Does artificial intelligence dream of natriuretic peptide?: The machine learning model improves human diagnostic performance of heart failure using chest radiography

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

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

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

Short title: *Kagawa et al.; AI, BNP, and chest X-ray*

Total word count: 4812 words

There are no conflicts of interest. There is no financial support for this study.

Address for correspondence:

Eisuke Kagawa, MD, PhD

Department of Cardiology, Hiroshima City Asa Hospital

1-2-1, Kameyamaminami, Asakita-ku, Hiroshima, 731-0293 Japan

Phone: (81) 82-815-5211

Fax: (81) 82-814-1791

E-mail: ekagawa007@gmail.com

ABSTRACT

BACKGROUND: Heart failure pandemic is a medical issue, and its diagnosis can be challenging. We developed an artificial intelligence (AI) model for predicting elevated brain natriuretic peptide (BNP) levels from chest X-ray images and assessed the hypothesis that the AI model has sufficient performance and could improve the diagnostic performance of humans.

METHODS: Patients who underwent chest radiography and BNP testing on the same day were enrolled in this study. The patients were divided into three dataset groups. The images of the chest X-ray were labeled based on plasma BNP levels >= 200 pg/mL. We trained neural network models that predicted the BNP label using the fine-tuning method, which were refined to control bias using featured maps, and the final ensemble model was constructed. Humans were trained with the training dataset and tested to predict BNP 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 (accuracy 0.818±0.04 and precision-recall (PR) area-under-curve (AUC) 0.909±0.03) as the final model with an accuracy of 0.855, precision of 0.873, recall of 0.827, specificity of 882, f1 score of 0.850, receiver-operating-characteristics AUC of 0.929, and PR AUC of 0.934. The 35 persons underwent the test, and the accuracy of the human significantly improved from 0.708±0.049 to 0.829±0.069 ($P < 0.001$) with AI assistance, with an accuracy of 0.92. The accuracy of the experts was higher than that of non-experts in the non-assisted test (0.692±0.042 vs. 0.728±0.051, $P = 0.030$); however, with the AI assistance, the accuracy of the non-experts were rather higher than those of the experts (0.851±0.074 vs. 0.803±0.054, $P = 0.033$).

CONCLUSIONS: The AI model can predict elevated BNP levels from chest X-ray images with better performance than experienced cardiologists and could improve the diagnostic performance of humans, ranging from non-experts to experienced cardiologists. Further studies are necessary for distrust of innovative technologies and utilization gap.

Key Words: deep learning, neural network, chest X-ray, brain natriuretic peptide, heart failure

Non-Standard Abbreviations and Acronyms

AI	Artificial intelligence
AUC	Area-under curve
BNP	Brain natriuretic peptide
GRAD-CAM	Gradient-class activation maps
PR	Precision-recall
ROC	Receiver-operating-characteristics
SAM	Sharpness aware minimization
SD	Standard deviation

Clinical Perspective

What is new?

- The AI model outperform expert cardiologists in predicting elevated BNP values from chest X-ray images.
- The AI model could improve both non-expert and expert diagnostic performance, with non-experts achieving the same or higher performance as experts.
- This study showed that the AI model improves human performance, and the degree of improvement is different for different individuals, although previous medical reports concerning prediction models only discussed the performance itself.

What are the clinical implications?

- The AI model may be able to support and improve our daily practice not only for non-expert, but for experienced cardiologists to pay attention from X-ray images to heart failure status requiring care.
- It would be easy to incorporate hard equipment or software into X-ray machines or external devices to assist our daily practice.
- Utilization gap is one of raising issue, and we must understand and adapt to innovative ideas, technologies, tools, and methods, including AI.

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 may 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. Of course, it also requires a decision to conduct 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.[7-9] 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.[10-13] 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

The source codes and weights of the models used in this study will be available on GitHub.

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.[14] The chest X-ray images were labeled binary according to the level of plasma BNP < 200 pg/mL or >= 200 pg/mL. 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.

Data Collection Image Preprocessing

Chest X-ray images were acquired using Digital Imaging and Communication in Medicine, PNG, or BMP images, and non-squared 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 with a featured map of gradient-class activation maps (GRAD-CAM) or attention maps, and if they seemed to focus on regions essentially unrelated to heart failure, their weights were discarded, and they were retrained anew. Second, we created a soft ensemble prediction model using the weak learners.

Model Awakens (Weak Learners)

The summary of the fine-tuning models that were based on well-known image recognition models were shown on GitHub. These models were loaded with weights that were pre-trained by ImageNet, which is a large non-medical image dataset.[15] The top layers of the fine-tuning models were designed as two full connected layers with neurons 2048 and 2, and each layer had a previous dropout layer.[16] The provability of the BNP >= 200 pg/mL was determined by two neurons with softmax activation function in the final output layers, which generated numbers ranging from 0 and 1. Data augmentation with random width shifts ranging from 10%, height shifts ranging from 15%, rotation ranging from 15 °, horizontal flips, Mixup, RandomErasing, and CutMix techniques were utilized in the models' training.[17-19] We regularized the neural networks with Label Smoothing.[20] All of the models in this study were optimized using sharpness aware minimization (SAM) using RAdam or Adam.[21-23] The hyperparameters are shown in Table S1. The input image sizes of the models were determined based on the pre-training image sizes. The training of models was visualized with the accuracy and loss of the training and validation datasets on every epoch, and all models were trained for a minimum of 100 epochs. If overfitting or learning seemed to have converged 100 epochs, training was terminated. If training did not seem to have converged at 100 epochs, training was continued for an additional 50 epochs until overfitting or learning convergence was confirmed. In the training of the fine-tuning models, all weights of the layers in the model except the top three (two full connected layers and one dropout layer) were frozen (set as non-trainable), and only the top three layers and layers of batch normalization were trained in the first 10 epochs. In the next 20 epochs, the specific numbers of the top layers were set to trainable and trained (Table S1). After

31 epochs, all layers were set as trainable and trained. The weights of the models with the minimum loss of the validation dataset at the end of the epoch were saved and utilized for the study. These models were candidates for weak learners in the final ensemble model.

