The Effectiveness of a Deep Learning Model to Detect Left Ventricular Systolic Dysfunction from Electrocardiograms

Susumu Katsushika,1 MD, Satoshi Kodera,1 MD, Mitsuhiko Nakamoto,2 MSc, Kota Ninomiya,1 MSc, Shunsuke Inoue,1 MD, Shinnosuke Sawano,1 MD, Nobutaka Kakuda,1 MD, Hiroshi Takiguchi,1 MD, Hiroyuki Shinohara,1 MD, Ryo Matsuoka,1 MD, Hirotaka Ieki,1 MD, Yasutomi Higashikuni,1 MD, Koki Nakanishi,1 MD, Tomoko Nakao,1,2 MD, Tomohisa Seki,1 MD, Norifumi Takeda,1 MD, Katsuhiro Fujiu,1,4 MD, Masao Daimon,1,2 MD, Hiroshi Akazawa,1 MD, Hiroyuki Morita,1 MD and Issei Komuro,1 MD

Summary

Deep learning models can be applied to electrocardiograms (ECGs) to detect left ventricular (LV) dysfunction. We hypothesized that applying a deep learning model may improve the diagnostic accuracy of cardiologists in predicting LV dysfunction from ECGs. We acquired 37,103 paired ECG and echocardiography data records of patients who underwent echocardiography between January 2015 and December 2019. We trained a convolutional neural network to identify the data records of patients with LV dysfunction (ejection fraction < 40%) using a dataset of 23,801 ECGs. When tested on an independent set of 7,196 ECGs, we found the area under the receiver operating characteristic curve was 0.945 (95% confidence interval: 0.936–0.954). When 7 cardiologists interpreted 50 randomly selected ECGs from the test dataset of 7,196 ECGs, their accuracy for predicting LV dysfunction was 78.0% ± 6.0%. By referring to the model’s output, the cardiologist accuracy improved to 88.0% ± 3.7%, which indicates that model support significantly improved the cardiologist diagnostic accuracy (P = 0.02). A sensitivity map demonstrated that the model focused on the QRS complex when detecting LV dysfunction on ECGs. We developed a deep learning model that can detect LV dysfunction on ECGs with high accuracy. Furthermore, we demonstrated that support from a deep learning model can help cardiologists to identify LV dysfunction on ECGs.

Key words: Echocardiography, Artificial intelligence

Asymptomatic left ventricular (LV) systolic dysfunction exists in 3%-6% of the general population.3 Asymptomatic LV systolic dysfunction is related to increasing incidence of symptomatic heart failure and elevated mortality rate.5 In patients with asymptomatic LV systolic dysfunction who were included in the Framingham Heart Study, the annual incidence of symptomatic heart failure was 6% and the mortality rate was 8%.5 In these patients, medications (angiotensin-converting enzyme inhibitors, angiotensin receptor blockers, and beta blockers) and therapeutic device therapies (implantable cardioverter-defibrillators and cardiac resynchronization systems) were effective in preventing the progression of heart failure, which reduced their mortality.5,41 Therefore, early diagnosis and treatment of asymptomatic LV systolic dysfunction are essential.

However, the available screening tools for asymptomatic LV systolic dysfunction are limited. Currently, measuring plasma levels of B-type natriuretic peptide7 is the best screening tool for asymptomatic LV systolic dysfunction, with area under the receiver operating characteristic curve (AUROC) values from 0.82 to 0.92;8 however, blood tests are invasive, and results are not immediately available. In addition, in clinical practice, this blood test is only performed in a limited number of patients in whom heart disease is suspected. Additional screening tools with good performance should be developed.

