EXPERIMENTAL STUDY

Automatic Detection of Left Ventricular Dilatation and Hypertrophy from Electrocardiograms Using Deep Learning

Takahiro Kokubo,1 MPH, Satoshi Kodera,2 MD, Shimnosuke Sawano,2 MD, Susumu Katsushika,3 MD, Mitsuhiko Nakamoto,2 MD, Hiroshi Takeuchi,3 MD, Nisei Kimura,1 MD, Hiroki Shinohara,2 MD, Ryo Matsuoka,2 MD, Koki Nakanishi,2 MD, Tomoko Nakao,2 MD, Yasutomi Higashikuni,2 MD, Norifumi Takeda,2 MD, Katsuhito Fujii,2 MD, Masao Daimon,2 MD, Hiroshi Akazawa,2 MD, Hiroyuki Morita,2 MD, Yutaka Matsuyama,1 PhD and Issei Komuro,2 MD

Summary

Left ventricular dilatation (LVD) and left ventricular hypertrophy (LVH) are risk factors for heart failure, and their detection improves heart failure screening. This study aimed to investigate the ability of deep learning to detect LVD and LVH from a 12-lead electrocardiogram (ECG). Using ECG and echocardiographic data, we developed deep learning and machine learning models to detect LVD and LVH. We also examined conventional ECG criteria for the diagnosis of LVH. We calculated the area under the receiver operating characteristic (AUROC) curve, sensitivity, specificity, and accuracy of each model and compared the performance of the models. We analyzed data for 18,954 patients (mean age (standard deviation): 64.2 (16.5) years, men: 56.7%). For the detection of LVD, the value (95% confidence interval) of the AUROC was 0.810 (0.801-0.819) for the deep learning model, and this was significantly higher than that of the logistic regression and random forest methods (P < 0.001). The AUROCs for the logistic regression and random forest methods (machine learning models) were 0.770 (0.761-0.779) and 0.757 (0.747-0.767), respectively. For the detection of LVH, the AUROC was 0.784 (0.777-0.791) for the deep learning model, and this was significantly higher than that of the logistic regression and random forest methods and conventional ECG criteria (P < 0.001). The AUROCs for the logistic regression and random forest methods were 0.758 (0.751-0.765) and 0.716 (0.708-0.724), respectively. This study suggests that deep learning is a useful method to detect LVD and LVH from 12-lead ECGs.

Key words: Artificial intelligence, Heart failure screening

Heart failure is a global pandemic affecting approximately 26 million people,1 and its prevalence is increasing.2 The incidence of heart failure increases with age, and a rapid increase in the number of patients with heart failure may result in societal burden and increased spending on medical expenses.3 Although early therapeutic intervention is recommended for the treatment and prevention of heart failure,4 asymptomatic heart failure is difficult to diagnose at an early stage.5 Therefore, improving the performance of heart failure screening is important for public health. Left ventricular dilatation (LVD) is a risk factor for the development of heart failure, with a risk ratio of 4.64.6 Left ventricular hypertrophy (LVH) is another risk factor for heart failure, with a risk ratio of 2.29.7 Echocardiography is the gold standard for the diagnosis of LVD and LVH.7 However, the high cost and equipment required for echocardiography make it impractical for screening. Therefore, an electrocardiogram (ECG) is often used to measure left ventricular activity in a non-invasive, rapid, and inexpensive manner. An ECG is useful for the diagnosis of arrhythmia and myocardial infarction. However, there are no established diagnostic criteria or screening methods based on ECG for LVD. Classical Sokolow-Lyon,8 Cornell,9 and Romhilt-Estes ECG criteria10 have been widely used to diagnose LVH, but are considered unsuitable for screening.11,12 To solve these problems, deep learning methods have been used to analyze ECGs. The deep learning method can extract potentially useful features that are not visually understandable; therefore, a high prediction performance can be expected. Deep learning can detect significant aortic regurgitation and left ventricular systolic dysfunction from a 12-lead ECG.13,14 However, there are no established LVD diagnostic criteria or screening methods based on ECG data, and the usefulness of deep learning for the
Figure 1. Study flowchart. The study flowchart shows the creation of the training, validation, and test datasets. ECG indicates electrocardiogram; TTE, transthoracic echocardiogram; LVDd, left ventricular end-diastolic diameter; and LVMI, left ventricular mass index.

detection of LVD has not been investigated. For the detection of LVH, a detailed comparison of deep learning with conventional criteria and other machine learning methods needs to be made. Visualizing the decision-making process of deep learning can clarify which ECG features are useful for the diagnosis and screening of LVD. The detection of LVD and LVH at an early stage is clinically significant. Therefore, in this study, we developed and validated an algorithm on the basis of deep learning for the detection of LVD and LVH from 12-lead ECGs.