Model Refinement to Control Bias by Featured Map

We refined the model using a featured map to control for bias. We refine the models based on our hypothesis that if the model strongly featured objects that were not essentially irrelevant to the theorem or the mechanism that differentiated images to each label, we would not use the model's weights for the final ensemble model and instead train a new one until the model did not feature in them and we could obtain an improved model. In this study, we generated 10 images that were processed with GRAD-CAM for the models with convolutional architecture or attention maps for those with attention architecture and visualized where the models featured in the images. Several objects were not supposed to be irrelevant to heart failure, such as captions, calipers, leads, electrodes, tubes, and catheters. We defined only captions as fundamentally unrelated to heart failure in this study because the other things were difficult to distinguish in the featured map. We defined "strongly associated" as focusing on 15 or more captions in 10 images with 29 captions. The cut-off of the captions was not set to 0 because there are millions to hundreds of millions of trainable parameters for a single model, and it was inevitable that some would focus on irrelevant ones. The trained fine-tuned models based on MLP-mixer were not utilized this method and were enrolled as weak learners as-are because we could not interpret the featured maps by the architecture.

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 >= 200 pg/mL. 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 >= 200 pg/mL, 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 Protocol and AI Assistance

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 < 200 pg/mL and 50 with BNP >= 200 pg/mL 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 found to be 86% accurate in the test dataset, which may be 10 – 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 and divided into three datasets by patient; 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 [Q1 – Q3]) 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.[24] The models were trained on a personal computer with an NVIDIA GeForce RTX 3080Ti GPU (NVIDIA Corporation, Santa Clara, CA) or 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.

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 main diagnoses or complaints 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.

Training and Refining Models to Control Bias by Featured Map

The 31 models (weak learners) were generated and trained (Table 2). The finetuned models based on InceptionResNetV2, MobileNetV3Small, MobileNetV3Large, ResNet-RS200, EfficientNetV2B0, EfficientNetV2B2, EfficientNetV2B3, EfficientNetV2S, and EfficientNetV2L were strongly featured in the captions of 10 prespecified images; the models were retrained (Figure 2). The MobileNetV3Large-based-model was retrained two times.

Results of the AI

The performances of the models are shown in Table 2 and Figure 3. The performance of the models that were pre-refined by the featured map had an accuracy of 0.817±0.042 and a PR

AUC of 0.908±0.021. The performance of the 10 models that were strongly featured in the caption appeared to be slightly improved after the refining with a featured map with an accuracy of 0.824±0.011 to 0.827±0.010 (P = 0.721), precision of 0.851±0.019 to 0.855±0.025 (P = 0.746), recall of 0.778±0.040 to 0.786±0.041 (P = 0.705), specificity of 0.864±0.023 to 0.866±0.034 (P = 0.869), F1 score of 0.815±0.014 to 0.818±0.015 (P = 0.688), ROC AUC of 0.904±0.007 to 0.908±0.005 (P = 0.097), and PR of 0.909±0.010 to 0.911±0.006 (P = 0.584). 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 of the final ensemble model was the accuracy was 0.855, the precision was 0.873, the recall (sensitivity) was 0.827, the specificity was 0.882, the F1 score was 0.850, the ROC AUC was 0.929, and the PR AUC was 0.934, respectively. The accuracy of the ensemble model with pre-refined to control bias by featured map was 0.851, and the refining to control bias by featured map and the ensemble method showed improved performance compared to the weak learners and those with strong features in the captions of chest X-ray images.

Performance of Human and the Non-Experts Strikes Back

Thirty-five persons including 15 non-healthcare providers, 20 medical doctors, and 13 cardiologists, were tested using the 100 images. The examinee's medical career lasted 10.2±9.0 years, and 16 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 4 and 5). 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). 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 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 Figure 3).

Discussion

This study is the first to report accurate neural network models to predict heart failure from a chest X-ray image, which was superior to that of experienced cardiologists and can improve human diagnostic performance.

The strengths of our study are as follows: First, this is the first report to show that the AI model can predict heart failure from chest X-ray images with superior performance compared to experienced cardiologists. Our models were based on the hypothesis that chest X-ray images show elevated BNP levels, which are associated with 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, so our hypothesis must be true to some extent. The featured map images revealed that our models were featured in the above chest X-ray findings (Figure 2), and the sensitivity of our final model is also relatively lower than its specificity, which is similar to previous human reports. 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.[25] 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.[26] One of the issues of their study was that heart failure was diagnosed by two cardiologists, and our study showed that physicians' performance 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.[27,28] 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 and exist in a short twilight moment of the state-of-the-art technology.