Machine learning is a method of algorithmic data analysis in which a computer learns features contained in the data. Recently, the machine learning method of deep learning has achieved state-of-the-art performance in several medical fields.9,10 Deep learning algorithms have been developed for use on electrocardiograms (ECGs) to predict the incidence of atrial fibrillation14 and the presence of mitral regurgitation15 and aortic stenosis.16 In each study, tens of thousands of raw 12-lead ECG data were
used to train a model based on a 2-dimensional convolutional neural network, and the results demonstrated highly accurate predictability. In addition, Attia, et al. reported a deep learning model that immediately detects LV dysfunction, which they defined as an ejection fraction of less than 35%, which had high accuracy. Their model was trained on 35,970 ECGs, although the dataset and their model are not available. We validated whether applying deep learning to ECG data could predict LV systolic dysfunction with high accuracy. We hypothesized that a deep learning model may improve the diagnostic accuracy of the cardiologists in detecting LV systolic dysfunction from ECGs.

**Methods**

**Study sample:** We included patients aged 18 years or older who underwent echocardiography at The University of Tokyo Hospital between January 2015 and December 2019 and had an ECG performed within 28 days of their echocardiography. The ECG and matched echocardiogram were paired (one-to-one correspondence between the ECG and the echocardiogram). When a single ECG corresponded to multiple echocardiograms, we selected one echocardiogram according to the following criteria: 1) the latest echocardiogram performed after the ECG, or 2) the latest echocardiography performed before that ECG if one was not performed after. Consequently, we obtained a dataset of 37,285 pairs (ECG and echocardiogram) from 18,954 patients. We then excluded 182 ECG-echocardiogram pairs for which the LV ejection fraction had not been recorded. We then randomly divided the remaining 37,103 pairs into the training dataset (23,801 pairs from 12,108 patients: 64.1%), the validation dataset (6,106 pairs from 3,010 patients: 16.5%), and the test dataset (7,196 pairs from 3,784 patients: 19.4%). Patients with multiple pairs were included in the same dataset (Figure 1). The study was conducted in accordance with the revised Declaration of Helsinki and approved by the Institutional Review Board of the University of Tokyo (reference number 2020312NI). Informed consent was obtained in the form of an opt-out selection on the web-site.

**Electrocardiography and echocardiography data acquisition:** ECGs were acquired at a sampling rate of 500 Hz (FCP-8700/FCP-8800; Fukuda Denshi, Tokyo, Japan). The voltage waveform was recorded for a 10-second interval and exported as a comma-separated value (.CSV) file. Echocardiography was performed by skilled ultrasound sonographers or cardiologists (Vivid 7/Vivid E9/Vivid E95, GE Healthcare, Waukesha, WI, USA; iE33/EPIQ7, Philips, Amsterdam, The Netherlands; Xario/Artida/Aplio 300/Aplio XV, Toshiba, Tokyo, Japan; or Acuson SC2000, Siemens, Munich, Germany). Echocardiographic measurements were obtained according to the American Society of Echocardiography recommendations at the time of acquisition, and each echocardiogram was interpreted by one or two experienced echocardiologists. The measurements were exported as a .csv file.

Ejection fraction is typically measured using either the modified Simpson method or a 2-D method. If the ejection fraction was measured by both methods, the value measured by the modified Simpson method was used in this study. LV dysfunction was defined as an ejection fraction of less than 40%. ECG-echocardiogram pairs that indicated LV dysfunction were labeled 1, and the rest were labeled 0.

**Model development:** ECGs comprise two types of information: spatial and temporal. In particular, the spatial in-

---

**Figure 1.** Training, validation, and test datasets. Flow chart showing how the 3 datasets used for model training and evaluation were created. To avoid cross-contamination among the training, validation, and test datasets, multiple data pairs from a single patient were included in the same dataset. To prevent overfitting and bias from imbalanced datasets, the training and validation datasets were augmented at different rates for each label.
formation differs between each ECG lead. For example, in determining where an acute myocardial infarction has occurred, significant ST changes in leads II, III, and aVF may reflect an inferior wall infarction, whereas significant ST changes in leads I, aVL, V5, and V6 may indicate a lateral wall infarction. Therefore, to make effective use of the spatial information, before inputting the ECG waveform information into the model, the leads were rearranged as follows: II, III, aVF, I, aVL, V6, aVR, V1, V2, V3, V4, and V5, respectively (Figure 2).

