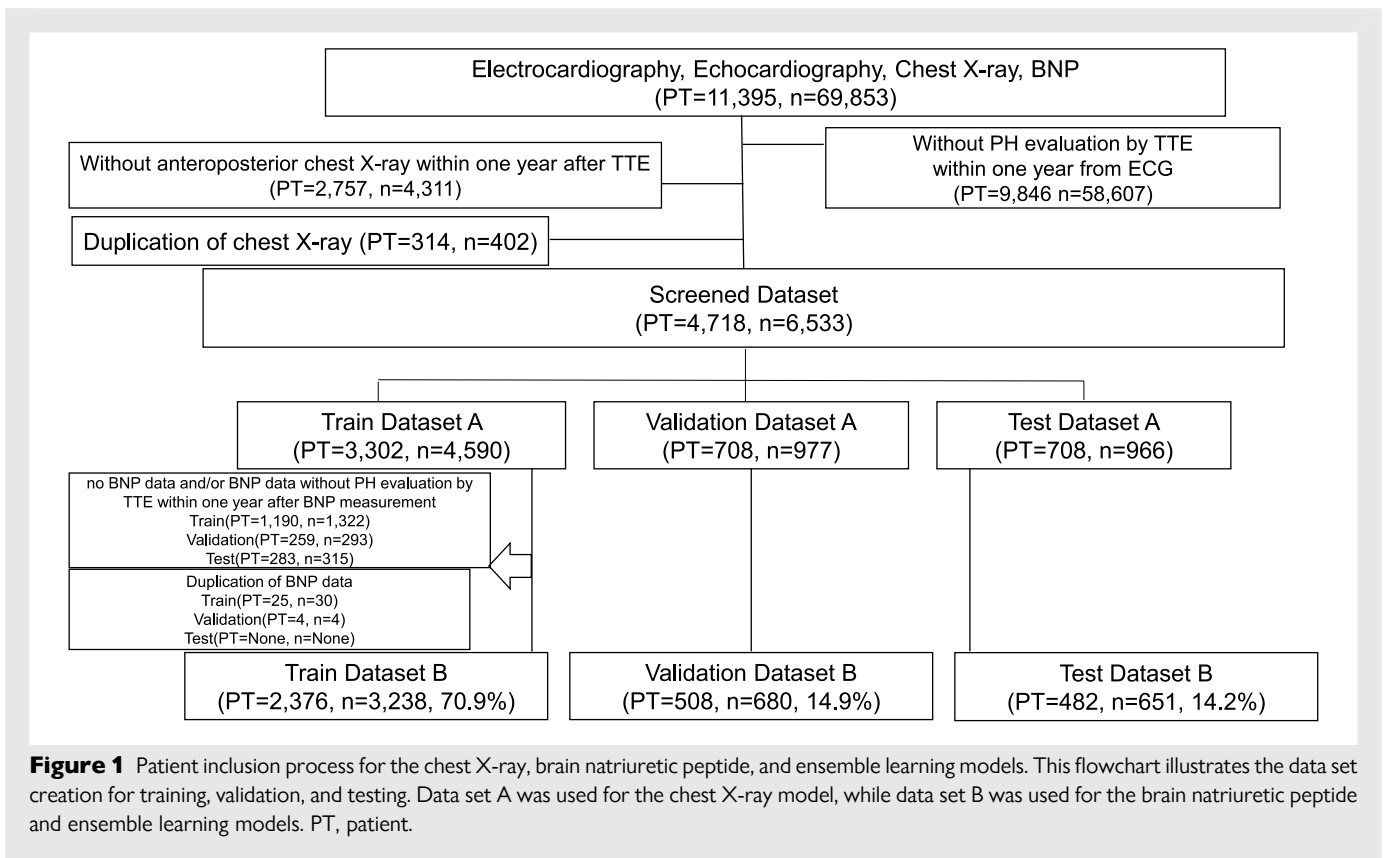
## Statistical analysis

Continuous numerical data were analysed using Welch's analysis of variance (ANOVA), while categorical data were analysed using the  $\chi^2$  test. The accuracy of the models was evaluated using receiver operating characteristic (ROC) curves, and the area under the ROC curve (AUROC) was calculated. The 95% confidence intervals (CIs) were computed using DeLong's method.<sup>32</sup> DeLong's method was used to compare the AUROCs within the same data set.<sup>33</sup> The Z-test was used to assess the difference in difficulty between cases with and without AI support. In the interpretation test, the accuracy rates for the cardiologists were evaluated with and without viewing of the AI model results. Differences in accuracy rates with and without AI support were analysed using Student's t-test. For cases with AI support, we calculated the agreement rate between the cardiologists' responses and the ensemble learning model's classification based on the cut-off value that maximized the Youden Index in the ensemble model's test data set. The predicted values of the ensemble learning model were divided into three groups: 0–0.33, 0.33–0.66, and 0.66–1.00. The ANOVA was performed to assess the presence of significant differences in agreement rates among these groups. Statistical tests were performed using Python, and a  $P < 0.05$  was considered statistically significant.

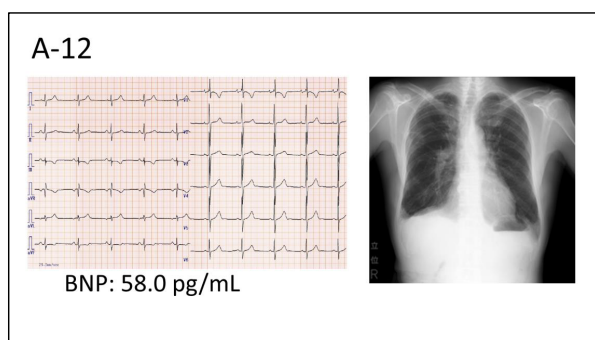
## Results

### Patient characteristics

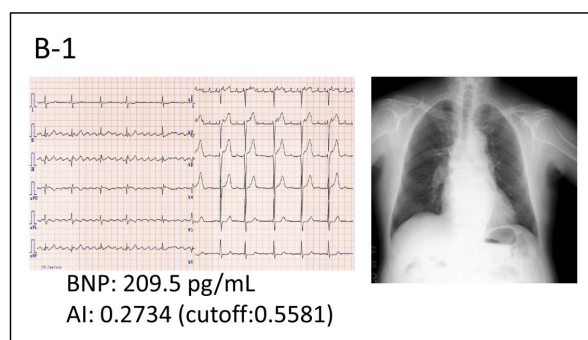
For the ECG model, data from 71 826 patients comprising 123 260 data points were used (see [Supplementary material online, Table S1](#)). For the CXR model, data from 4718 patients comprising 6533 data points were included (see [Supplementary material online, Table S2](#)). The BNP and ensemble learning models used data derived from the same data set as the CXR model. [Table 1](#) details the demographic backgrounds of the training, validation, and testing data sets. Following random allocation to these groups, the average age ranged from 62.9 to 63.4 years. In the data set, 56.1% of participants were male, and 43.9% were female. Some inconsistencies in the totals exist due to the retrospective collection of data, which lacked records of some examination findings.