Methods

Datasets: We included patients aged 18 years or older who had a transthoracic echocardiogram (TTE) at The University of Tokyo Hospital between January 2015 and December 2019, and they had an ECG performed within 28 days of the TTE. The ECG and matched TTE were paired (i.e., there was a one-to-one correspondence between the ECG and the TTE). When a single ECG corresponded to multiple TTEs, we selected either the latest TTE performed after the ECG or, if a TTE was not performed after the ECG, we selected the latest TTE performed before the ECG.

The initial dataset comprised 37,285 ECG-TTE pairs from 18,954 patients. We excluded 518 pairs in which the left ventricular end-diastolic diameter (LVDd) and left ventricular mass index (LVMI) had not been recorded. We then randomly divided the remaining 36,767 pairs from 18,763 patients into a training dataset (23,676 pairs from 12,008 patients: 64.4%), validation dataset (5,733 pairs from 3,002 patients: 15.6%), and test dataset (7,358 pairs from 3,753 patients: 20.0%) (Figure 1). Patients with multiple pairs were included in the same dataset to prevent data leaking. ECGs were acquired at a sampling rate of 500 Hz using 12-lead instruments (FCP-8700/FCP-8800; Fukuda Denshi, Tokyo). The voltage waveform was recorded for 10 seconds and exported to a comma-separated value (CSV) file. ECG measurement data were also exported to a CSV file. Echocardiography was performed by skilled ultrasound sonographers or cardiologists using a variety of instruments (Vivid 7/Vivid E9/Vivid E95, GE Healthcare, Waukesha, WI, USA; iE33/EPIQ7, Philips, Amsterdam, The Netherlands; Xario/Artida/Aploio300/AploioXV, Toshiba, Tokyo; or Acuson SC2000, Siemens, Munich, Germany). Echocardiographic measurements were performed by following American Society of Echocardiography recommendations, and each TTE was interpreted by 1 or 2 experienced echocardiologists. Measurement data were exported to a CSV file. The LVDd, interventricular septum thickness, and posterior left ventricular wall thickness were typically measured using a 2-dimensional method. The LVMI was calculated using the linear method.

Normal ranges of the mass and size of the left ventricle depend upon gender, race, and height. Therefore, the criterion of LVD and LVH was defined in accordance with the Japanese data of the Journal of the American Society of Echocardiography (JASE) study in the main analysis. LVD was defined as an indexed LVDd > 31 mm/
Table I. Baseline Characteristics of ECG-Transthoracic Echocardiographic Data

| Variable | Training ($n=23,676$) | Validation ($n=5,733$) | Test ($n=7,358$) | $P$ value |
|----------|------------------------|------------------------|------------------|-----------|
| Patient characteristics | | | | |
| Age (years) | 63.38 ± 6.93 | 63.07 ± 16.64 | 63.31 ± 17.07 | 0.488 |
| < 40 | 2,636 (11.13) | 634 (11.06) | 855 (11.62) | |
| 40–49 | 2,400 (10.14) | 556 (9.70) | 780 (10.60) | |
| 50–59 | 3,377 (14.26) | 738 (12.87) | 871 (11.84) | |
| 60–69 | 4,994 (21.09) | 1,254 (21.87) | 1,574 (21.39) | |
| 70–79 | 6,421 (27.12) | 1,672 (29.16) | 2,077 (28.23) | |
| 80+ | 3,848 (16.25) | 879 (15.33) | 1,201 (16.32) | |
| Male | 13,390 (56.56) | 3,243 (56.57) | 4,199 (57.07) | 0.732 |
| Height (cm) | 162.14 ± 26.13 | 161.69 ± 9.66 | 161.71 ± 0.12 | 0.774 |
| Weight (kg) | 60.08 ± 13.95 | 60.09 ± 14.76 | 59.96 ± 14.84 | 0.471 |
| ECG findings | | | | |
| AF | 1,968 (8.31) | 552 (9.63) | 621 (8.44) | 0.006 |
| HR | 73.86 ± 15.21 | 74.48 ± 15.86 | 74.24 ± 15.38 | 0.07 |
| R-R | 845.11 ± 166.95 | 838.16 ± 166.85 | 842.12 ± 165.31 | <0.001 |
| P-R | 171.07 ± 42.67 | 171.15 ± 46.86 | 171.41 ± 44.47 | <0.001 |
| QT | 105.67 ± 22.38 | 106.31 ± 23.18 | 106.08 ± 22.93 | 0.059 |
| QTc | 396.81 ± 39.52 | 396.88 ± 41.02 | 397.28 ± 39.91 | 0.428 |
| QRS axis | 28.31 ± 45.33 | 28.99 ± 44.86 | 27.18 ± 44.37 | 0.086 |
| ECG-TTE pairs were labeled with binary LVD and LVH data. Model development: We developed deep learning models and machine learning models, and validated their detection of LVD and LVH. Conventional ECG-based criteria were also used to diagnose LVH.