Based on the method from a previous study, we developed a deep learning model to detect LV dysfunction from ECGs using convolutional neural network architecture with 7 convolutional blocks followed by 2 fully connected layers. The first 6 convolutional blocks were called temporal blocks: they were designed to learn temporal features within each lead. After the last temporal block, a spatial convolutional block was used to fuse the data from all leads and learn the features into one convolutional layer. We called this Model A (Figure 2). However, ECG spatial information may be partially neglected using Model A because the data from all leads are fused into one convolutional layer. Therefore, we modified the spatial block into which the ECG waveform information from each of 3 leads was convolved and called the result Model B (Figure 2). Furthermore, the average value of the outputs of the 2 models was calculated as Model ensemble (Figure 2).

Before training the models, the training and validation datasets were augmented. We extracted multiple intervals of waveform information shaped 4,000 × 12 in regular strides from a single interval of ECG waveform information of shape 5,000 × 12. In addition, because training a model with imbalanced data is likely to bias the results, the distribution of labels contained in both datasets was equalized by adjusting the cutting stride for each label (Figure 1).

The ECG data from the augmented training dataset were inputted into Models A and B. Both models were trained to minimize binary cross-entropy loss between the models’ predictions and ground truth labels using the Adam optimizer, which is an effective variant of stochastic gradient descent optimization, with a batch size of 256 for 100 epochs. The initial learning rate was selected from among 0.00001, 0.000006, 0.000003, 0.000001, or 0.0000005 (we selected the initial learning rate with the highest accuracy at the time of inference). The learning rate was reduced by a factor of 2 if the loss plateaued after 3 epochs. If the loss did not decrease for 20 consecutive epochs, model training was stopped even if 100 epochs had not been completed, and the model’s weights at the lowest loss value were saved. The trained models were applied to the first 8 seconds of each ECG interval included in the test dataset. The predicted value (i.e., the probability of the presence of LV dysfunction) was calculated for each model. Models were developed in Python using the PyTorch deep learning library and an Nvidia Tesla V-100 32 GB graphics processing unit.

Model performance evaluation: We evaluated the diagnostic performance of the 3 models - Model A, Model B, and Model ensemble - on the test dataset to calculate the accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) when the cutoff value was set to 0.5, as well as the AUROC and area under the precision-recall curve (AUPRC). We also calculated the AUPRC because the AUROC can be misleading when evaluating models trained with imbalanced data.

We also conducted an interpretation test to evaluate the ability of Model ensemble to support our cardiologists’ interpretations in predicting LV dysfunction from an ECG. Seven board-certified cardiologists, each with over 7 years of clinical experience, participated in the interpretation test. Each cardiologist independently classified 50 randomly selected ECGs from the test dataset, 10% of which exhibited LV dysfunction, into normal LV function (labeled 0) or LV dysfunction (labeled 1). Judgments during ECG classification were made based on the impression of each cardiologist because there are no established ECG criteria for LV dysfunction. Second, we informed them of the AUROC and predicted values of Model ensemble. Then, they classified the same randomly sorted ECGs as either normal LV function (labeled 0) or LV dysfunction (labeled 1) with reference to the predicted values of Model ensemble (i.e., model support). We did not share the information about the actual prevalence of LV dysfunction in this dataset. We compared the accuracy, sensitivity, specificity of each classification test. Furthermore, to visualize which regions affected the interpretations of the developed model, we used the gradient-weighted class activation mapping (Grad-CAM) method.

Statistical analysis: Continuous variables are presented as the mean and standard deviation and were compared using 2-tailed analyses of variance. Categorical variables are expressed as frequencies and percentages and were compared using Chi-squared tests. We evaluated the 95% confidence intervals (CIs) of accuracy, sensitivity, specificity, PPV, and NPV using bootstrapping (resampling 10,000 times with replacement). To evaluate the diagnostic performance of the cardiologists’ interpretations with and without model support, we used the Obuchowski method, which extends the McNemar test to the case in which the observations are sampled in clusters. We used the DeLong method to compare the models’ AUROC values. Statistical analysis was performed using R, and statistical significance was defined as a P-value of < 0.05.