## Case without AI model prediction



## Case with ensemble learning model prediction



**Figure 3** Screen examples for the interpretation test. Cardiologists were shown electrocardiography, chest X-ray, and brain natriuretic peptide data on a computer screen. The left picture shows a case without artificial intelligence prediction, while the right image shows a case with the ensemble learning model prediction score and its corresponding cut-off value. Each case was presented individually.

## Model performance

The performance of the models, including AUROC, accuracy, sensitivity, and specificity of the ECG, CXR, BNP, and ensemble learning models for the testing data sets, is documented in [Table 2](#), and each ROC curve is presented in [Figure 4](#). For the 12-lead ECG model, the AUROC was 0.818 (95% CI: 0.808–0.828), accuracy was 72.9% (95% CI: 72.3–73.5), sensitivity was 75.6% (95% CI: 75.0–76.1), and specificity was 72.5% (95% CI: 71.9–73.1). For the CXR model, the AUROC was 0.823 (95% CI: 0.780–0.866), accuracy was 85.5% (95% CI: 83.1–87.6), sensitivity was 64.6% (95% CI: 61.6–67.6), and specificity was 89.3% (95% CI: 87.1–91.1). For the BNP model, AUROC was 0.724 (95% CI: 0.668–0.780), accuracy was 70.0% (95% CI: 66.4–73.4), sensitivity was 66.7% (95% CI: 63.0–70.2), and specificity was 70.8% (95% CI: 67.2–74.1). For the ensemble learning model, the AUROC was 0.872 (95% CI: 0.829–0.915), accuracy was 83.7% (95% CI: 80.7–86.4), sensitivity was 74.6% (95% CI: 71.1–77.8), and specificity was 85.7% (95% CI: 82.8–88.1). In order to directly compare each model's performance, we calculated AUROCs and performed the DeLong test using each model's prediction values for the test data set of the ensemble learning model. The AUROCs and *P*-values between each model are presented in [Supplementary material online, Figure S3](#) and [Table S3](#).

