Does artificial intelligence dream of elevated natriuretic peptide in chest radiography?

Brief title: AI, BNP, and chest X-ray

Eisuke Kagawa, MD, PhD,[a] Masaya Kato, MD, PhD,[a] Noboru Oda, MD, PhD,[a] Eiji Kunita, MD, PhD,[a] Michiaki Nagai, MD, PhD,[a] Aya Yamane, MD,[a] Shogo Matsui, MD, PhD,[a] Yuki Yoshitomi, MD,[a] Hiroto Shimajiri, MD,[a] Tatsuya Hirokawa, MD,[a] Shunsuke Ishida, MD,[a] Genki Kurimoto, MD,[a] Keigo Dote, MD, PhD[a,b]

[a] Department of Cardiology, Hiroshima City Asa Hospital, Hiroshima, Japan

[b] Department of Cardiology, Hiroshima City North Medical Center Asa Citizens Hospital, Hiroshima, Japan

Total word count: 4758 words

Funding: There is no financial support for this study.

Disclosures: There are no conflicts of interest.

Address for correspondence:

Eisuke Kagawa, MD, PhD

Department of Cardiology, Hiroshima City Asa Hospital

1-2-1, Kameyamaminami, Asakita-ku, Hiroshima, 7310293 Japan

Phone: (81) 82-815-5211

Fax: (81) 82-814-1791

E-mail: ekagawa007@gmail.com

Tweet: We developed the AI model to predict elevated BNP levels from chest X-rays. The model outperformed cardiologists in predicting elevated BNP levels, and the AI assistance improved the diagnostic performance of both experts and non-experts. #HeartFailure, #CHF, #cvImaging

Acknowledgements:

We appreciate all the individuals who participated in this study and underwent testing. They might have bad feelings about being tested.

ABSTRACT

BACKGROUND: Heart failure pandemic is a medical issue, and its diagnosis can be challenging.

OBJECTIVES: To assess the hypothesis that the artificial intelligence (AI) model for predicting elevated brain natriuretic peptide (BNP) levels from chest X-ray images has sufficient performance and could improve the diagnostic performance of human.

METHODS: Patients who underwent chest radiography and BNP testing on the same day were enrolled in this study. The chest X-ray images were labeled based on plasma BNP levels >= 200 pg/mL. We trained models that predicted the label, and the final ensemble model was constructed. Humans were trained and tested to predict labels, followed by the same test, referring to the prediction of the AI model. Diagnostic performance was assessed.

RESULTS: Among the 1607 patients, 10103 chest X-ray images and BNP values were collected. We assembled 31 models as the final model with an accuracy of 0.855, precision of 0.873, recall of 0.827, receiver-operating-characteristics area-under-curve (AUC) of 0.929, and precision-recall AUC of 0.934. The accuracy of the testing by 35 individuals significantly improved from 0.708±0.049 to 0.829±0.069 ($P < 0.001$) with AI assistance. The accuracy of the experts was higher than that of non-experts (0.692±0.042 vs. 0.728±0.051, $P = 0.030$); however, with the AI assistance, the accuracy of the non-experts was rather higher than that of the experts (0.851±0.074 vs. 0.803±0.054, $P = 0.033$).

CONCLUSIONS: The AI model can predict elevated BNP levels from chest X-ray images and could improve the diagnostic performance of humans.

Condensed Abstract: We developed the AI model to predict elevated natriuretic peptide levels from chest radiography and evaluated its performance. We tested the performance of humans and examined whether their performance could be improved by referring the AI predictions. The AI model achieved an accuracy of 0.855 and a receiver-operating-characteristics area-under-curve of 0.929. The accuracy of human testing was significantly improved from 0.708±0.049 to 0.829±0.069 ($P < 0.001$) with AI assistance. The accuracy of the experts was higher than that of non-experts; however, with the AI assistance, the accuracy of the non-experts was higher than that of the experts.

Keywords: deep learning, neural network, machine learning, chest radiography, brain natriuretic peptide, heart failure

Abbreviations:

AI	Artificial intelligence
AUC	Area-under curve
BNP	Brain natriuretic peptide
GRAD-CAM	Gradient-class activation maps
MAE	Mean absolute errors
PR	Precision-recall
ROC	Receiver-operating-characteristics
SD	Standard deviation

Heart failure is one of the major causes of visits to medical facilities in both unplanned emergency situations and routine medical checkups. It is also a growing and called heart failure pandemic because of the aging population (1). However, diagnosis of heart failure, which should be managed medically can be challenging due to the variability of symptoms, especially in heart failure which is complicated by other diseases for those visiting medical facilities or attending physicians who are not familiar with cardiovascular disease (2). Several tests are used to evaluate heart failure (1). Chest X-rays are widely available, and the images can be obtained quickly; however, the evaluation of chest X-ray images requires experience and has limited sensitivity and specificity for heart failure (3). Natriuretic peptide is useful for evaluating heart failure not only in emergency situations for diagnosis, but also be for heart failure management (4-6). However, natriuretic peptide testing requires equipment, and even if the facility has such equipment, it may not be available at all hours, as many facilities do not offer testing at night or on weekends. Furthermore, it is essential to make the decision to perform the test itself. We hope that automated support tools will be developed to assist in our daily practice of heart failure quickly and inexpensively. The rise of artificial intelligence (AI) and the evolution of computer hardware provide novel findings and solutions (7,8). With regard to image recognition, deep neural networks (deep learning) provide relatively good performance compared to those of previous architectures, and some of them can be utilized in the real clinical world (9-11). The aim of our study was to diagnose heart failure by chest X-ray imaging using an AI model and to support clinical practice. We assessed the hypothesis that AI models that predict elevated brain natriuretic peptide (BNP) levels from chest X-ray images could provide excellent performance compared to experienced cardiologists and could improve the diagnostic performance of

humans.