Results

Patient characteristics: This study included 37,103 ECG-echocardiography pairs (37,103 ECGs paired with 28,802 echocardiograms) from 18,919 patients. Among the 28,802 echocardiograms included in this study, the ejection fraction value was measured by the modified Simpson method on 7,550 echocardiograms, whereas the 2-D method was used on 28,050 echocardiograms. The median period between the acquisition of the paired ECG and echocardiography was 1 day. In total, 7,607 patients had multiple ECG and echocardiography pairs. The characteristics of the patients included in this study are shown in Table I. The mean age of the study population was 63.4 ± 16.9 years. There were 21,025 ECGs from 10,403 male participants and 16,078 ECGs from 8,516 female partici-
Figure 2. Selected model architectures. Summary of 3 model architectures (Model A, Model B, and Model ensemble). Models A and B each comprised 6 temporal blocks, a spatial block, and 2 fully connected layers. Model ensemble calculated the average value of the outputs of Models A and B.

We enrolled 3,501 ECGs from 1,116 patients with LV dysfunction in this study. The training, validation, and test datasets included 23,801 ECGs from 12,108 patients (64.1%); 6,106 ECGs from 3,027 patients (16.5%); and 7,196 ECGs from 3,784 patients (19.4%), respectively (Figure 1, Table I). The distribution of LV dysfunction in each dataset is also shown in Table I.

Diagnostic performance: The diagnostic performance
values of Model A, Model B, and Model ensemble on the test dataset with a cut-off value of 0.5 are shown in Table II. The accuracies of Models A, B, and ensemble were 88.9% (95% CI: 88.1%-89.6%), 89.6% (95% CI: 88.8%-90.3%), and 89.6% (95% CI: 88.8%-90.3%), respectively. The specificities of Models A, B, and ensemble were 86.4% (95% CI: 83.8%-88.8%), 84.1% (95% CI: 81.3%-86.7%), and 85.8% (95% CI: 83.0%-88.2%), respectively. The specificities of Models A, B, and ensemble were 89.2% (95% CI: 88.4%-89.9%), 90.2% (95% CI: 89.4%-90.9%), and 90.0% (95% CI: 89.2%-90.7%), respectively. The AUROC values of the 3 models are shown in Figure 3A. There was no significant difference between the AUROC values of Models A and B (0.942, 95% CI: 0.933-0.951 versus 0.941, 95% CI: 0.932-0.950, respectively; P = 0.54). However, the AUROC of Model ensemble (0.945, 95% CI: 0.936-0.954) was significantly higher than those of Models A (0.942, 95% CI: 0.933-0.951; P < 0.01) and B (0.941, 95% CI: 0.932-0.950; P < 0.01). The AUPRC values of Models A, B, and ensemble were 0.716 (95% CI: 0.694-0.738), 0.724 (95% CI: 0.703-0.746), and 0.740 (95% CI: 0.719-0.761), respectively (Figure 3B).

**Ability of Model ensemble to assist in predicting LV dysfunction:** The diagnostic performance values of the cardiologists’ interpretations with and without model support (by Model ensemble) of about 50 randomly selected ECGs from the test dataset are shown in Table III. The accuracy, sensitivity, and specificity of the cardiologists’ interpretations were 78.0% ± 6.0%, 51.4% ± 18.1%, and 81.0% ± 7.7%, respectively. The accuracy, sensitivity, and specificity of the cardiologists’ interpretations with model support were 88.0% ± 3.7%, 74.3% ± 9.0%, and 89.5% ± 3.9%, respectively. Each diagnostic performance value was improved with support from Model ensemble (Figure 4). In particular, there were significant improvements in accuracy and specificity (Table III; P = 0.02 and P = 0.04, respectively).