Deep learning models: We developed an ensemble neural network (ENN) model, which consisted of a convolutional neural network (CNN) and a deep neural network (DNN). The CNN used raw ECG data as input (Supplemental Figure). The first 6 layers were designed to learn temporal features within each ECG lead. Each temporal layer comprised a convolution layer, followed by a batch normalization layer, a nonlinear rectified linear unit activation function layer, and a max-pooling layer. After the last temporal layer, a spatial convolutional layer was used to fuse the data from all leads (i.e., all features were convolved into 1 layer). The DNN consisted of fully connected layers (Supplemental Figure), and 19 parameters comprising ECG features and demographic information were used as input data (Table I). Each fully connected layer was followed by a batch normalization layer, a rectified linear unit activation layer, and a dropout layer. In the ENN model, the CNN and DNN were combined immediately after the convolution layer to improve the performance (Supplemental Figure). We also developed CNN-only and DNN-only models.

Using a sigmoid output layer, the deep learning model predicted 1 (LVD- or LVH-positive) if the output continuous value was ≥ 0.5 and predicted 0 (LVD- or LVH-negative) otherwise. If the loss did not decrease for 10 consecutive epochs, model training was stopped even if 100 epochs had not been completed and the model weights at the lowest validation loss value were saved. We

m² for men and > 32 mm/m² for women. LVH was defined as an LVMI > 101 g/m² for men and > 85 g/m² for women. The results may change depending on the definitions of LVD and LVH. Therefore, we also used the definitions of LVD and LVH in accordance with the American Society of Echocardiography (ASE) in subanalysis because it is one of the most popular criteria throughout the world. For these criteria, LVD was defined as an LVDd > 58 mm for men and > 52 mm for women. LVH was defined as an LVMI > 125 g/m² for men and > 110 g/m² for women. ECG-TTE pairs were labeled with binary LVD and LVH data.

Model development: We developed deep learning models and machine learning models, and validated their detection of LVD and LVH. Conventional ECG-based criteria were also used to diagnose LVH.

Deep learning models: We developed an ensemble neural network (ENN) model, which consisted of a convolutional neural network (CNN) and a deep neural network (DNN). The CNN used raw ECG data as input (Supplemental Figure). The first 6 layers were designed to learn temporal features within each ECG lead. Each temporal layer comprised a convolution layer, followed by a batch normalization layer, a nonlinear rectified linear unit activation function layer, and a max-pooling layer. After the last temporal layer, a spatial convolutional layer was used to fuse the data from all leads (i.e., all features were convolved into 1 layer). The DNN consisted of fully connected layers (Supplemental Figure), and 19 parameters comprising ECG features and demographic information were used as input data (Table I). Each fully connected layer was followed by a batch normalization layer, a rectified linear unit activation layer, and a dropout layer. In the ENN model, the CNN and DNN were combined immediately after the convolution layer to improve the performance (Supplemental Figure). We also developed CNN-only and DNN-only models.