Methods

Study Patients and Datasets

Patients who underwent chest radiography and BNP-testing on the same day at Hiroshima City Asa Hospital and Hiroshima City North Medical Center Asa Citizens Hospital from October 2021 to September 2022 were eligible for this study. Since BNP-testing is the first-choice for natriuretic-peptide testing in these hospitals but not N-terminal pro-brain natriuretic peptide, we selected BNP for this study. We reviewed the medical records of eligible patients, including periods other than the above, and when chest X-ray and BNP-testing were performed on the same day, the chest X-ray images, and plasma BNP values were collected. To increase the robustness and generalizability of the AI models, all conditions of the frontal view of chest X-ray images were collected, including anterior-posterior, posterior-anterior, standing, sitting, and supine positions, with or without inspiration, and any diseases or conditions. Lateral chest radiographs were not included in this study. Patients taking an angiotensin receptor neprilysin inhibitor were included because we intended to make models robust and widely available in real clinical settings, and it might not increase BNP (12). According to the statement of the Japanese Heart Failure Society, we used a BNP cut-off value of 200 pg/mL for the main study and 100 pg/mL for the sub study (13). The chest X-ray images were labeled binary according to the cut-off value. The patients from Hiroshima City Asa Hospital were used for the training and validation datasets, while the patients from Hiroshima City North Medical Center Asa Citizens Hospital were used for the test dataset. The study patients from Hiroshima City Asa Hospital

were randomly divided into two datasets for training and validation. The patients were assigned to datasets in the training: validation: testing dataset ratio of approximately 0.66:0.17:0.17. Many patients had multiple pairs of chest X-ray images and BNP labels, and each patient's data were assigned to only one dataset to avoid overfitting. The patients who were duplicated in both hospitals were assigned to only one dataset. The study was approved by the local institutional review board.

Image Preprocessing

Non-squared chest X-ray images were transformed to square using 0 padding (black field), not cropping. We planned to pre-process the images in which the head was in the upper half of the image and the spine was at 60 ° - 120 ° to the horizontal line of images; however, all the images met these requirements, and these image pre-processing was not utilized.

Outline of an AI Model

We designed AI models to predict higher BNP levels as follows: First, as weak learners, we fine-tuned 31 modified pre-trained image recognition models. The bias of the models was checked using a featured map of gradient-class activation maps (GRAD-CAM) or attention maps. If the model strongly focused on irrelevant objects for the classification, we discarded it and retrained a new model until the model no longer focused on them. Second, we created a soft ensemble prediction model using the weak learners. Details of training for models are provided in the Supplemental Methods.

Evaluation of Models

After the 31 models (weak learners) were obtained, we constructed the final soft ensemble model with the averaged probabilities of the 31 models >= 0.5, which were considered BNP >= cut-off value. The accuracy, precision, recall (sensitivity), specificity, F1 score, receiver-operating-characteristics (ROC) curves, and precision-recall (PR) curves were calculated using the test dataset. ROC and PR curves were constructed using probability of BNP >= cut-off value, and the area under the curve (AUC) was calculated. The results of the performance tests using the test dataset were not used to train or select the models back in time to avoid overfitting the test dataset.

Human Testing

We evaluated human performance to predict elevated BNP levels from chest X-ray images. The general findings of heart failure on chest X-ray images, as well as those of BNP, were taught to those who were tested. The examinee was shown chest X-ray images and their BNP label in the training dataset as human training and then evaluated 100 chest X-ray images from the test dataset and answered their binary prediction. The 100 images for human testing consisted of 50 images with BNP < cut-off value and 50 with BNP >= cut-off value and were randomly sorted. After the first test, to assess whether the AI assistance could improve human diagnostic performance, the examinee evaluated the same 100 images again with reference to the predictions of the final ensemble model. The performance of the AI model was explained to have the accuracy of 86% on the test dataset, which is approximately 10 to 20% higher than that of humans before the second test. The accuracy of the 100 images of the AI model and the ratio

of the two labels were not noticed by the examinee until all tests were performed. An expert was defined as having a medical career duration of >= 10 years.

Sub Study

We generated models to predict the BNP values. Images with BNP values >= 2000 pg/mL were excluded from the three datasets; the top two full connected layers were replaced with three full connected layers (two layers of 1024 neurons with activation function mish and one layer with one neuron with sigmoid function). We trained the models using the last 11 layers, which are trainable for fine-tuning. In addition, we trained models to predict elevated BNP values with the cut-off as 100 pg/mL by modified hyperparameters with reference to the training from the main study.

Statistical Analysis and Calculations

Continuous variables are presented as medians (with first and third quartiles) or mean±standard deviation (SD) and categorical variables as numbers and percentages as appropriate. The Python 3.10.7 (Python Software Foundation, Delaware) and TensorFlow 2.10.1 (Google LCC, Mountain View, CA) were utilized for our machine learning and statistical analysis (14). The models were trained using an NVIDIA GeForce RTX 3080Ti GPU (NVIDIA Corporation, Santa Clara, CA) or the Google Colaboratory platform (Google LLC, Mountain View, CA). The difference in the accuracies of the first and second human tests was tested using Welch's t-test or paired t-test as appropriate. A P value < 0.05 was considered statistically significant. The models used in this study and the sample code to utilize them will be available on GitHub.

Results

Baseline Characteristics

An overview of this study is shown in Figure 1, and the baseline characteristics of the study patients are shown in Table 1. The diagnoses of the study patients (N = 1607) were heart failure (N = 471), acute heart failure (N =320), coronary artery disease (N = 517), acute coronary syndrome (N = 176), hypertrophic cardiomyopathy (N = 32), interstitial pneumonia (N = 64), and hemodialysis (N = 23). A total of 10103 chest X-ray images were collected, 1061 (66%) patients with 6697 (66%) images, 273 (17%) patients with 1693 (17%) images, and 273 (17%) patients with 1713 (17%) images were assigned to the training, validation, and test datasets, respectively.

Performance of the AI

The 31 models (weak learners) were generated and trained (Figure 2). The performances of the models are shown in Table 2, Figure 3, and Central Illustration. The accuracy and PR-AUC of the post-refined models using featured map are 0.818±0.042 and 0.909±0.021, respectively. The performance metrics of the final ensemble model were as follows: accuracy was 0.855, precision was 0.873, recall (sensitivity) was 0.827, specificity was 0.882, F1 score was 0.850, ROC AUC was 0.929, and PR AUC was 0.934.