## Physician interpretation test

No significant difference was observed in the percentage of correct answers to the AI model between the two question sets (part A and part B) ( $P = 0.67$ ). The results of the cardiologists' interpretation tests are presented in [Table 3](#). The accuracy rates  $\pm$  standard deviation of cardiologists without AI predictions were  $65.0\% \pm 4.7\%$ , and with AI support, it was  $74.0\% \pm 2.7\%$  ( $P < 0.01$ ). Additionally, sensitivity was  $53.7\% \pm 17.1\%$  without AI and  $66.0\% \pm 12.4\%$  with AI support ( $P < 0.01$ ), while specificity was  $76.3\% \pm 14.8\%$  without AI and  $82.0\% \pm 9.1\%$  with AI support ( $P = 0.04$ ). The agreement rates between the cardiologists' responses and the ensemble learning model's classification for cases with AI support were plotted for each case (see [Supplementary material online, Figure S4](#)). The ANOVA performed for the agreement rates among the three groups, divided by the ensemble learning model's prediction values at 0.33 and 0.66, showed a significant difference ( $F = 4.900$ ,  $P = 0.011$ ). Classification outcomes in the interpretation test 'with AI part' cases are presented in [Supplementary material online, Table S4](#).

## Discussion

This multimodal model effectively predicts the possibility of PH, representing a novel approach as no existing models use multiple clinical examinations for this purpose. We demonstrated that our model improves cardiologists' diagnostic accuracy in detecting PH.

The AUROC of the ensemble learning model was 0.872, comparable to those of previously reported models predicting PH from 12-lead ECG (0.87–0.90)<sup>11–14</sup> and CXR (0.71 and 0.988).<sup>15,16</sup> In our study, the ensemble learning model achieved a higher AUROC than did the ECG, CXR, and BNP models. This improvement may be attributed to the increased number of modalities used.<sup>34</sup> However, because of differences in the data sets used for each model, precise accuracy comparisons among these four models were not possible. According to the comparison results in the test data set for ensemble model (see [Supplementary material online, Figure S3](#) and [Table S3](#)), our ensemble learning model outperformed the other models, followed by the CXR, ECG, and BNP models. Previous studies did not compare the detection accuracy of doctors using the AI model with those not using them, leaving it unclear whether these models improved doctors' detection accuracy. Our ensemble learning model increased cardiologists' detection accuracy from 65.0 to 74.0%. Additionally, cardiologists with lower accuracy without the model experienced a greater increase in accuracy when using this model.

Currently, cardiologists evaluate ECGs, CXR, and BNP values without AI support to determine whether patients should undergo further examination. Using our model enables cardiologists to detect PH more accurately, aiding in deciding whether to proceed with detailed examinations. In the interpretation test, the agreement between the cardiologists' responses and the ensemble learning model's classification results was higher when the ensemble learning model's prediction value was closer to 0 or 1. The cardiologists were shown the prediction value and the cut-off value of 0.5581, and it is possible that they relied on the model's classification when the ensemble learning model strongly indicated the presence or absence of PH. This suggests that the cardiologists may have trusted the model more when the prediction values clearly indicated a positive or negative classification, potentially using the cut-off value as a guide.

Although it was cardiologists that conducted interpretation tests in this study, primary care physicians, likely non-cardiologists, may also benefit from our model. Our model encourages primary care physicians to make rapid consultations with cardiologists by improving their detection accuracy, potentially reducing diagnostic delays. Earlier