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. It means that inexperienced ones are proficient in using a useful diagnostic tool, they can perform better than experienced ones could do. 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 source codes and weights of the models were 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 SAM, the architecture without convolutional layers such as vision transformer and MLP-mixer, and novel model refinement to control bias using featured map. Finally, an ensemble of these models

is used to create a strong learner. To our knowledge, this is the first report on the use of GRAD-CAM or attention maps, to improve diagnostic performance. Featured maps such as GRAD-CAM or attention maps, are typically utilized to determine the important location in the input image. In this study, we utilized the GRAD-CAM and attention map to control bias, retrain some models, and improve the performances of the models. This improvement might be linked to the generalization of the model to avoid overfitting and make it robust. In the history of machine learning, the need for humans to define features one by one to improve the accuracy of predictions has been major challenge. In deep learning, it is possible to automatically extract features from training data and improve accuracy. The above refining method is a “hybrid learning” method that adjusts the features learned by deep learning. However, it is possible that some of the models discarded in this study pointed to truly novel features of heart failure that have not been reported, even if the performance of the refined models was improved. 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.[29] 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 the 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. Because sensitivity-specificity is a tradeoff, a weak learner can sometimes be considered a good learner for an ensemble if it has a specific performance, even if its overall performance is not as good. 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 our daily clinical practice with all-comers, we suppose that there is some generalizability and robustness for clinical settings concerning PR curves and AUC. In the sub study predicting absolute BNP values, there was no statistical difference between the predicted and true BNP value; however, the MAE was 207 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. 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). We set the cut-off BNP value 200 pg/mL, and the models with other cut-offs such as 100 pg/mL, had similar performance to those with 200 pg/mL in our sub study.

Conclusions

The AI model can predict elevated BNP from chest X-ray images with better performance than experienced cardiologists and can improve the diagnostic performance of humans, ranging from non-experts to experienced cardiologists. These models can assist in daily practice. Further studies are necessary for distrust of innovative technologies and utilization gap.

Acknowledgements

We appreciate all the tested person in this study. They might have bad feelings about being tested.

Source of Funding

None.

Conflict of Interest

There is no conflict of interest.

References

1. Heidenreich PA, Bozkurt B, Aguilar D, Allen LA, Byun JJ, Colvin MM, Deswal A, Drazner MH, Dunlay SM, Evers LR, 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. *Circulation*. 2022;145:e895-e1032. doi: 10.1161/cir.0000000000001063

2. Stevenson LW, Perloff JK. The limited reliability of physical signs for estimating hemodynamics in chronic heart failure. *Jama*. 1989;261:884-888.

3. Chakko S, Woska D, Martinez H, de Marchena E, Futterman L, Kessler KM, Myerberg RJ. Clinical, radiographic, and hemodynamic correlations in chronic congestive heart failure: conflicting results may lead to inappropriate care. *Am J Med*. 1991;90:353-359. doi: 10.1016/0002-9343(91)80016-f

4. Maisel AS, Krishnaswamy P, Nowak RM, McCord J, Hollander JE, Duc P, Omland T, Storrow AB, Abraham WT, Wu AH, et al. Rapid measurement of B-type natriuretic peptide in the emergency diagnosis of heart failure. *N Engl J Med*. 2002;347:161-167. doi: 10.1056/NEJMoa020233

5. Mueller C, Scholer A, Laule-Kilian K, Martina B, Schindler C, Buser P, Pfisterer M, Perruchoud AP. Use of B-type natriuretic peptide in the evaluation and management of acute dyspnea. *N Engl J Med*. 2004;350:647-654. doi: 10.1056/NEJMoa031681

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. doi: 10.1136/bmjebm-2019-111208

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. doi: 10.1016/j.jacc.2020.11.030

8. Tsay D, Patterson C. From Machine Learning to Artificial Intelligence Applications in Cardiac Care. *Circulation*. 2018;138:2569-2575. doi: 10.1161/circulationaha.118.031734

9. Deo RC. Machine Learning in Medicine. *Circulation*. 2015;132:1920-1930. doi: 10.1161/circulationaha.115.001593

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

11. Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, Carter RE, Yao X, Rabinstein AA, Erickson BJ, 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. doi: 10.1016/s0140-6736(19)31721-0

12. Adedinsewo D, Carter RE, Attia Z, Johnson P, Kashou AH, Dugan JL, Albus M, Sheele JM, Bellolio F, Friedman PA, et al. Artificial Intelligence-Enabled ECG Algorithm to Identify Patients With Left Ventricular Systolic Dysfunction Presenting to the Emergency Department With Dyspnea. *Circ Arrhythm Electrophysiol*. 2020;13:e008437. doi: 10.1161/circep.120.008437

13. Gessert N, Lutz M, Heyder M, Latus S, Leistner DM, Abdelwahed YS, Schlaefer A. Automatic Plaque Detection in IVOCT Pullbacks Using Convolutional Neural Networks. *IEEE Trans Med Imaging*. 2019;38:426-434. doi: 10.1109/tmi.2018.2865659

14. Myhre PL, Prescott MF, Murphy SP, Fang JC, Mitchell GF, Ward JH, Claggett B, Desai AS, Solomon SD, Januzzi JL. 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. doi: 10.1016/j.jchf.2021.09.007

15. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. Paper/Poster presented at: 2009 IEEE conference on computer vision and pattern recognition; 2009;

16. Misra D. Mish: A self regularized non-monotonic activation function. *arXiv preprint arXiv:190808681*. 2019.

17. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:171009412*. 2017.

18. Zhong Z, Zheng L, Kang G, Li S, Yang Y. Random Erasing Data Augmentation. *arXiv preprint arXiv:170804896*. 2017.

19. Yun S, Han D, Oh SJ, Chun S, Choe J, Yoo Y. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. *arXiv e-prints*. 2019:arXiv: 1905.04899.

20. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. *arXiv preprint arXiv:151200567*. 2015.

21. Foret P, Kleiner A, Mobahi H, Neyshabur B. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:201001412*. 2020.

22. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:14126980*. 2014.

23. Liu L, Jiang H, He P, Chen W, Liu X, Gao J, Han J. On the variance of the adaptive

learning rate and beyond. *arXiv preprint arXiv:190803265*. 2019.