**Visualizing the model’s decision making:** The Grad-CAM method was used to depict the focus of the developed model in interpreting ECGs. Both Models A and B focused on the QRS complex in leads V1-V6 to detect LV dysfunction from ECG data (Figure 5).

**Discussion**

We developed a deep learning model to detect LV dysfunction from ECGs with high accuracy. Furthermore, we showed that cardiologists had improved diagnostic accuracy in predicting LV dysfunction from ECG when referring to our model (i.e., receiving model support). Our model appears to focus on the QRS complex in determining whether or not the ejection fraction is reduced.

Our model’s AUROC, accuracy, sensitivity, and specificity on the test dataset were 0.945, 89.6%, 85.8%, and 90.0%, respectively. In a previous study, Attia, et al demonstrated that the application of deep learning to ECGs could be useful as a screening tool by developing a deep learning model that predicted LV dysfunction, defined as an ejection fraction of less than 35%, from ECGs with high accuracy. Their model’s AUROC, accuracy, sen-

---

**Table I.** Patient Characteristics at the Time of ECG and Echocardiogram Acquisition

|                          | Training dataset (n = 23,801) | Validation dataset (n = 6,106) | Test dataset (n = 7,196) | P-value  |
|--------------------------|-------------------------------|-------------------------------|--------------------------|----------|
| Age, years               | 63.3 (17.0)                  | 63.5 (16.7)                  | 63.5 (16.7)              | 0.45     |
| Age groups               |                               |                               |                          | 0.09     |
| < 40                     | 2,760 (11.6%)                | 657 (10.8%)                  | 748 (10.4%)              |          |
| 40-49                    | 2,358 (9.9%)                | 619 (10.1%)                  | 781 (10.9%)              |          |
| 50-59                    | 3,188 (13.4%)                | 835 (13.7%)                  | 995 (13.8%)              |          |
| 60-69                    | 5,067 (21.3%)                | 1,318 (21.6%)                | 1,484 (20.6%)            |          |
| 70-79                    | 6,562 (27.6%)                | 1,670 (27.4%)                | 2,005 (27.9%)            |          |
| 80+                      | 3,866 (16.2%)                | 1,007 (16.5%)                | 1,183 (16.4%)            |          |
| Sex                      |                               |                               |                          | 0.22     |
| Female                   | 10,385 (43.6%)               | 2,636 (43.2%)                | 3,057 (42.5%)            |          |
| Male                     | 13,416 (56.4%)               | 3,470 (56.8%)                | 4,139 (57.5%)            |          |
| Body Height, cm          | 161.7 (9.8)                  | 161.8 (10.0)                 | 161.8 (9.9)              | 0.91     |
| Body Weight, kg          | 60.0 (13.6)                  | 59.8 (13.4)                  | 60.3 (13.6)              | 0.11     |
| Mean EF, %               | 61.4 (14.6)                  | 60.5 (15.4)                  | 60.9 (15.2)              | < 0.01   |
| LV dysfunction           | 2,153 (9.0%)                 | 611 (10.0%)                  | 737 (10.2%)              | < 0.01   |

Data are presented as mean (standard deviation). P-values are from two-tailed analysis of variance or the Chi-squared test for differences in the distribution of values between the 3 datasets.

**Table II.** Models’ Diagnostic Performance on the Test Dataset

| Model name       | Accuracy   | Sensitivity | Specificity | PPV        | NPV        |
|------------------|------------|-------------|-------------|------------|------------|
| Model A          | 88.9% (88.1-89.6) | 86.4% (83.8-88.8) | 89.2% (88.4-89.9) | 47.6% (45.8-49.5) | 98.3% (98.0-98.6) |
| Model B          | 89.6% (88.8-90.3) | 84.1% (81.3-86.7) | 90.2% (89.4-90.9) | 49.4% (47.4-51.4) | 98.0% (97.7-98.3) |
| Model ensemble   | 89.6% (88.8-90.3) | 85.8% (83.0-88.2) | 90.0% (89.2-90.7) | 49.4% (47.5-51.4) | 98.2% (97.9-98.5) |