Performance of Human

Thirty-five persons including 20 medical doctors, and 13 cardiologists, were tested using the

100 images. The duration of medical practice among the subjects was 10.2±9.0 years, and 16 of them were experts. The accuracy of the AI model for this test was 0.92, the accuracy of the first human test was 0.708±0.049, and that of the AI-assisted test was 0.829±0.069, which was significantly higher than the non-assisted test ($P < 0.001$) (Figure 3 and Central Illustration). The accuracy of the medical doctors and experts was higher than that of non-medical doctors and non-experts in the non-assisted rest, respectively (0.687±0.032 vs. 0.725±0.054, $P = 0.014$; 0.692±0.042 vs. 0.728±0.051, $P = 0.030$); however, with the AI assistance, the accuracy of the medical doctors was similar to that of the non-medical doctors (0.843±0.074 vs. 0.818±0.064, $P = 0.289$), and the accuracy of the non-experts were rather higher than those of the experts (0.851±0.074 vs. 0.803±0.054, $P = 0.033$, Figure 4). In the non-assisted test, the accuracy had a weak positive correlation with the duration of medical careers ($r = 0.414$, $P = 0.014$), while in the AI-assisted test it had a weak negative correlation ($r = -0.347$, $P = 0.041$). For the eight images that were incorrectly predicted by the AI model, the accuracy for humans was 0.301±0.124. The hard ensemble prediction of the human with majority voting, the accuracy was 0.80 with the first test and 0.88 with the AI-assisted test.

Sub Study

We generated a model to predict BNP values. For weak learners, the mean absolute errors (MAE) of the predicted BNP values and the true BNP values were 208 pg/mL, their mean squared errors were $1.01*10^5$ pg^2/mL^2, and there was no significant difference between the predicted and true BNP values ($P = 0.274$). The performances of the models to predict elevated BNP level with a cut-off as 100 pg/mL were also as well as the models with those of 200 pg/mL

(Table 2 and Central Illustration).

Discussion

This study reports the development of a high-performing model for predicting elevated BNP levels from chest X-ray images, which can improve human diagnostic accuracy.

The strengths of our study are as follows: First, this is the first report to show that the AI model can predict elevated BNP levels from chest X-ray images with superior performance compared to experienced cardiologists. Our models were based on the hypothesis that chest X-ray images contain features that are associated with elevated BNP levels, which may indicate heart failure. Findings of chest X-ray imaging predicting heart failure have reported cardiomegaly, pulmonary venous congestion, interstitial or alveolar edema, and cephalization (1,3). Cephalization and interstitial and alveolar edema are reported to be modestly specific for heart failure but not very sensitive (sensitivity of 47 – 61% for elevated pulmonary capillary wedge pressure). In these previous reports, chest X-ray images were evaluated in humans, and our study demonstrated that the AI models had higher diagnostic performance than humans regarding predicting elevated BNP levels. Therefore, the first part of the hypothesis must be true. The featured map images revealed that our models were featured in the above chest X-ray findings (Figure 2), which are similar to previous human reports about chest radiography and heart failure. However, plasma natriuretic peptide levels are affected by several factors such as age, sex, hemoglobin level, renal function, left ventricular end-diastolic pressure, and left ventricular ejection fraction, and elevated BNP must not be evaluated perfectly by chest X-ray imaging (15). On the other hand, it could be said that the predictability, as measured by the ROC

AUC, is higher than 0.929 in our test dataset population. Matsumoto et al. reported that transfer-learned VGG16 models predict heart failure from the chest X-ray images labeled by two cardiologists (16). One of the issues of their study was that heart failure was diagnosed by two cardiologists, and our study showed that physicians' performance for predicting elevated BNP levels is relatively poor compared to the AI models, even if they are experienced. Other reports showed that chest radiography could predict pulmonary hypertension or a higher pulmonary artery wedge pressure (17,18). One of the weaknesses is the use of pulmonary hypertension, which may be affected by posture, such as prone or sitting, and the time interval between catheterization and X-ray. Another issue is the relatively low predictive performance, which may have been caused by the small sample size. Zou et al. reported that their model predicted pulmonary hypertension from a chest X-ray image with an ROC AUC of 0.967; however, their models were trained and tested in selected images that excluded pleural/pericardial effusion, pneumothorax, pulmonary atelectasis, and massive consolidation obstruction of the right ventricular outflow tract, or pulmonary artery were excluded; our study consisted of all available chest X-ray images, including trauma, interstitial pneumonia, mechanically ventilated patients, acute or chronic respiratory failure, implanted cardiac devices, dextrocardia, congenital heart diseases, dialysis, and hypertrophic cardiomyopathy in addition to the above excluded conditions, and our model may be robust.

Second, the performance of the models was compared with that of front-line physicians. Several AI studies have reported that the AI model could accurately predict medical features from traditional tests; however, most of them did not assess physicians' performance. Our study showed that not only the ensemble models but also weak learners had higher performance than

experienced physicians and revealed non-experts' performance in diagnosing elevated BNP.

Third, we showed that the suggestions of the AI model could improve the diagnostic performance of humans. Even if a supportive tool exhibits excellent performance, it is useless if users do not trust or use it. There is no report that the AI model's assistance improves the diagnostic performance of humans, and previous reports concerning medical prediction models only discussed the models' performance itself. This study revealed that performance was positively associated with the duration of medical careers; however, AI assistance improves the diagnostic performance of both inexperienced and experienced ones, with the inexperienced ones achieving the same or higher performance as the experienced ones. This means that inexperienced individuals, with the help of a useful diagnostic tool, can perform as well as or better than experienced individuals alone. The limited improvement in the expert may be due to their self-confidence from their experience and distrust of the AI model, although they were told that the AI model's performance is superior to that of any human. On the other hand, a less experienced person might accept the AI's prediction without resistance due to a lack of self-confidence. Distrust for new technology or findings and self-confidence in them will be a new issue in AI; there has always been a similar distrust for the new in other categories. We must understand and adapt to new ideas, technologies, tools, and methods, including AI.

Fourth, the models of this study and sample codes to utilize them will be available on GitHub. This means that anyone can test and refine the models and test new ideas that the old to make way for the new.