**Table 1 Patient characteristics for the brain natriuretic peptide and ensemble learning models**

	Train	Validation	Test	P-value
Number of studies	3238	680	651	
Number of patients	2376	508	482	
Age, years	62.9 ± 17.8	62.9 ± 18.9	63.4 ± 17.6	0.80
Sex				0.18
Male, <i>n</i>	1831 (56.5)	359 (52.8)	371 (57.0)	
Body height, cm	162.3 ± 29.3	161.4 ± 10.0	161.7 ± 10.0	0.65
Body weight, kg	59.8 ± 13.3	58.7 ± 12.5	59.7 ± 13.0	0.15
Echocardiographic findings				
LVEF, %	59.2 ± 15.9	60.3 ± 15.2	58.7 ± 15.6	0.13
LA diameter, mm	40.8 ± 22.4	39.1 ± 12.2	40.9 ± 20.4	0.16
LAVI, mL/m <sup>2</sup>	43.4 ± 29.5	41.5 ± 28.8	44.4 ± 31.9	0.22
TR				0.83
TR = 0	2968 (91.7)	618 (90.9)	590 (90.6)	
TR = 1	195 (6.0)	42 (6.2)	43 (6.6)	
RVSP, mmHg	31.9 ± 15.6	31.5 ± 14.0	30.9 ± 15.1	0.28
PH				0.34
PH = 0	2597 (80.2)	541 (79.6)	537 (82.5)	
PH = 1	641 (19.8)	139 (20.4)	114 (17.5)	
ECG findings				
HR, b.p.m.	72.1 ± 14.8	72.0 ± 15.7	71.8 ± 15.5	0.93
PR interval, ms	175.9 ± 50.1	178.4 ± 49.3	175.8 ± 50.3	0.48
QRS interval, ms	110.8 ± 25.2	109.8 ± 24.6	109.9 ± 25.1	0.52
QT interval, ms	405.8 ± 40.1	405.5 ± 40.9	408.3 ± 41.7	0.33
QTc, ms	439.3 ± 36.9	437.9 ± 34.5	440.7 ± 36.2	0.38
QRS axis	28.9 ± 48.1	31.6 ± 47.1	26.4 ± 47.5	0.14
Laboratory data				
BNP, pg/mL	181.7 ± 361.1	177.4 ± 358.0	161.9 ± 270.6	0.42

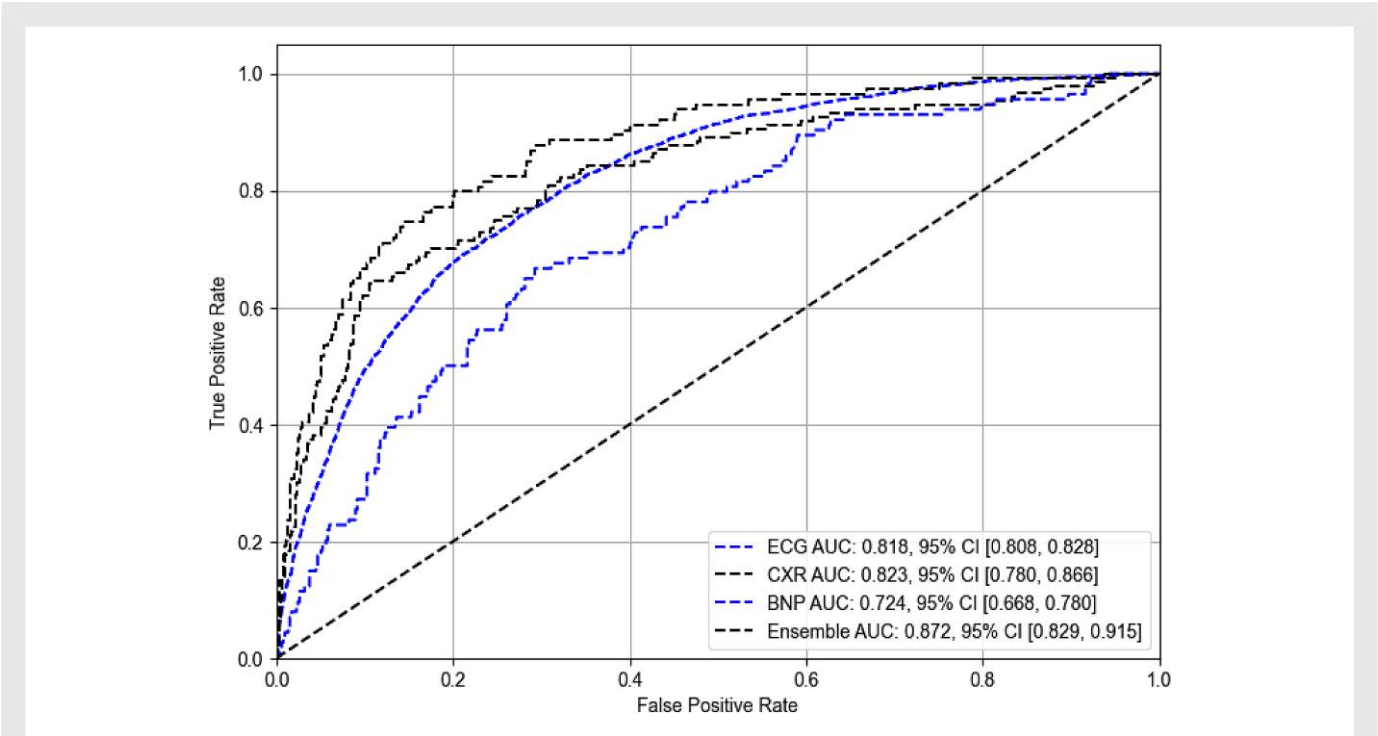
This patient cohort corresponds to data set B in [Figure 1](#). Data are presented as mean with standard deviation or *n* (%). *P*-values indicate differences between the training, validation, and test data sets, calculated by Welch's analysis of variance for continuous numerical data or  $\chi^2$  test for categorical data. Tricuspid regurgitation (TR) was defined as moderate or higher (TR = 1). LVEF, left ventricular ejection fraction; LA, left atrium; LAVI, left atrium volume index; HR, heart rate; RVSP, right ventricular systolic pressure; PH, pulmonary hypertension; HR, heart rate, PR interval, time from the onset of the P-wave to the start of the QRS complex; QT interval, time from the start of the Q-wave to the end of the T-wave; QTc, corrected QT interval; QRS interval, time taken for the ventricle to depolarize; BNP, brain natriuretic peptide.