Data are presented as percentage (95% confidence interval). PPV indicates positive predictive value; and NPV, negative predictive value.
Figure 3. Receiver operating characteristic curve and precision-recall curve of the 3 models for the test dataset. A: Receiver operating characteristic curves (95% confidence intervals) of the 3 models for the test dataset. B: Precision-recall curves (95% confidence intervals) of the 3 models for the test dataset. Red, blue, and yellow lines represent Models A, B, and ensemble, respectively. AUROC indicates area under the receiver operating characteristic curve; and AUPRC, area under the precision-recall curve. *Interaction between the AUROC values of the Models A and ensemble, $P < 0.01$. **Interaction between the AUROC values of Models B and ensemble, $P < 0.01$.

However, Attia, et al. did not verify whether their model could be used in clinical practice. We investigated whether our model could help cardiologists interpret ECGs. The accuracy and specificity of cardiologist interpretations were significantly improved by referring to the developed model's output. Furthermore, their sensitivity also improved, though not significantly. Therefore, our model may contribute to improving the diagnostic accu-

sitivity, and specificity on an internal validation dataset were 0.93, 85.7%, 86.3%, and 85.7%, respectively. Our model’s diagnostic accuracy was comparable to that of their model.

However, Attia, et al. did not verify whether their model could be used in clinical practice. We investigated whether our model could help cardiologists interpret ECGs. The accuracy and specificity of cardiologist interpretations were significantly improved by referring to the developed model’s output. Furthermore, their sensitivity also improved, though not significantly. Therefore, our model may contribute to improving the diagnostic accu-
Table III. Cardiologist Diagnostic Performance with or Without Model Support

| Cardiologist | Interpretation Accuracy | Sensitivity | Specificity | Supported by Model | Accuray | Sensitivity | Specificity |
|--------------|--------------------------|-------------|-------------|--------------------|----------|-------------|-------------|
| 1            | 82.0%                    | 40.0%       | 86.7%       | 90.0%              | 80.0%    | 91.1%       |
| 2            | 82.0%                    | 20.0%       | 88.9%       | 96.0%              | 80.0%    | 97.8%       |
| 3            | 74.0%                    | 40.0%       | 77.8%       | 86.0%              | 60.0%    | 88.9%       |
| 4            | 70.0%                    | 80.0%       | 68.9%       | 86.0%              | 80.0%    | 86.7%       |
| 5            | 88.0%                    | 60.0%       | 91.1%       | 88.0%              | 80.0%    | 88.9%       |
| 6            | 72.0%                    | 60.0%       | 73.3%       | 84.0%              | 80.0%    | 84.4%       |
| 7            | 78.0%                    | 60.0%       | 80.0%       | 86.0%              | 60.0%    | 88.9%       |

Mean ± SD: 78.0 ± 6.0%*  51.4 ± 18.1%  81.0 ± 7.7%**  88.0 ± 3.7%*  74.3 ± 9.0%  89.5 ± 3.9%**

Cardiologist diagnostic performance with or without model support (by Model ensemble) are shown as percentages. The mean value and standard deviation (SD) are also shown. *Interaction between the accuracy of cardiologist interpretations with or without model support, \( P = 0.02 \). ** Interaction between the specificity of cardiologist interpretations with or without model support, \( P = 0.04 \).

Cardiologist accuracy of LV dysfunction from ECG in clinical practice.