Fifth, we improved the models using several techniques. In addition to classical standard methods such as data augmentation with shift, flip, and rotations, activation function "relu",

optimizer “Adam”, and convolutional layers, our model was designed to incorporate current or past state-of-the-art techniques, including data augmentations of MixUp, Random Erasing, and Cutmix, the activation function “mish”, regularization of label smoothing, optimizer of sharpness aware minimization, the architecture without convolutional layers such as vision transformer and MLP-mixer, and model refinement to control bias using featured map. Finally, the ensemble of these models was used to create the strong learner. Because our models were based on these technical aspects and our collected dataset was all-comer and had not been used in the previous reports concerning heart failure and chest X-ray, we thought this was one of the advantages of our study and may increase the model’s robustness and generalizability.

Sixth, good old chest radiographs are widely available in medical facilities. It has the potential to change the world because the prediction of a single chest X-ray image with one weak learner or small ensemble does not take much computing time, even in our personal computer without a graphics processing unit, has no time to wait with it, and can be easily integrated into an X-ray machine or smart phone.

The ensemble prediction could improve performance (19). In our study, soft voting was utilized and improved the performance from the single models (weak leaners) with an accuracy of 0.818 to the ensemble model of 0.855. For example, using the ensemble method, in the case of three weak learners with simple majority voting for binary prediction, if more than two weak learners predict the correct answer, the ensemble’s decision will be correct. If each weak learner is assumed to be independent and their probability is uniformly 0.8, the probability that more than two weak learners are correct is 0.896 (0.8 * 0.8 * 0.8 + 0.2 * 0.8 * 0.8 * 3). To improve the performance of the ensemble, the models’ performances must not be poor, and the correlation of

model predictions must not be too strong. It is similar to the performance of a heart-team in that members who totally concur with others, or are silent members, are equal to nothing in that they do not contribute to improving the quality of decisions, and active members without a certain level of performance deteriorate the performance. With respect to an ensemble model, weak learners should be different from each other because the correlation should not be extremely high, although we intended to make all models good in a way that seemed best. Sensitivity and specificity are in a trade-off relationship, so even if a weak leaner is not the best in terms of overall accuracy, using it as the weak leaner in an ensemble model can improve the accuracy and generalization performance of the ensemble model because of the diversity of weak learners. Possible approaches to increasing the accuracy of the ensemble model include replacing inaccurate learners with more accurate models, adding more accurate models to the ensemble model, or selecting or retraining weak learners according to the ensemble model's performance. However, we did not perform these because all of them resulted in overfitting to the test dataset and loose generalizability and robustness, finally leading to deteriorated performance. This is one of the difficulties in implementing the ensemble method. As the correlation of models' predictions was not known before the termination of training models, we cannot connect the dots looking forward; we can only connect them looking backwards, so we have to trust that the weak learners will somehow connect in future and have to train them with a variety of sophisticated methods, as well as for teaching humans.

Study Limitations

The accuracy of the models was affected by the proportion of difficult-to-diagnose images in the

test dataset. When evaluating other images, if the proportion of such images is different from our test dataset, the accuracy may be different from that in our study. However, as our dataset was collected from daily clinical practice with all-comers, we suppose that there is a certain degree of generalizability and robustness. In the sub study predicting absolute BNP values, there was no statistical difference between the predicted and true BNP value; however, the MAE was 208 pg/mL, about the same as the cut-off for BNP in this study, which is not as well as the binary prediction of BNP levels. One reason for this may be sample size, and performance could be improved with a larger dataset. Natriuretic peptides are utilized for the diagnosis of heart failure with absolute values or relative changes and for management with sequential relative changes. We performed this study using binary prediction, mainly because our main objective was to diagnose heart failure in this study. The clinical impact of the gaps between the predicted and true BNP values may depends on how they can be utilized in clinical situations. In this regard, human performance in predicting absolute BNP values from chest X-ray images was extremely poor beyond prediction in preliminary testing (data not shown). The cut-off BNP value is one of issues to be discussed. The cut-off values should be defined according to purpose while balancing sensitivity-specificity. We set the cut-off BNP values 200 pg/ml to diagnose unrecognized heart failure which should be managed early; however, a cut-off value of 100 or 125 pg/ml would be considered for other purpose or applications. In this study, the models with the cut-off of 100 pg/ml showed similar excellent performance to those with the cut-off of 200 pg/ml. We labeled the X-ray images based on the BNP cut-off value, but it should be noted that heart failure is not diagnosed solely based on natriuretic peptide values. One of our goals is to diagnose unrecognized heart failure which should be managed early, and it is not equal to

diagnosing elevated BNP levels alone.

Conclusions

The AI model can predict elevated BNP levels from chest X-ray images with superior performance than experienced cardiologists and can improve the diagnostic performance of humans, ranging from non-experts to experienced cardiologists. These models may be useful in clinical settings. Further studies are necessary for distrust of innovative technologies and utilization gap.

Clinical Perspective

Competency in Systems-Based Practice: The AI model outperform expert cardiologists in predicting elevated BNP values from chest X-ray images, potentially improving diagnostic performance for both non-experts and experts.

Translational Outlook: Utilization gap is one of raising issue, and we must understand and adapt to innovative ideas, technologies, tools, and methods, including AI.

References

1. Heidenreich PA, Bozkurt B, Aguilar D et al. 2022 AHA/ACC/HFSA Guideline for the Management of Heart Failure: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. J Am Coll Cardiol 2022;79:e263-e421.
2. Stevenson LW, Perloff JK. The limited reliability of physical signs for estimating hemodynamics in chronic heart failure. JAMA 1989;261:884-8.
3. Chakko S, Woska D, Martinez H et al. Clinical, radiographic, and hemodynamic correlations in chronic congestive heart failure: conflicting results may lead to inappropriate care. Am J Med 1991;90:353-9.
4. Maisel AS, Krishnaswamy P, Nowak RM et al. Rapid measurement of B-type natriuretic peptide in the emergency diagnosis of heart failure. N Engl J Med 2002;347:161-7.
5. Mueller C, Scholer A, Laule-Kilian K et al. Use of B-type natriuretic peptide in the evaluation and management of acute dyspnea. N Engl J Med 2004;350:647-54.
6. McLellan J, Bankhead CR, Oke JL, Hobbs FDR, Taylor CJ, Perera R. Natriuretic peptide-guided treatment for heart failure: a systematic review and meta-analysis. BMJ Evid Based Med 2020;25:33-37.
7. Quer G, Arnaout R, Henne M, Arnaout R. Machine Learning and the Future of Cardiovascular Care: JACC State-of-the-Art Review. J Am Coll Cardiol 2021;77:300-313.
8. Fuster V, Bozkurt B, Chandrashekhar YS et al. JACC Journals' Pathway Forward With AI Tools: The Future Is Now. J Am Coll Cardiol 2023.

9. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 2012;25.

10. 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.

11. Gessert N, Lutz M, Heyder M et al. Automatic Plaque Detection in IVOCT Pullbacks Using Convolutional Neural Networks. IEEE Trans Med Imaging 2019;38:426-434.

12. Myhre PL, Prescott MF, Murphy SP et al. Early B-Type Natriuretic Peptide Change in HFrEF Patients Treated With Sacubitril/Valsartan: A Pooled Analysis of EVALUATE-HF and PROVE-HF. JACC Heart failure 2022;10:119-128.

13. The Japanese Heart Failure Society Committee on Heart Failure Prevention, Sato Y, Yoshimura M et al. Guidelines regarding management for heart failure using blood BNP and NT-proBNP levels. 2014. Available at: http://www.asas.or.jp/jhfs/english/outline/guidelines_20180822.html. Accessed March 31, 2023.

14. Kluyver T, Ragan-Kelley B, Pérez F et al. Jupyter Notebooks-a publishing format for reproducible computational workflows, 2016.

15. Tsutamoto T, Wada A, Sakai H et al. Relationship between renal function and plasma brain natriuretic peptide in patients with heart failure. J Am Coll Cardiol 2006;47:582-6.

16. Matsumoto T, Kodera S, Shinohara H et al. Diagnosing Heart Failure from Chest X-Ray Images Using Deep Learning. Int Heart J 2020;61:781-786.

17. Zou XL, Ren Y, Feng DY et al. A promising approach for screening pulmonary

hypertension based on frontal chest radiographs using deep learning: A retrospective study. PloS one 2020;15:e0236378.

18. Hirata Y, Kusunose K, Tsuji T, Fujimori K, Kotoku J, Sata M. Deep Learning for Detection of Elevated Pulmonary Artery Wedge Pressure Using Standard Chest X-Ray. Can J Cardiol 2021;37:1198-1206.

19. Lam L, Suen SY. Application of majority voting to pattern recognition: an analysis of its behavior and performance. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 1997;27:553-568.

Figure Legends

Figure 1: Overview of Making AI Models

We designed AI models to predict higher BNP levels as follows: First, as weak learners, we fine-tuned 31 modified pre-trained image recognition models. The bias of the models was checked using a featured map of GRAD-CAM or attention maps. If the model strongly focused on irrelevant objects for the classification, we discarded it and retrained a new model until the model no longer focused on them. Second, we created a soft ensemble prediction model using the weak learners.

Figure 2: Chest X-Ray Images and Featured Map of Weak Learners

(**A**) X-ray image of the chest with BNP 9 pg/mL and GRAD-CAM image generated by the EfficientNetV2L-based-model. (**B**) The EfficientNetV2L-based-model featured pulmonary congestion and pacemaker, with BNP of 798 pg/mL. (**C**) The attention map revealed the Vit-b16-based-model featured in the caption, the pulmonary vessel, and the diaphragmatic line with BNP 413 pg/mL. The caption in the upper left corner is a Kanji meaning supine position. (**D**) The EfficientNetV2S-based-model featured in the captions of images with BNP 824 pg/mL. The caption in the upper left corner is a Kanji meaning sitting position. (**E**) The model utilized in **D** was retrained, and the new model was not featured in the captions but in the pleural effusion, cardiomegaly, pulmonary artery, air bronchogram, and Kerley A lines. The caption in the upper left corner is a Kanji meaning sitting position.

Figure 3: The Representative Images and Predictions of the AI Models and Humans

The chest X-ray images and their featured map images, age, sex, BNP value, AI prediction, and human accuracy are shown. The featured map images were generated by the EfficientNetV2S-based-model using GRAD-CAM. The caption in the upper right corner of A and the upper left corner of B, C, and D are Kanji meaning sitting and standing position, respectively.

Figure 4: The Accuracy of the Humans' Testing

(A and B) Both the non-experts (0.692±0.042 vs. 0.851±0.074, $P < 0.001$) and experts (0.728±0.051 vs. 0.803±0.054, $P < 0.001$) improved their accuracy with the assistance of the AI model. (**C**) The increase in the accuracy with the AI assistance was greater for the non-experts than for the experts (0.159±0.069 vs. 0.074±0.052, $P < 0.001$). The data was presented as mean±SD.

Central Illustration: AI Model to Predict Elevated BNP Improved Diagnostic Performance of Human

We developed the final ensemble model to predict elevated BNP levels from 31 weak learners. The AI model achieved a ROC AUC of 0.929. The accuracy of human testing was significantly improved from 0.708±0.049 to 0.829±0.069 ($P < 0.001$) with AI assistance. In the non-assisted test, the accuracy had a weak positive correlation with the duration of medical careers ($r = 0.414$, $P = 0.014$). In the AI-assisted test it had a weak negative correlation ($r = -0.347$, $P = 0.041$).

Table 1. Characteristics of the Study Patients and Materials

Patient	
Study patients	1607
Male sex	980 (61)
Diagnosis	
Heart failure	471
Acute heart failure	320
Coronary artery disease	517
Acute coronary syndrome	176
Hypertrophic cardiomyopathy	32
Congenital heart disease	14
Atrial fibrillation	253
Peripheral artery disease	97
Pericardial effusion	6
Aortic disease	68
Hemodialysis	23
Chronic kidney disease	171
Pneumonia	62
Interstitial pneumonia	64
Venous thromboembolism	43
Pneumothorax	9
Lung carcinoma	16

Trauma	5
Chest pain syndrome	29
Dataset, patients	
Train	1061 (66)
Valid	273 (17)
Test	273 (17)
Images	
Chest X-ray images, N	10103
Age, y	74 (64 – 81)
Plasma BNP, pg/ml	158 (47 – 569)
Dataset, images	
Train	6697 (66)
BNP >= 200 pg/ml	2994 (45)
Valid	1693 (17)
BNP >= 200 pg/ml	726 (43)
Test	1713 (17)
BNP >= 200 pg/ml	852 (50)

Data presented as N (%) or median (interquartile range).