24. Kluyver T, Ragan-Kelley B, Pérez F, Granger BE, Bussonnier M, Frederic J, Kelley K, Hamrick JB, Grout J, Corlay S. *Jupyter Notebooks-a publishing format for reproducible computational workflows*. 2016.

25. Tsutamoto T, Wada A, Sakai H, Ishikawa C, Tanaka T, Hayashi M, Fujii M, Yamamoto T, Dohke T, Ohnishi M, et al. Relationship between renal function and plasma brain natriuretic peptide in patients with heart failure. *J Am Coll Cardiol*. 2006;47:582-586. doi: 10.1016/j.jacc.2005.10.038

26. Matsumoto T, Kodera S, Shinohara H, Ieki H, Yamaguchi T, Higashikuni Y, Kiyosue A, Ito K, Ando J, Takimoto E, et al. Diagnosing Heart Failure from Chest X-Ray Images Using Deep Learning. *Int Heart J*. 2020;61:781-786. doi: 10.1536/ihj.19-714

27. Zou XL, Ren Y, Feng DY, He XQ, Guo YF, Yang HL, Li X, Fang J, Li Q, Ye JJ, 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. doi: 10.1371/journal.pone.0236378

28. 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. doi: 10.1016/j.cjca.2021.02.007

29. 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. doi: 10.1109/3468.618255

Figure Legends

Figure 1. Overview of the study.

We designed AI models to predict elevated 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 with the featured maps, and if they seemed to focus on regions essentially unrelated to heart failure (caption of the images), their weights were discarded, and they were retrained anew. Second, we created a soft ensemble prediction model as the final model using the weak learners. The performances of the models were evaluated using the test dataset. For human testing, 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. 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.

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. Performance of the models.

ROC curves (A) and PR curves (B) of the AI models for predicting BNP >= 200 pg/ml. ROC curves (C) and PR curves (D) for predicting BNP >= 100 pg/ml. The ROC and PR curves were shown for the final ensemble model and for 3 of the 31 weak leaners. ROC indicates receiver-operating characteristics; PR, precision recall.

Figure 4. 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 5. The accuracy of the humans' testing.

A, The accuracy of the non-assisted 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 ($P < 0.001$). **B**, In the non-assisted test, the accuracy had a weak positive correlation with the duration of medical careers ($r = 0.414$, $P = 0.014$). **C**, In the AI-assisted test it had a weak negative correlation (r =

-0.347, P = 0.041). **D**, **E**, 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. **F**, 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.

Table 1. Characteristics of the Study Patients and Materials

Patient	
Study patients	1607
Male sex	980 (61)
Main diagnosis / complaints	
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)
Valid	1693 (17)
Test	1713 (17)

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	29	0.839	0.839	0.835	0.842	0.837	0.911	0.915
InceptionResNetV2 retrained	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	21	0.817	0.831	0.793	0.841	0.812	0.907	0.912
MobileNetV3Small retrained	224	5	0.809	0.876	0.717	0.900	0.789	0.914	0.918
MobileNetV3Large	224	15	0.842	0.886	0.784	0.900	0.832	0.912	0.920
MobileNetV3Large retrained	224	22	0.821	0.889	0.731	0.910	0.802	0.904	0.913
MobileNetV3Large retrained_2	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	23	0.823	0.830	0.808	0.836	0.819	0.894	0.901
ResNet-RS200 retrained	224	7	0.823	0.879	0.745	0.899	0.807	0.903	0.906
EfficientNetV2B0	224	27	0.803	0.838	0.747	0.857	0.790	0.897	0.899
EfficientNetV2B0 retrained	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	29	0.832	0.855	0.798	0.867	0.826	0.910	0.916
EfficientNetV2B2 retrained	260	0	0.821	0.801	0.852	0.791	0.826	0.907	0.912
EfficientNetV2B3	300	15	0.830	0.853	0.796	0.864	0.823	0.904	0.912
EfficientNetV2B3 retrained	300	0	0.825	0.873	0.758	0.891	0.811	0.908	0.913
EfficientNetV2S	384	23	0.827	0.870	0.765	0.887	0.814	0.905	0.914
EfficientNetV2S retrained	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	15	0.813	0.855	0.701	0.875	0.800	0.893	0.888
EfficientNetV2L retrained	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
Pre/post refined 10 models †									
Pre-refined			0.824±0.011	0.851±0.019	0.778±0.040	0.864±0.023	0.815±0.014	0.904±0.007‡	0.909±0.010
Post-refined			0.827±0.010	0.855±0.025	0.786±0.041	0.866±0.034	0.818±0.015	0.908±0.005‡	0.911±0.006
Overall †									
Pre-refined			0.817±0.042	0.844±0.063	0.789±0.060	0.843±0.124	0.812±0.026	0.905±0.017	0.909±0.021
Post-refined			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 †									
Pre-refined			0.851	0.873	0.819	0.883	0.845	0.928	0.932
Post-refined (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