Using a sigmoid output layer, the deep learning model predicted 1 (LVD- or LVH-positive) if the output continuous value was ≥ 0.5 and predicted 0 (LVD- or LVH-negative) otherwise. If the loss did not decrease for 10 consecutive epochs, model training was stopped even if 100 epochs had not been completed and the model weights at the lowest validation loss value were saved. We
implemented a multi-input neural network in Python version 3.7.4 using the open-source PyTorch version 1.6.0 deep learning library and the NVIDIA Tesla V100 32 Gb graphics processing unit (NVIDIA Corporation, Santa Clara, CA, USA). For all 3 models, the following parameters were used: loss function, binary cross-entropy with logits loss (BCEWithLogitsLoss); optimizer, Adam; learning rate, 0.00005; and batch size, 128 (based on the grid search of previous studies\(^{13,14}\)).

**Machine learning models:** Logistic regression and random forest models, which have been used in previous studies,\(^{16,19}\) were used as machine learning models. Nineteen parameters comprising ECG features and demographic variables were used as input data, and “LVD” and “LVH” were used as predictor variables (Table I). Training was performed using the combined training and validation datasets, totaling 29,409 records from 15,010 patients. We developed the models using the open-source scikit-learn 0.21.3 module in Python. To search the random forest hyperparameters, we used a grid search and random search to maximize the area under the receiver operating characteristic (AUROC) curve.

**Conventional criteria:** Conventional Sokolow-Lyon, Cornell, Cornell product, and Romhilt-Estes criteria were used to diagnose LVH from ECGs. In this study, Sokolow-Lyon and Cornell diagnostic criteria, which have been adopted by The Japanese Society of Hypertension,\(^{20}\) were used for comparison with the machine learning models. The Sokolow-Lyon criterion is used to diagnose LVH when SV1 + RV5 > 35 mm or max{RV5, RV6} ≥ 26 mm. The Cornell diagnostic criterion is used when SV3 + RaVL > 28 mm (men) or > 20 mm (women).

**Statistical analysis:** To quantify the baseline characteristics of the patients, continuous variables are expressed as the mean and standard deviation, and categorical variables are expressed as percentages. To quantify the performance of the deep learning and machine learning models, we calculated the AUROC, sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV). We used the method of Hanley, et al.\(^{21}\) to calculate the 95% confidence interval (CI) of the AUROC and the normal approximation to the binomial distribution to calculate the 95% CIs of the sensitivity, specificity, accuracy, PPV, and NPV. We compared the performance of the deep learning and machine learning models with the conventional criteria using the AUROC. We also compared the performance of the CNN, DNN, and ENN models to determine which type of data is important to detect LVD and LVH. The Delong test\(^{22}\) was used to assess differences between the AUROC values of the ENN model and the other models. The R version 4.0.5 software environment (R Foundation for Statistical Computing, Vienna, Austria) was used for these statistical analyses. For hypothesis testing, statistical significance was defined as \(P < 0.05\).

To compare model outputs with existing medical knowledge, we needed to identify which parts of the ECG waveforms had the largest effect on algorithm decision-making. Therefore, we used gradient-weighted class activation mapping (Grad-CAM)\(^{23}\) to visualize which ECG leads and waveforms the ENN model focused on during prediction.

**Ethical considerations:** 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 a website.

**Results**

The mean (standard deviation) of age, height, and weight of the patients was 64.2 (16.5) years, 162.1 (24.0) cm, and 60.0 (14.2) kg, respectively. The percentage of men was 56.66%. Among 36,767 patient records, 6,073 fitted LVH and 4,007 records fitted LVD. No significant differences in baseline characteristics of the patients were observed among the training, validation, and test groups (Table I).

**Detection performance of various methods for LVD:** Figure 2 and Table II show the prediction performance for LVD using JASE criteria of the CNN, DNN, ENN, logistic regression, and random forest models. The ENN model had the highest AUROC (0.810, 95% CI: 0.801-0.819), sensitivity (0.429, 0.413-0.445), accuracy (0.800, 0.796-0.804), and NPV (0.890 (0.888-0.892)). The AUROC values of the CNN, DNN, logistic regression, and random forest models were 0.737 (0.727-0.747), 0.784 (0.775-0.793), 0.77 (0.761-0.779), and 0.757 (0.747-0.767), respectively. The specificity of the ENN model was 0.915 (0.912-0.918) and its PPV was 0.608 (0.601-0.615). The DNN model had the highest specificity (0.959, 0.957-0.961) and PPV (0.688, 0.680-0.696). The AUROC of the ENN model was significantly higher than that of the other models (all \(P < 0.001\)), and the AUROC of the CNN model was significantly higher than that of the DNN model (\(P < 0.001\)). The heat map produced by applying Grad-CAM to the ENN model (Figure 3) shows that the ENN model particularly focused on the QRS waves of V5 and V6 when detecting LVD. The prediction performance for LVD using ASE criteria is shown in Supplemental Table I.