Table 2. Performance of Models with BNP Cut-Off as 200 pg/ml

Model*	**Input size**	**Featured caption**	**Accuracy**	**Precision**	**Recall (Sensitivity)**	**Specificity**	**F1 score**	**ROC AUC**	**PR AUC**
Cut-off: 200 pg/ml (main study)									
VGG16	224	0	0.842	0.868	0.805	0.879	0.835	0.919	0.926
VGG16	384	0	0.823	0.877	0.749	0.897	0.808	0.907	0.914
VGG19	224	0	0.838	0.869	0.793	0.882	0.829	0.913	0.918
VGG19	384	0	0.841	0.871	0.799	0.883	0.833	0.915	0.923
VGG19_2	384	0	0.834	0.844	0.816	0.852	0.830	0.910	0.917
InceptionResNetV2	299	1	0.829	0.852	0.793	0.864	0.822	0.909	0.909
Xception	299	3	0.791	0.916	0.638	0.942	0.752	0.899	0.904
Xception_2	299	6	0.840	0.862	0.807	0.873	0.834	0.919	0.926
MobileNetV3Small	224	5	0.809	0.876	0.717	0.900	0.789	0.914	0.918
MobileNetV3Large	224	1	0.837	0.863	0.799	0.875	0.830	0.912	0.917

ResNet-RS101	224	3	0.823	0.884	0.741	0.904	0.806	0.902	0.909
ResNet-RS200	224	7	0.823	0.879	0.745	0.899	0.807	0.903	0.906
EfficientNetV2B0	224	0	0.819	0.841	0.784	0.854	0.811	0.900	0.900
EfficientNetV2B1	240	0	0.831	0.847	0.806	0.856	0.826	0.904	0.909
EfficientNetV2B2	260	0	0.821	0.801	0.852	0.791	0.826	0.907	0.912
EfficientNetV2B3	300	0	0.825	0.873	0.758	0.891	0.811	0.908	0.913
EfficientNetV2S	384	0	0.844	0.867	0.810	0.878	0.837	0.914	0.919
EfficientNetV2M	480	12	0.835	0.853	0.807	0.863	0.830	0.915	0.919
EfficientNetV2L	480	7	0.832	0.839	0.819	0.845	0.829	0.909	0.906
ConvNeXtTiny	224	0	0.831	0.865	0.781	0.879	0.821	0.922	0.924
ConvNeXtSmall	224	0	0.827	0.866	0.772	0.882	0.816	0.918	0.922
ConvNeXtBase	224	8	0.830	0.834	0.820	0.839	0.827	0.910	0.913
ConvNeXtLarge	224	8	0.832	0.835	0.825	0.839	0.830	0.906	0.908
ConvNeXtXLarge	224	0	0.820	0.889	0.727	0.911	0.800	0.913	0.919
Vit-b16	224	10	0.822	0.816	0.828	0.816	0.822	0.906	0.909
Vit-b16_2	224	10	0.827	0.816	0.843	0.812	0.829	0.906	0.912

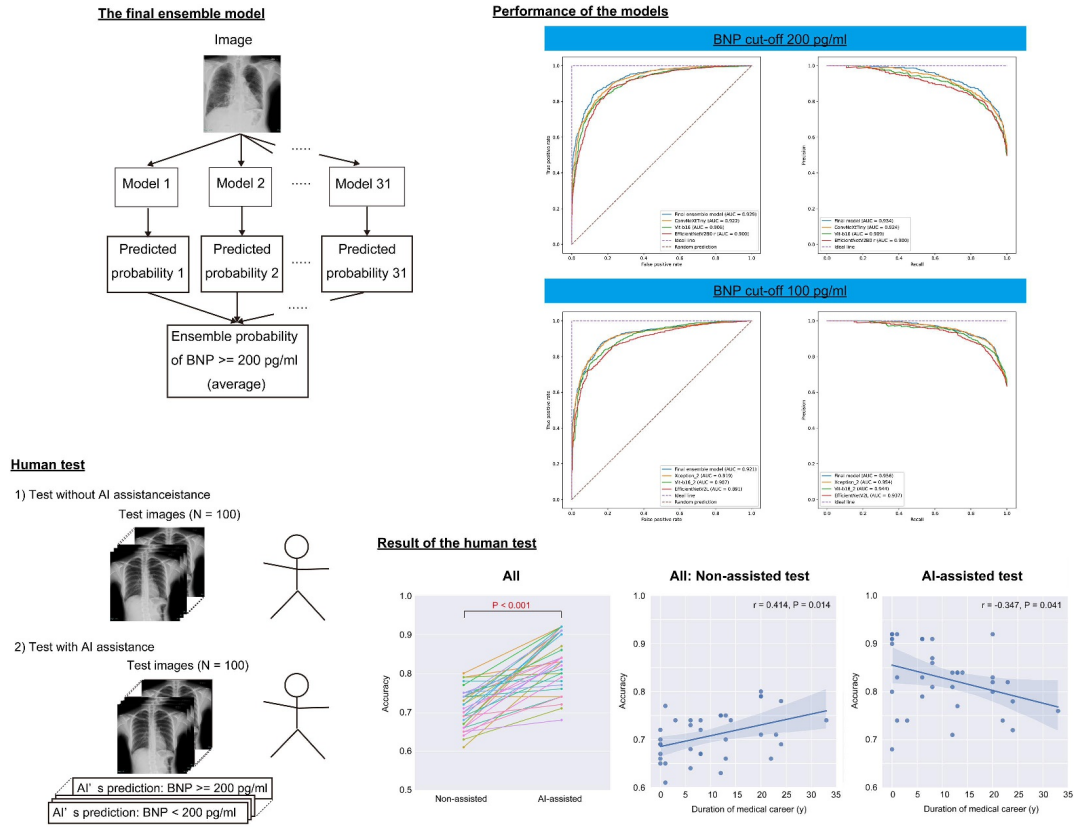
Vit-b16	320	12	0.811	0.863	0.736	0.885	0.794	0.905	0.909
Vit-b16	384	10	0.828	0.903	0.733	0.922	0.809	0.919	0.924
MLPMixerB32	224	-	0.817	0.856	0.759	0.874	0.804	0.901	0.898
MLPMixerB32_2	224	-	0.605	0.558	0.985	0.231	0.713	0.834	0.817
MLPMixerB32_3	224	-	0.768	0.716	0.884	0.653	0.791	0.864	0.862
Overall									
			0.818±0.042	0.845±0.064	0.791±0.060	0.844±0.125	0.813±0.026	0.906±0.017	0.909±0.021
Ensemble model									
(Final model)			0.855	0.873	0.827	0.882	0.850	0.929	0.934
Cut-off: 100 pg/ml (sub study)									
VGG16	224	3	0.838	0.843	0.915	0.703	0.877	0.915	0.952
VGG16	384	5	0.827	0.836	0.904	0.692	0.869	0.907	0.948
VGG19	224	2	0.838	0.852	0.902	0.728	0.876	0.913	0.951
VGG19	384	9	0.841	0.862	0.891	0.752	0.877	0.907	0.947
InceptionResNetV2	299	6	0.847	0.861	0.905	0.746	0.883	0.912	0.951