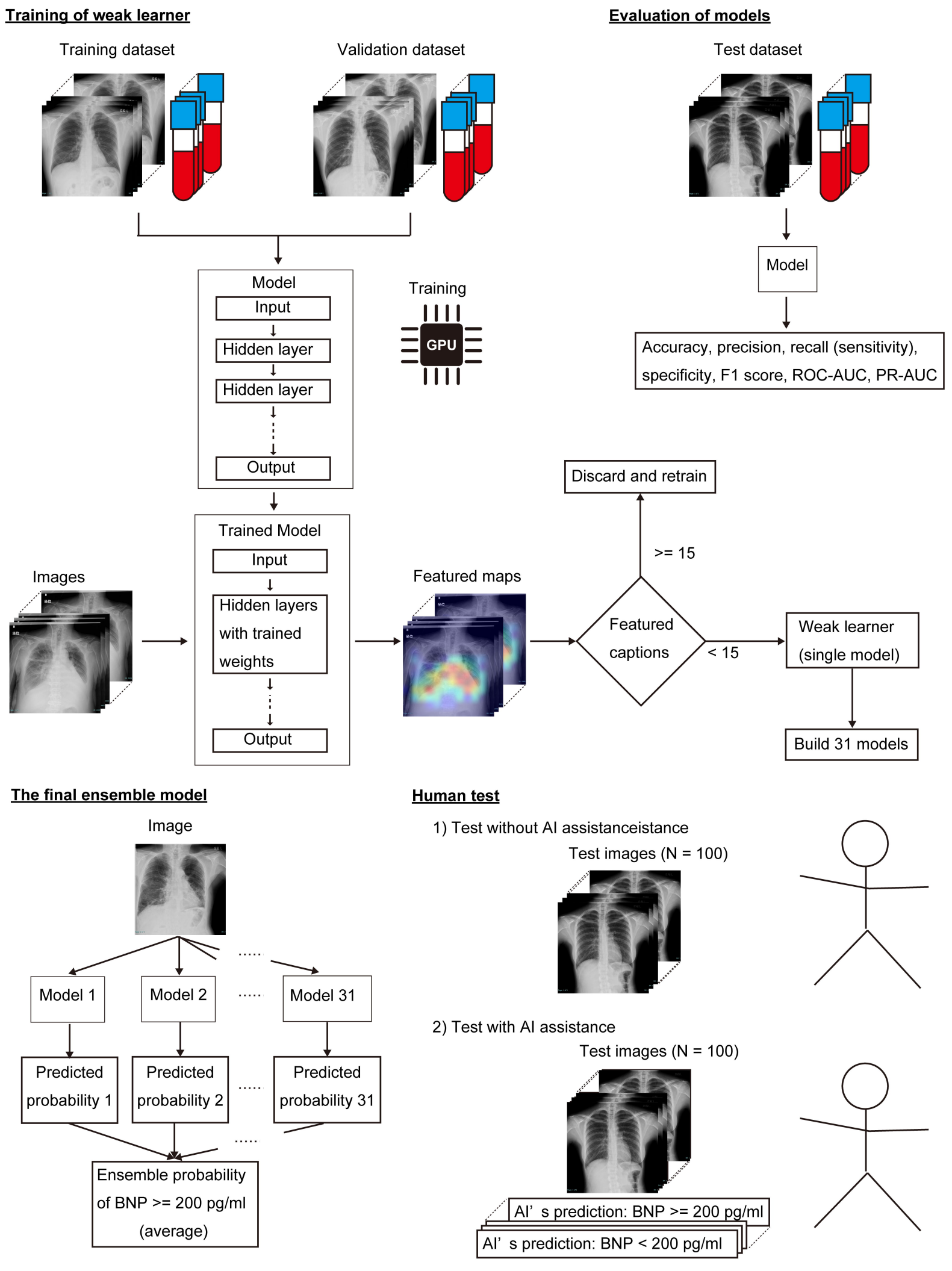
Vit-b16_2	320	9	0.822	0.816	0.929	0.636	0.869	0.899	0.938
Vit-b16	384	11	0.83	0.836	0.91	0.69	0.871	0.908	0.946
MLPMixerB32	224		0.843	0.88	0.871	0.794	0.876	0.907	0.948
MLPMixerB32_2	224		0.843	0.866	0.891	0.76	0.878	0.908	0.947
MLPMixerB32_3	224		0.838	0.851	0.903	0.725	0.876	0.906	0.945
Overall †									
			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.

†Pre / post refined means models were not refined by the featured map or those after refined.

‡Paired t-test: $P < 0.10$.


Training of weak learner
Training dataset
Validation dataset
Model
Input
Hidden layer
Hidden layer
Output
Training
GPU
Trained Model
Input
Hidden layers with trained weights
Output
Images
Featured maps
Featured captions
>= 15
Discard and retrain
< 15
Weak learner (single model)
Build 31 models
Evaluation of models
Test dataset
Model
Accuracy, precision, recall (sensitivity), specificity, F1 score, ROC-AUC, PR-AUC
The final ensemble model
Image
Model 1
Model 2
Model 31
Predicted probability 1
Predicted probability 2
Predicted probability 31
Ensemble probability of BNP >= 200 pg/ml (average)
Human test
1) Test without AI assistanceistance
Test images (N = 100)
2) Test with AI assistance
Test images (N = 100)
AI' s prediction: BNP >= 200 pg/ml
AI' s prediction: BNP < 200 pg/ml