**Table 2 Model performance**

Modality	AUROC	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
ECG	0.818 (0.808–0.828)	72.9 (72.3–73.5)	75.6 (75.0–76.1)	72.5 (71.9–73.1)	28.9 (28.3–29.5)	95.3 (95.0–95.5)
CXR	0.823 (0.780–0.866)	85.5 (83.1–87.6)	64.6 (61.6–67.6)	89.3 (87.1–91.1)	51.9 (48.8–55.0)	93.4 (91.6–94.8)
BNP	0.724 (0.668–0.780)	70.0 (66.4–73.4)	66.7 (63.0–70.2)	70.8 (67.2–74.1)	32.6 (29.1–36.3)	90.9 (88.5–92.9)
Ensemble learning	0.872 (0.829–0.915)	83.7 (80.7–86.4)	74.6 (71.1–77.8)	85.7 (82.8–88.1)	52.5 (48.6–56.3)	94.1 (92.0–95.6)

AUROC data are shown with point estimates and 95% confidence intervals in parenthesis. Accuracy, sensitivity, specificity, PPV, and NPV are shown as percentage with 95% confidence interval in parenthesis.

PPV, positive predictive value; NPV, negative predictive value.



**Figure 4** Area under the receiver operating characteristic curves for each model on the test data set. The receiver operating characteristic curves represent the performance of the individual models: electrocardiography, chest X-ray, brain natriuretic peptide, and the ensemble learning model on each test data set. The area under the receiver operating characteristic curve with 95% confidence interval is shown for each model. The diagonal dashed line represents the line of no discrimination (area under the receiver operating characteristic curve = 0.5), indicating random classification performance.

**Table 3** Physician interpretation of test results

		Accuracy (%)	Sensitivity (%)	Specificity (%)
Part A	Cardiologists	65.0 ± 4.7	53.7 ± 17.1	76.3 ± 14.8
Part B	Cardiologists with AI support	74.0 ± 2.7	66.0 ± 12.4	82.0 ± 9.1
	P-value	<0.01	<0.01	0.04

Accuracy, sensitivity, and specificity of cardiologists are shown with standard deviation. P-values indicate differences between part A without AI support and part B with AI support, derived from Student's t-test.

diagnosis can improve patient outcomes. In a previous study, 40.9% patients with PH were misdiagnosed before receiving a correct diagnosis, causing diagnostic delays with an average interval of 7.7 months (0.2–155.9 months) between the first consultation and PH diagnosis.<sup>10</sup> Here, our ensemble learning model can assist doctors in detecting PH from basic clinical examinations, potentially leading to more rapid consultations, further examination, and diagnoses. Additionally, our ensemble learning model enhances both the sensitivity and specificity of PH diagnosis, reducing false negatives without additional clinical tests and thereby offering cost benefits.