**Detection performance of various methods for LVH:** Figure 4 and Table III show the prediction performance for LVH using JASE criteria of the CNN, DNN, ENN, logistic regression, and random forest models and Sokolow-Lyon and Cornell diagnostic criteria. The ENN model showed a higher performance than the conventional criteria regarding the AUROC, sensitivity, specificity, accuracy, PPV, and NPV. In particular, the AUROC, sensitivity, accuracy, and NPV of the ENN model were higher than those of all the other models, with values (95% CIs) of 0.784 (0.777-0.791), 0.605 (0.593-0.617), 0.732 (0.727-0.737), and 0.912 (0.911-0.914), respectively. The AUROC values of the CNN, DNN, logistic regression, and random forest models and Sokolow-Lyon and Cornell criteria were 0.776 (0.769-0.783), 0.760 (0.753-0.767), 0.758 (0.751-0.765), 0.716 (0.708-0.724), 0.584 (0.576-0.592), and 0.610 (0.602-0.618), respectively. The specificity of the ENN model was 0.813 (0.809-0.817) and its PPV was 0.673 (0.668-0.678). The DNN model had the highest specificity (0.907, 0.904-0.910) and PPV (0.729,
Advance Publication AUTOMATIC DETECTION OF LVD AND LVH FROM ECG

Figure 2. Predictive performance of various methods for left ventricular dilatation. Receiver operating characteristic curves were produced using 5 methods for the diagnosis of left ventricular dilatation (LVD). CNN indicates convolutional neural network; DNN, deep neural network; and ENN, ensemble neural network.

Table II. Predictive Performance of Various Methods for LVD

| Method          | AUROC (95% CI) | P value | Sensitivity (95% CI) | Specificity | Accuracy (95% CI) | PPV (95% CI) | NPV (95% CI) |
|-----------------|----------------|---------|----------------------|-------------|-------------------|--------------|--------------|
| CNN             | 0.737 (0.727-0.747) | < 0.001 | 0.202 (0.189–0.215)  | 0.971       | 0.789 (0.785–0.793) | 0.680 (0.670–0.690) | 0.860 (0.858–0.862) |
| DNN             | 0.784 (0.775-0.793) | < 0.001 | 0.262 (0.250–0.274)  | 0.959       | 0.785 (0.782–0.788) | 0.688 (0.680–0.696) | 0.868 |
| ENN             | 0.810 (0.801-0.819) | 0.429 (0.413–0.445) | 0.915 (0.912–0.918) | 0.800       | 0.608 (0.601–0.615) | 0.890        |
| Logistic Regression | 0.770 (0.761-0.779) | < 0.001 | 0.260 (0.246–0.274)  | 0.952       | 0.789 (0.785–0.793) | 0.626        | 0.867 |
| Random Forest   | 0.757 (0.747-0.767) | < 0.001 | 0.284 (0.269–0.299)  | 0.940       | 0.786 (0.782–0.790) | 0.596        | 0.869 |

The AUROC curve, sensitivity, specificity, accuracy, PPV, and NPV of 5 methods for diagnosing LVD are shown. The 95% confidence intervals are shown in parentheses. The P values correspond to comparisons with the ENN model using the Delong test. AUROC indicates area under the receiver operating characteristic; PPV, positive predictive value; NPV, negative predictive value; CNN, convolutional neural network; DNN, deep neural network; and ENN, ensemble neural network.

The AUROC of the ENN model was significantly higher than that of the other models (all P < 0.001), and the AUROC of the CNN model was significantly higher than that of the DNN model (P = 0.007). The heat map produced by applying Grad-CAM to the ENN model (Figure 5) shows that the ENN model particularly focused on QRS waves of V5 and V6 when detecting LVH. The prediction performance for LVH using the ASE criteria is shown in Supplemental Table II.