The strengths of this interpretation test are as follows. First, the ejection fraction values used in this study were evaluated by multiple experienced echocardiologists. Accordingly, the estimates are likely reliable. Second, depending on the data selected for training and testing, it was possible that data leakage could result in our model having higher diagnostic accuracy. Therefore, when the ECG-echocardiogram pairs were divided into 3 datasets, patients with multiple pairs were included in the same dataset. Third, information about the prevalence of LV dysfunction and the accuracy of Model ensemble were a factor that could affect the accuracy of the interpretations by the cardiologists. Given the clinical practice in which Model ensemble would be deployed, it is necessary to determine the accuracy of the model used for interpretation, although the prevalence of LV dysfunction cannot be known exactly. Therefore, we did not provide information about the prevalence of LV dysfunction to the cardiologists, but we provided information about the diagnostic accuracy of Model ensemble. Fourth, the reliability of the cardiologists’ interpretation accuracy was high because of their experience and expertise in interpreting ECGs. When inexperienced doctors or experts from other disciplines interpret ECGs, referring to the model’s output for support may yield a greater improvement in interpretation accuracy.

In addition, by using the Grad-CAM method, we found that the developed model likely focused on the QRS complex in the precordial leads to predict LV dysfunction. Previous studies have identified ECG predictors of LV dysfunction as LV hypertrophy, left bundle branch...
block, prolongation of QRS duration, axial deviation, and R-peak delay, which was defined as the R-peak time in lead V5 or V6 exceeding the S-peak time in leads V1, V2, and V3. The metabolic or structural changes associated with cardiomyopathic processes, such as cardiomyocyte hypertrophy, inflammation, and fibrosis, would result in ECG changes. For example, prolongation of QRS duration and R-peak delays reflect ventricular depolarization abnormalities arising from conduction delays that are the result of increased LV end-diastolic volume. These changes may have been detected by the developed model as changes in the QRS complex on ECG.

This study has several limitations. First, because the ECG-echocardiography data pairs were not acquired simultaneously, there was a slight temporal delay between the components of the paired data. However, this temporal delay was small, as both assessments were obtained within a few days for most pairs. Second, we did not clarify which QRS complex features are important to the model in diagnosing LV dysfunction, although the Grad-CAM images for Models A and B show the areas of highest focus for each model.
CAM method showed that the developed model focused on the QRS complex. To solve this problem inherent to the nature of deep learning, which is often called "the black box of artificial intelligence", technology that can explain the criteria of the deep learning model in more detail is required. Third, the analysis in this study was performed using ECG waveform information obtained from a single type of ECG machine. Therefore, to verify the generalizability of this study’s results, it is necessary to use ECG waveform information acquired from other ECG machines. Finally, this was a single-center retrospective study, which limits the generalizability of the findings. Further study using external data is needed.

Conclusion

We developed a deep learning model that can detect LV dysfunction (defined as an ejection fraction of less than 40%) from an ECG data with high accuracy. Furthermore, we demonstrated that applying a deep learning model to ECG data was useful in helping cardiologists identify LV dysfunction from ECGs.

Acknowledgments

We thank Nisei Kimura and Gyo Uchida for supporting deep learning model development, and Coren Walters-Stewart, PhD, and Richard Lipkin, PhD, from Edanz Group (https://en-author-services.edanz.com/ac) for editing drafts of this manuscript.

Disclosure

Conflicts of interest: The authors have no financial relationships relevant to this article to disclose.

Data availability statement: The deidentified data of participants will not be shared.