No specific clinical tests are available for PH diagnosis, except for right heart catheterization, which is invasive and inappropriate for screening.<sup>1</sup> The modalities used in this study (ECGs, CXR, and BNP), though having low specificity, are advantageous because they are simple, less invasive, cost-effective, and do not require special skills, such as those required for echocardiography. It was reported that the

increased number of modalities does not necessarily enhance prediction accuracy.<sup>35</sup> In the current medical AI field, it is preferable to select modalities based on background knowledge, such as disease characteristics and clinical examinations. We utilized three clinical examinations (ECG, CXR, and BNP) in this study, considering the diagnostic algorithm in current guidelines.<sup>1</sup> To improve the prognosis of patients with PH, a method to support doctors' diagnoses should be established. We demonstrated that our ensemble learning model supports cardiologists in detecting PH more accurately and helps physicians consult with cardiologists earlier. This allows patients to receive appropriate treatment earlier, improving their prognosis.

**Limitations**

In this study, we used echocardiographic measurements as criteria for PH. Although the standard criterion for diagnosing PH is right heart

catheterization, it poses potential risks, making it challenging to perform on individuals with low prior probabilities of having PH. Therefore, as observed in previous studies,<sup>11–14</sup> we utilized echocardiography.<sup>11–14</sup> Some patients with PH may have trivial tricuspid regurgitation, resulting in false negative results in screenings using RVSP as a criterion measured by echocardiography.<sup>36</sup> However, if we had developed a model based on a diagnosis of PH using right heart catheterization data, the model would have been limited to a population deemed appropriate for right heart catheterization, typically including patients with significant heart disease or highly suspected PH. By labelling PH using echocardiography, as in our study, we included a broader range of patients. Patient data in this study were collected from tertiary hospitals in Japan; therefore, they did not consist entirely of asymptomatic patients. Therefore, our model cannot be directly applied to PH screening in asymptomatic individuals. External validation in asymptomatic individuals and additional tuning are required. The model was developed, validated, and assessed using data obtained only from Japanese facilities. The accuracy of AI diagnostics for X-rays can vary according to patient characteristics,<sup>37</sup> suggesting that additional training might be required to adapt the AI model for use in non-Japanese populations. Our ensemble learning model did not demonstrate significant superiority in classification performance compared to previous single-modality models. We hypothesize several reasons for this. First, we employed a lower threshold for RVSP, which made our data set more challenging, aiming to detect patients in earlier stages or those who may not yet exhibit significant changes on diagnostic tests. Second, the baseline models used for ECG, CXR, and BNP in our ensemble learning model might influence the overall performance, and alternative models could potentially yield better results. Additionally, in this research, we focused on ECG, CXR, and BNP, known to have strong associations with PH; future research could explore other modalities that might improve detection accuracy. In the interpretation test, only ECG, CXR, BNP, and prediction values were presented to the cardiologists. While this could be a limitation because it differs from the scenario in actual clinical practice, where patient symptoms and other background information are considered, it has been noted that even with such background information, accurate PH diagnosis remains a challenge. Therefore, we believe that the additional information provided by our model would be beneficial for doctors.

In conclusion, we developed an ensemble learning model that detects the presence of PH from ECG, CXR images, and BNP values. This model can enhance the accuracy of physicians in predicting PH, potentially improving patient outcomes by contributing to the early diagnosis of PH.

## Lead author biography



Dr Risa Kishikawa is a cardiologist at the University of Tokyo Hospital in Tokyo, Japan. She graduated from the Faculty of Medicine at the University of Tokyo in 2015. After completing her clinical training and cardiology residency at Showa General Hospital in Tokyo, she began her research in medical AI at the Graduate School of Medicine of the University of Tokyo, obtaining her PhD in Medicine in 2024. She conducts research with the goal of using AI to diagnose cardiovascular diseases and to aid in preventive care.

## Supplementary material

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

## Acknowledgements

We would like to thank Editage ([www.editage.jp](http://www.editage.jp)) for English language editing. We also acknowledge the use of Grammarly and ChatGPT for their assistance in proofreading and refining the manuscript.

## Funding

This study was supported by the Cross-ministerial Strategic Innovation Promotion Program (SIP) on “Integrated Health Care System” (Grant Number JPJ012425) and the Japan Agency for Medical Research and Development (Grant Number JP23hk0102078h0003).

**Conflict of interest:** R.K. held stock in NVIDIA. The other authors declare no conflict of interest for this contribution.