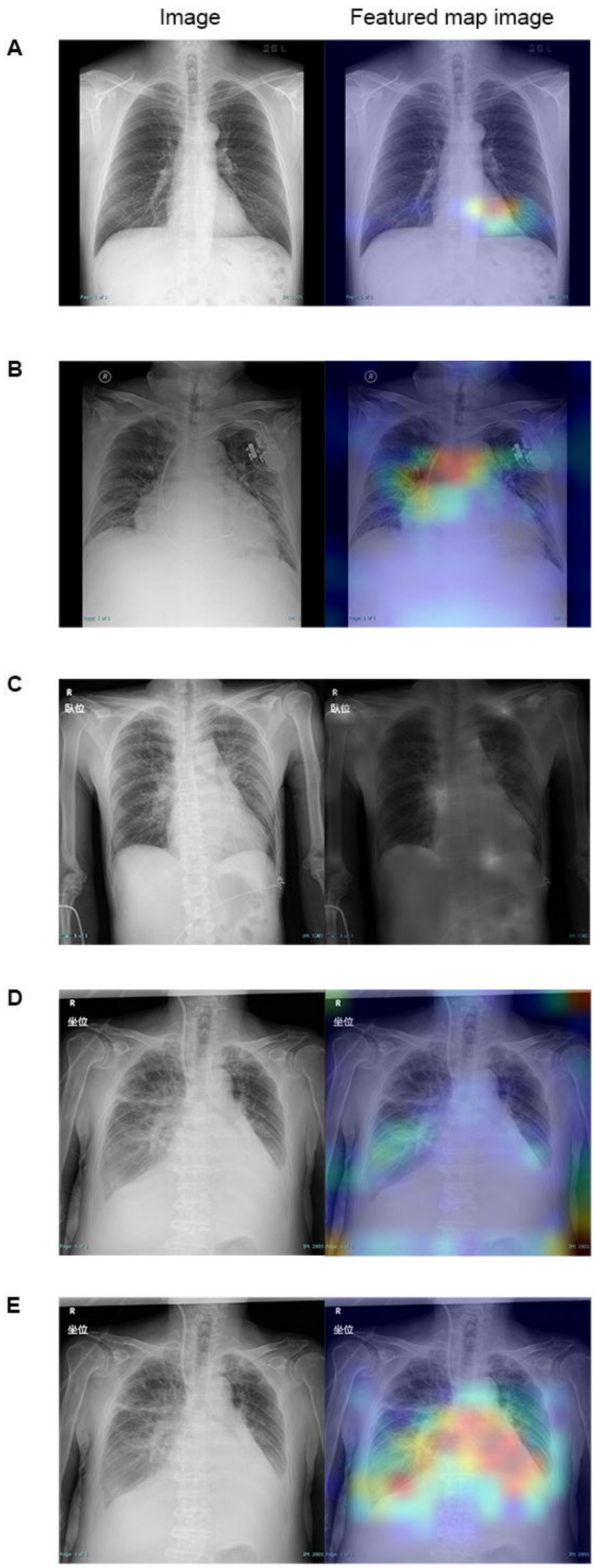


Image
Featured map image
A
B
C
R
卧位
D
R
坐位
E
R
坐位

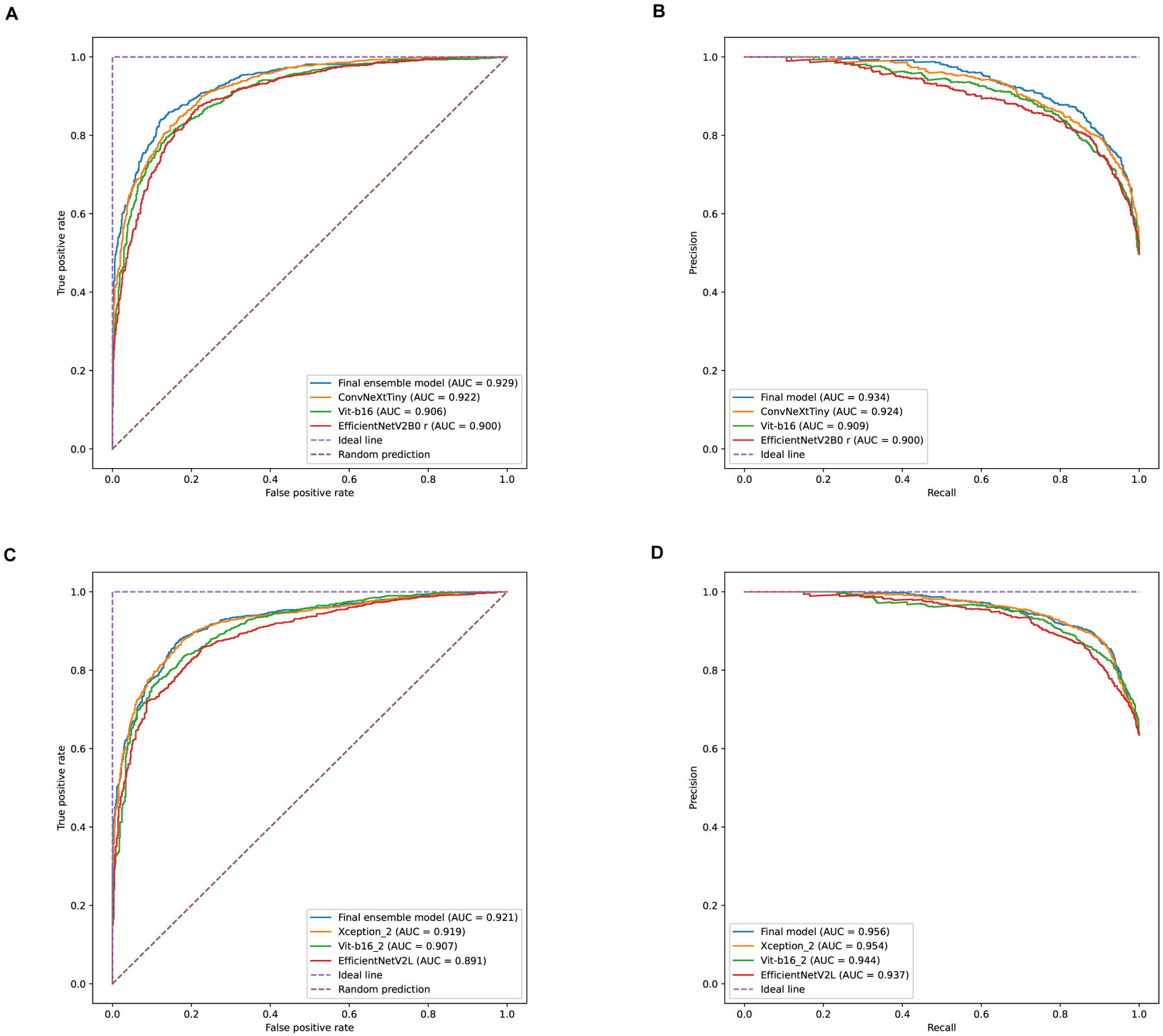
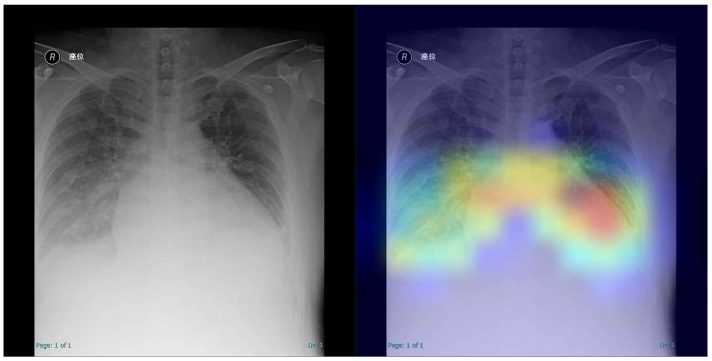