Xception	299	4	0.853	0.871	0.902	0.768	0.886	0.912	0.949
Xception_2	299	8	0.853	0.861	0.917	0.743	0.888	0.919	0.954
MobileNetV3Small	224	12	0.832	0.834	0.917	0.684	0.874	0.899	0.936
MobileNetV3Large	224	11	0.831	0.851	0.89	0.73	0.87	0.906	0.946
ResNet-RS101	224	0	0.828	0.845	0.892	0.716	0.868	0.891	0.93
ResNet-RS200	224	10	0.853	0.857	0.922	0.733	0.888	0.908	0.943
EfficientNetV2B0	224	0	0.837	0.859	0.889	0.748	0.874	0.898	0.936
EfficientNetV2B1	240	0	0.838	0.851	0.902	0.727	0.876	0.903	0.944
EfficientNetV2B2	260	1	0.842	0.864	0.892	0.756	0.878	0.908	0.946
EfficientNetV2B3	300	1	0.83	0.85	0.889	0.727	0.869	0.907	0.947
EfficientNetV2S	384	5	0.834	0.859	0.883	0.749	0.871	0.914	0.951
EfficientNetV2S_2	384	6	0.838	0.853	0.9	0.732	0.876	0.914	0.952
EfficientNetV2M	480	1	0.842	0.866	0.889	0.762	0.877	0.914	0.952
EfficientNetV2L	480	4	0.818	0.877	0.83	0.797	0.853	0.891	0.937
ConvNeXtTiny	224	3	0.815	0.837	0.88	0.701	0.858	0.886	0.927
ConvNeXtSmall	224	10	0.846	0.864	0.899	0.754	0.881	0.905	0.943

ConvNeXtBase	224	2	0.842	0.85	0.912	0.72	0.88	0.903	0.942
ConvNeXtLarge	224	13	0.825	0.867	0.855	0.773	0.861	0.89	0.931
Vit-b16	224	13	0.846	0.849	0.922	0.716	0.884	0.915	0.95
Vit-b16_2	224	12	0.831	0.849	0.891	0.725	0.87	0.907	0.944
Vit-b16	320	7	0.842	0.852	0.908	0.727	0.879	0.911	0.948
Vit-b16_2	320	9	0.822	0.816	0.929	0.636	0.869	0.899	0.938
Vit-b16	384	11	0.83	0.836	0.91	0.69	0.871	0.908	0.946
MLPMixerB32	224		0.843	0.88	0.871	0.794	0.876	0.907	0.948
MLPMixerB32_2	224		0.843	0.866	0.891	0.76	0.878	0.908	0.947
MLPMixerB32_3	224		0.838	0.851	0.903	0.725	0.876	0.906	0.945
Overall									
			0.837±0.010	0.854±0.014	0.897±0.020	0.733±0.033	0.875±0.008	0.906±0.008	0.945±0.007
Ensemble model									
			0.853	0.867	0.909	0.759	0.888	0.921	0.956

ROC indicates receiver-operating-characteristics curve; AUC, area under curve, PR, precision-recall curve.

*Model means the base-model of finetuning for the weak learners.



Central Illustration: AI Model to Predict Elevated BNP Improved Diagnostic Performance of Human