## Data availability

The data underlying this article cannot be shared publicly because, according to the informed consent obtained through an opt-out form on our website, participants were informed that their data, including those that were not identifiable, would not be disclosed to other researchers.

## References

- Humbert M, Kovacs G, Hoeper MM, Badagliacca R, Berger RMF, Brida M, et al. 2022 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension. *Eur Heart J* 2022;**43**:3618–3731.
- Simonneau G, Montani D, Celermajer DS, Denton CP, Gatzoulis MA, Krowka M, et al. Haemodynamic definitions and updated clinical classification of pulmonary hypertension. *Eur Respir J* 2019;**53**:1801913.
- Ley L, Hölzgen R, Bogossian H, Ghofrani HA, Bandorski D. Electrocardiogram in patients with pulmonary hypertension. *J Electrocardiol* 2023;**79**:24–29.
- Crisan S, Baghina RM, Luca SA, Cozlac AR, Negru AG, Vacarescu C, et al. Comprehensive imaging in patients with suspected pulmonary arterial hypertension. *Heart* 2024;**110**:228–234.
- Miniati M, Monti S, Airò E, Pancani R, Formichi B, Bauleo C, et al. Accuracy of chest radiography in predicting pulmonary hypertension: a case-control study. *Thromb Res* 2014;**133**:345–351.
- Leuchte HH, Holzappel M, Baumgartner RA, Ding I, Neurohr C, Vogeser M, et al. Clinical significance of brain natriuretic peptide in primary pulmonary hypertension. *J Am Coll Cardiol* 2004;**43**:764–770.
- Rudski LG, Lai WW, Afilalo J, Hua L, Handschumacher MD, Chandrasekaran K, et al. Guidelines for the echocardiographic assessment of the right heart in adults: a report from the American Society of Echocardiography endorsed by the European Association of Echocardiography, a registered branch of the European Society of Cardiology, and the Canadian Society of Echocardiography. *J Am Soc Echocardiogr* 2010;**23**:685–713. quiz 786–8.
- Torbicki A, Kurzyna M. The diagnostic approach to pulmonary hypertension. *Semin Respir Crit Care Med* 2023;**44**:728–737.
- Strange G, Gabbay E, Kermeen F, Williams T, Carrington M, Stewart S, et al. Time from symptoms to definitive diagnosis of idiopathic pulmonary arterial hypertension: the delay study. *Pulm Circ* 2013;**3**:89–94.
- Small M, Perchenet L, Bennett A, Linder J. The diagnostic journey of pulmonary arterial hypertension patients: results from a multinational real-world survey. *Ther Adv Respir Dis* 2024;**18**:17534666231218886.
- Kwon JM, Kim KH, Medina-Inojosa J, Jeon KH, Park J, Oh BH. Artificial intelligence for early prediction of pulmonary hypertension using electrocardiography. *J Heart Lung Transplant* 2020;**39**:805–814.
- Aras MA, Abreau S, Mills H, Radhakrishnan L, Klein L, Mantri N, et al. Electrocardiogram detection of pulmonary hypertension using deep learning. *J Card Fail* 2023;**29**:1017–1028.
- Liu CM, Shih ESC, Chen JY, Huang CH, Wu IC, Chen PF, et al. Artificial intelligence-enabled electrocardiogram improves the diagnosis and prediction of mortality in patients with pulmonary hypertension. *JACC Asia* 2022;**2**:258–270.
- Leha A, Hellenkamp K, Unsöld B, Mushemi-Blake S, Shah AM, Hasenfuß G, et al. A machine learning approach for the prediction of pulmonary hypertension. *PLoS One* 2019;**14**:e0224453.
- Imai S, Sakao S, Nagata J, Naito A, Sekine A, Sugiura T, et al. Artificial intelligence-based model for predicting pulmonary arterial hypertension on chest x-ray images. *BMC Pulm Med* 2024;**24**:101.