A
True positive rate
False positive rate
Final ensemble model (AUC = 0.929)
ConvNeXtTiny (AUC = 0.922)
Vit-b16 (AUC = 0.906)
EfficientNetV2B0 r (AUC = 0.900)
Ideal line
Random prediction
B
Precision
Recall
Final model (AUC = 0.934)
ConvNeXtTiny (AUC = 0.924)
Vit-b16 (AUC = 0.909)
EfficientNetV2B0 r (AUC = 0.900)
Ideal line
C
True positive rate
False positive rate
Final ensemble model (AUC = 0.921)
Xception_2 (AUC = 0.919)
Vit-b16_2 (AUC = 0.907)
EfficientNetV2L (AUC = 0.891)
Ideal line
Random prediction
D
Precision
Recall
Final model (AUC = 0.956)
Xception_2 (AUC = 0.954)
Vit-b16_2 (AUC = 0.944)
EfficientNetV2L (AUC = 0.937)
Ideal line

Image Featured map image

A



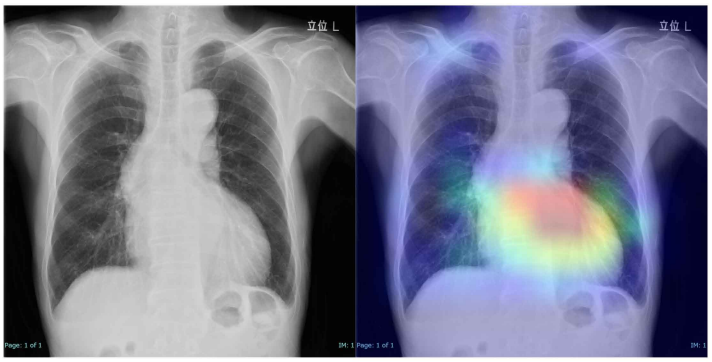
41 – 55 y.o. male

BNP 437 pg/mL

The AI prediction (BNP >= 200 pg/ml): 0.909

Humans' accuracy: 1.00

B



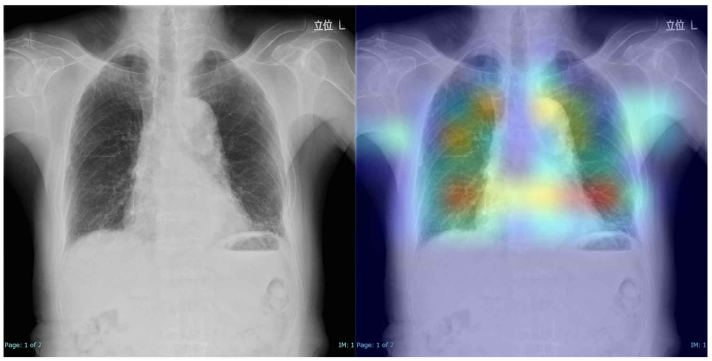
76 – 80 y.o. male

BNP 852 pg/mL

The AI prediction (BNP >= 200 pg/ml): 0.714