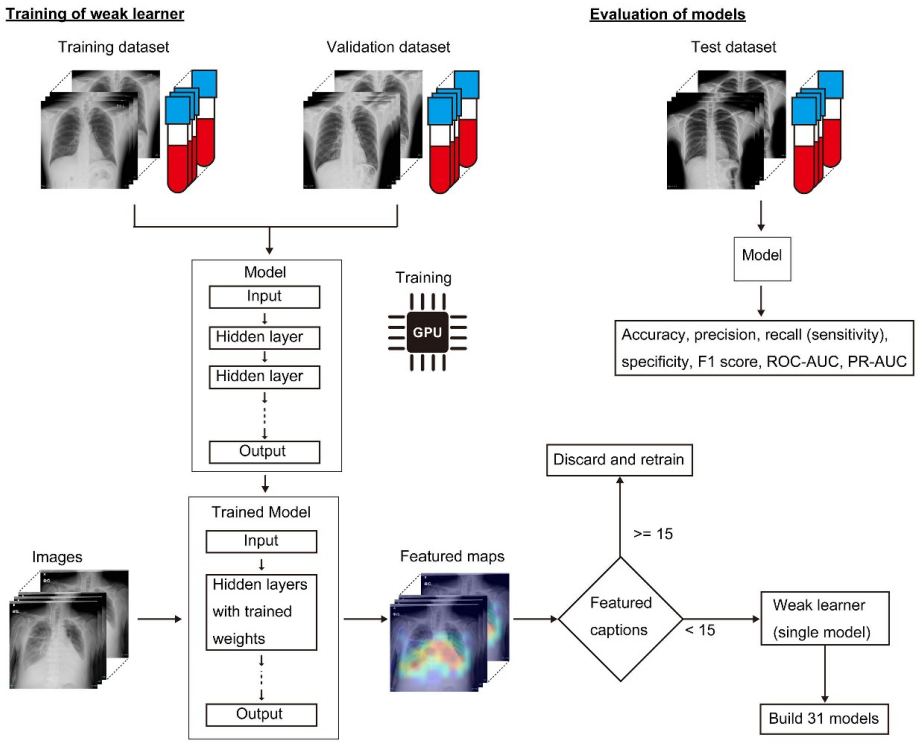


Figure 1: Overview of Making AI Models

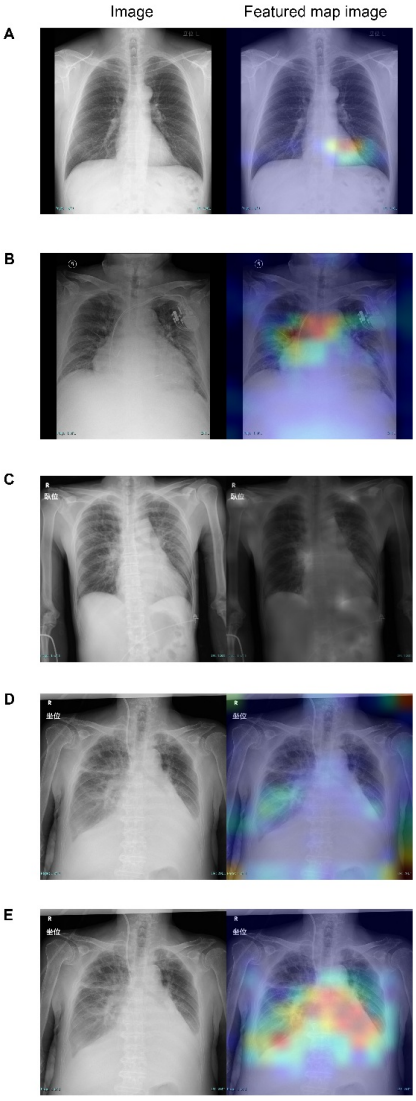


Figure 2: Chest X-Ray Images and Featured Map of Weak Learners

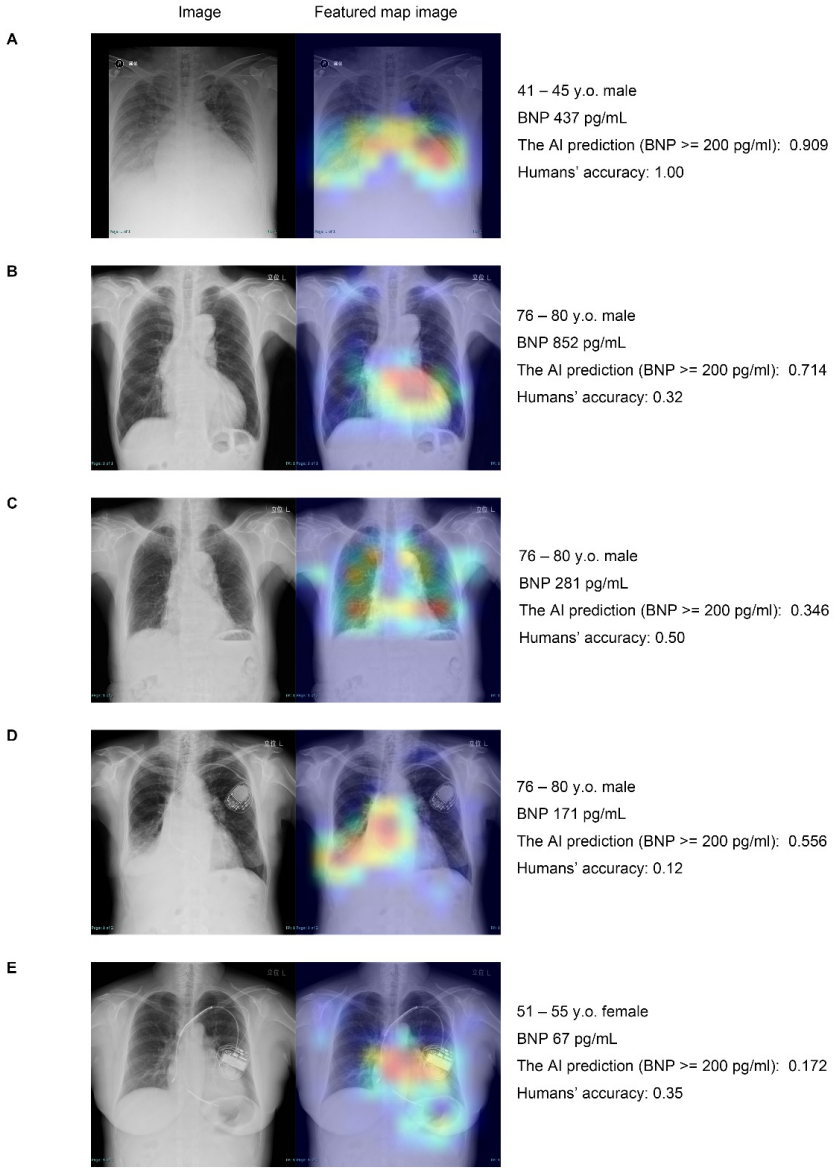


Figure 3: The Representative Images and Predictions of the AI Models and Humans

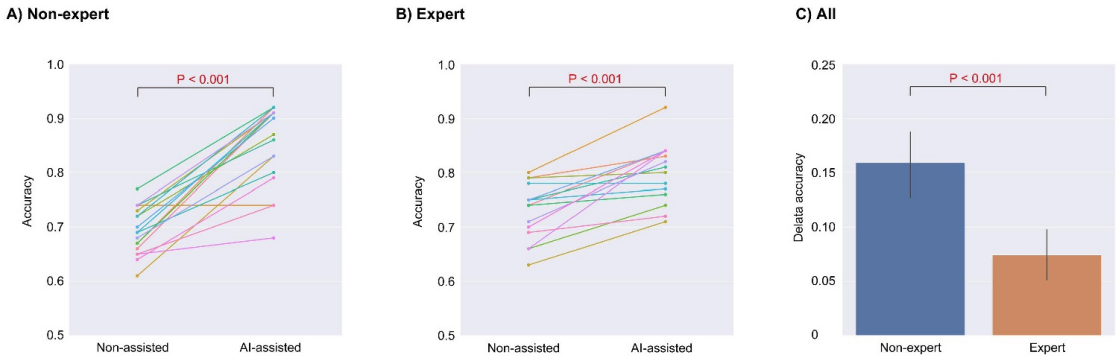


Figure 4: The Accuracy of the Humans' Testing