16. Kusunose K, Hirata Y, Tsuji T, Kotoku J, Sata M. Deep learning to predict elevated pulmonary artery pressure in patients with suspected pulmonary hypertension using standard chest X ray. *Sci Rep* 2020;**10**:19311.
17. Amal S, Safarnejad L, Omiye JA, Ghanzouri I, Cabot JH, Ross EG. Use of multi-modal data and machine learning to improve cardiovascular disease care. *Front Cardiovasc Med* 2022;**9**:840262.
18. Soto JT, Weston Hughes J, Sanchez PA, Perez M, Ouyang D, Ashley EA. Multimodal deep learning enhances diagnostic precision in left ventricular hypertrophy. *Eur Heart J Digit Health* 2022;**3**:380–389.
19. Hassan N, Slight R, Morgan G, Bates DW, Gallier S, Sapey E, et al. Road map for clinicians to develop and evaluate AI predictive models to inform clinical decision-making. *BMJ Health Care Inform* 2023;**30**:e100784.
20. de Hond AAH, Leeuwenberg AM, Hooft L, Kant IMJ, Nijman SWJ, van Os HJA, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *NPJ Digit Med* 2022;**5**:2.
21. Sato M, Kodera S, Setoguchi N, Tanabe K, Kushida S, Kanda J, et al. Deep learning models for predicting left heart abnormalities from single-lead electrocardiogram for the development of wearable devices. *Circ J* 2023;**88**:146–156.
22. Ueda D, Matsumoto T, Ehara S, Yamamoto A, Walston SL, Ito A, et al. Artificial intelligence-based model to classify cardiac functions from chest radiographs: a multi-institutional, retrospective model development and validation study. *Lancet Digit Health* 2023;**5**:e525–e533.
23. Matsumoto T, Walston SL, Miki Y, Ueda D. Nervus: a comprehensive deep learning classification, regression, and prognostication tool for both medical image and clinical data analysis 2022. *arXiv e-prints*. December 12, 2022:arXiv:2212.11113v1. <https://doi.org/10.48550/arXiv.2212.11113>.
24. Manasrah A, Alkayem A, Qasaimeh M, Nofal S. Assessment of machine learning security: the case of healthcare data. In: International Conference on Data Science, E-Learning and Information Systems 2021, p. 91–98. Association for Computing Machinery, Ma'an, Jordan. 2021.
25. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *JAIR* 2002;**16**:321–357.
26. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950;**3**:32–35.
27. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. *Biom J* 2005;**47**:458–472.
28. Azpiri-Lopez JR, Galarza-Delgado DA, Colunga-Pedraza IJ, Arvizu-Rivera RI, Cardenas-de la Garza JA, Vera-Pineda R, et al. Echocardiographic evaluation of pulmonary hypertension, right ventricular function, and right ventricular-pulmonary arterial coupling in patients with rheumatoid arthritis. *Clin Rheumatol* 2021;**40**:2651–2656.
29. McLaughlin VV, Archer SL, Badesch DB, Barst RJ, Farber HW, Lindner JR, et al. ACCF/AHA 2009 expert consensus document on pulmonary hypertension a report of the American College of Cardiology Foundation Task Force on Expert Consensus Documents and the American Heart Association developed in collaboration with the American College of Chest Physicians; American Thoracic Society, Inc.; and the Pulmonary Hypertension Association. *J Am Coll Cardiol* 2009;**53**:1573–1619.
30. Greiner S, Jud A, Aurich M, Hess A, Hilbel T, Hardt S, et al. Reliability of noninvasive assessment of systolic pulmonary artery pressure by Doppler echocardiography compared to right heart catheterization: analysis in a large patient population. *J Am Heart Assoc* 2014;**3**:e001103.
31. Bossone E, D'Andrea A, D'Alto M, Citro R, Argiento P, Ferrara F, et al. Echocardiography in pulmonary arterial hypertension: from diagnosis to prognosis. *J Am Soc Echocardiogr* 2013;**26**:1–14.
32. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;**44**:837–845.
33. Sun X, Xu VV. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Process Lett* 2014;**21**:1389–1393.
34. Kline A, Wang H, Li Y, Dennis S, Hutch M, Xu Z, et al. Multimodal machine learning in precision health: a scoping review. *NPJ Digit Med* 2022;**5**:171.
35. Remeseiro B, Bolon-Canedo V. A review of feature selection methods in medical applications. *Comput Biol Med* 2019;**112**:103375.
36. O'Leary JM, Assad TR, Xu M, Farber-Eger E, Wells QS, Hemnes AR, et al. Lack of a tricuspid regurgitation Doppler signal and pulmonary hypertension by invasive measurement. *J Am Heart Assoc* 2018;**7**:e009362.
37. Castro DC, Walker I, Glocker B. Causality matters in medical imaging. *Nat Commun* 2020;**11**:3673.