Humans' accuracy: 0.32

C



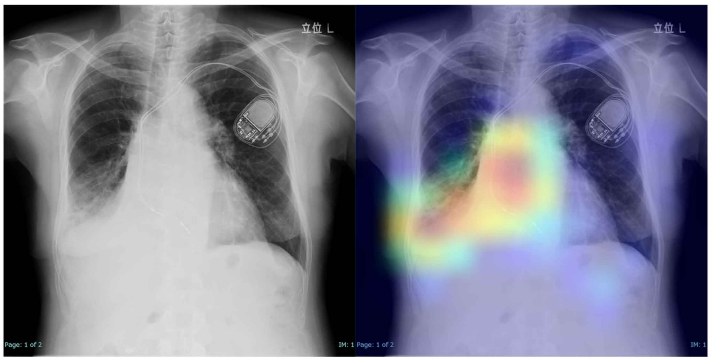
76 – 80 y.o. male

BNP 281 pg/mL

The AI prediction (BNP >= 200 pg/ml): 0.346

Humans' accuracy: 0.50

D



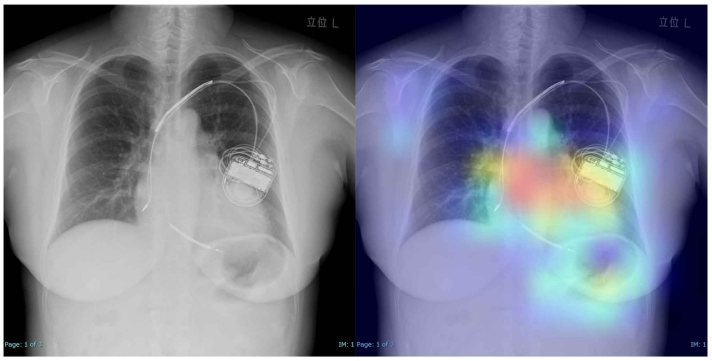
76 – 80 y.o. male

BNP 171 pg/mL

The AI prediction (BNP >= 200 pg/ml): 0.556

Humans' accuracy: 0.12

E

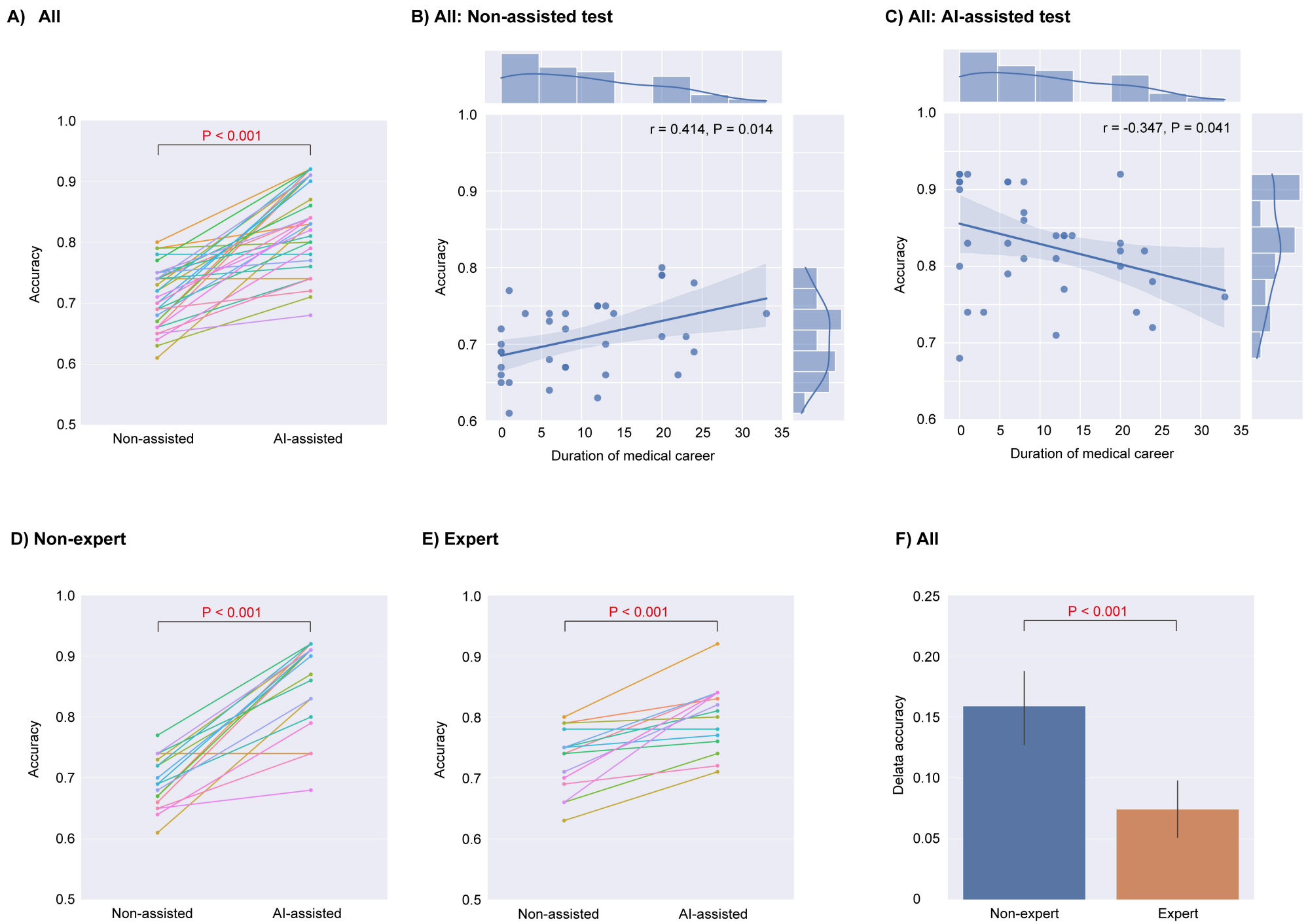


51 – 55 y.o. female

BNP 67 pg/mL

The AI prediction (BNP >= 200 pg/ml): 0.172

Humans' accuracy: 0.35


A) All
P < 0.001
Accuracy
1.0
0.9
0.8
0.7
0.6
0.5
Non-assisted
AI-assisted
B) All: Non-assisted test
r = 0.414, P = 0.014
Accuracy
Duration of medical career
0
5
10
15
20
25
30
35
C) All: AI-assisted test
r = -0.347, P = 0.041
Accuracy
Duration of medical career
D) Non-expert
P < 0.001
Accuracy
Non-assisted
AI-assisted
E) Expert
P < 0.001
Accuracy
Non-assisted
AI-assisted
F) All
P < 0.001
Delata accuracy
0.25
0.20
0.15
0.10
0.05
0
Non-expert
Expert