References

1. Wang TJ, Levy D, Benjamin EJ, Vasan RS. The epidemiology of “asymptomatic” left ventricular systolic dysfunction: implications for screening. Ann Intern Med 2003; 138: 907-16.
2. Wang TJ, Evans JC, Benjamin EJ, Levy D, LeRoy EC, Vasan RS. Natural history of asymptomatic left ventricular systolic dysfunction in the community. Circulation 2003; 108: 977-82.
3. Pfeffer MA, Braunwald E, Moyé LA, et al. Effect of captopril on mortality and morbidity in patients with left ventricular dysfunction after myocardial infarction. Results of the survival and ventricular enlargement trial. The SAVE Investigators. N Engl J Med 1992; 327: 669-77.
4. Dargie HJ. Effect of carvedilol on outcome after myocardial infarction in patients with left-ventricular dysfunction: the CAPRICORN randomised trial. Lancet 2001; 357: 1385-90.
5. Colucci WS, Kolas TJ, Adams KF, et al; REVERT Study Group. Metoprolol reverses left ventricular remodeling in patients with asymptomatic systolic dysfunction: the REVersal of VEntricular Remodeling with Toprol-XL (REVERT) trial. Circulation 2007; 116: 49-56.
6. Priori SG, Blomstrom-Lundqvist C, Mazzanti A, et al; ESC Scientific Document Group. 2015 ESC Guidelines for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death: The Task Force for the Management of Patients with Ventricular Arrhythmias and the Prevention of Sudden Cardiac Death of the European Society of Cardiology (ESC). Endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC). Eur Heart J 2015; 36: 2793-867.
7. McDonagh TA, McDonald K, Maisel AS. Screening for asymptomatic left ventricular dysfunction using B-type natriuretic Peptide. Congest Heart Fail 2008; 14: 5-8.
8. Redfield MM, Rodeheffer RJ, Jacobsen SJ, Mahoney DW, Bailey KR, Burnett JC Jr. Plasma brain natriuretic peptide to detect preclinical ventricular systolic or diastolic dysfunction: a community-based study. Circulation 2004; 109: 3176-81.
9. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015; 521: 436-44.
10. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016; 316: 2402-10.
11. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017; 542: 115-8.
12. Chan HP, Samala RK, Hadjiiski LM, Zhou C. Deep learning in medical image analysis. Adv Exp Med Biol 2020; 1213: 3-21.
13. Benincasa G, Marfella R, Della Mura N, Schiano C, Napoli C. Strengths and opportunities of network medicine in cardiovascular diseases. Circ J 2020; 84: 144-52.
14. 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: 861-7.
15. Kwon JM, Kim KH, Akkus Z, Jeon KH, Park J, Oh BH. Artificial intelligence for detecting mitral regurgitation using electrocardiography. J Electrocardiol 2020; 59: 151-7.
16. Kwon JM, Lee SY, Jeon KH, et al. Deep learning-based algorithm for detecting aortic stenosis using electrocardiography. J Am Heart Assoc 2020; 9: e014717.
17. Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. Nat Med 2019; 25: 70-4.
18. Lang RM, Badano LP, Mor-Avi V, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. J Am Soc Echocardiogr 2015; 28: 1-39.e14.
19. Bloom MW, Greenberg B, Jaarsma T, et al. Heart failure with reduced ejection fraction. Nat Rev Dis Primers 2017; 3: 17058.
20. Sgarbossa EB, Birnbaum Y, Parrillo JE. Electrocardiographic diagnosis of acute myocardial infarction: Current concepts for the clinician. Am Heart J 2001; 141: 507-17.
21. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations. San Diego, CA, USA: May 7-9, 2015.
22. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015; 10: e0118432.
23. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Proc IEEE Int Conf Comput Vis 2017: 1: 618-26.
24. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat Med 2000; 19: 1141-64.
25. Obuchowski NA. On the comparison of correlated proportions for clustered data. Stat Med 1998; 17: 1495-507.
26. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biométrics 1988; 44: 837-45.
27. Ho JE, Lyass A, Lee DS, et al. Predictors of new-onset heart failure: differences in preserved versus reduced ejection fraction.
28. O’Neal WT, Mazur M, Bertoni AG, et al. Electrocardiographic predictors of heart failure with reduced versus preserved ejection fraction: The multi-ethnic study of atherosclerosis. J Am Heart Assoc 2017; 6: e006023.

29. Recke SH. Left ventricular function-conduction impairment as reflected by the ECG in chronic aortic regurgitation. Wien Klin Wochenschr 2011; 123: 502-7.

30. Recke S, Gansser R, Marienhagen J, et al. R peak time prolongation and R peak delay in leads I, V5, or V6. Diagnostic values as signs of myocardial dysfunction in chronic mitral incompetence. J Electrocardiol 1994; 27: 129-36.

31. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov 2019; 9: e1312.