Discussion

In this study, we used deep learning to detect LVD and LVH from 12-lead ECGs. The AUROC of the ENN model was significantly higher than that of other machine learning models, which suggested that deep learning was better at detecting LVD than the other methods. The AUROC of the ENN model was also significantly higher than that of other machine learning models and conventional criteria, which suggested that deep learning was better at detecting LVH. The ENN model had a higher sensitivity, accuracy, and NPV than those in the other methods. This finding suggested that deep learning analysis of ECGs might be useful for detecting cases of LVD or LVH that have a low prevalence. The result that the ENN was better at detecting LVD and LVH than the other methods is robust because it was suggested by definitions of the JASE and the ASE.

To the best of our knowledge, this is the first study to apply machine learning methods, including deep learning, to 12-lead ECG analysis for the detection of LVD. Deep learning could be a useful method for detecting LVD because the AUROC of the ENN model was higher than that of machine learning models. The deep learning models also had a higher sensitivity and NPV than the other machine learning models. Because of the high NPV of the deep learning model, LVD-negative cases can be identified with a high probability, which should enable the develop-
Figure 3. Gradient-weighted class activation mapping-based diagnosis of left ventricular dilatation. Red and yellow areas of the heat maps indicate the features that the model focused on.

Figure 4. Predictive performance of various methods for left ventricular hypertrophy. Receiver operating characteristic curves produced using 7 methods for the diagnosis of left ventricular hypertrophy.

ment of more cost-effective screening. Because the diagnostic accuracy of the deep learning model was much higher than that previously achieved using chest radiograms, we found that the performance of this model was acceptable. Additionally, by visualizing heat maps of which leads and waveforms the deep learning model mainly focused on, we concluded that QRS waves of V5 and V6 may contain useful information for the diagnosis of LVD.

Deep leaning models may also useful for detecting
LVH because the AUROC of the ENN model was higher than those of machine learning models and conventional criteria. The ENN model had a higher sensitivity and NPV than the other machine learning models. The ENN model outperformed the conventional diagnostic criteria for LVH, with a higher AUROC, sensitivity, specificity, accuracy, PPV, and NPV. The conventional Sokolow-Lyon and Cornell criteria and machine learning models had high specificity and low sensitivity, as observed in previous studies. Additionally, the deep learning model focused on QRS waves of V5 and V6 when detecting LVH. This result is also consistent with previous findings and existing medical knowledge that V5 and V6 voltages are used for the Sokolow-Lyon diagnostic criteria.

LVD and LVH are important risk factors for the development of heart failure. However, in clinical practice, LVD and LVH do not tend to be effectively diagnosed from an ECG. In this study, the sensitivity of the deep learning model focused on QRS waves of V5 and V6 when detecting LVH.
learning model was insufficient for use of the model for screening of LVD and LVH. There are several methods to improve the sensitivity. The first method is to increase the sample size. In this study, we used the dataset for 18,954 patients; however, previous studies have used larger sample sizes and achieved higher sensitivities. Additionally, the second method is to improve the analytical method. In this study, the input data comprised only age, gender, height, and weight as demographic data; therefore, additional data, such as the subjects’ history of hypertension, heart failure, and cardiomyopathy might improve the performance of the statistical analysis. If the performance can be improved, simple detection using deep learning models may contribute to an accurate diagnosis of LVD and LVH. ß-Blockers and angiotensin receptor blockers can improve LVD and LVH. Therefore, if LVD and LVH can be detected at an early stage, preventing the onset of heart failure through treatment will be possible. If screening devices using our deep learning-based ECG analysis are approved, our models may contribute to the large-scale screening of heart failure and help alleviate the problem of the global pandemic of heart failure.

Some limitations of this study should be considered. First, because this study used data collected at only one facility, the external validity of our models needs to be examined using data collected at other facilities. An example of this limitation is that the diagnostic power of the Cornell criterion may vary depending on demographic variables, such as race, age, and gender. Additionally, LVH is the leading cause of sudden death in young athletes and hypertensive patients, and most previous LVH screening methods have been studied using subjects from these high-risk populations. However, the age of patients included in this study ranged from 18 to 102 years (mean age, 64.2 years), and they were enrolled in the hospital for outpatient visits or physical examinations. Therefore, they could not be considered a high-risk population for heart failure. Consequently, the applicability of our models may depend on the attributes of the population to which they are applied. The threshold values for deep learning and machine learning models may also require further validation. In this study, a neutral value was set to 0.5, and the sensitivity, specificity, accuracy, PPV, and NPV were calculated by predicting 0 (LVD- and LVH-negative) if the output continuous value was < 0.5 and predicting 1 (LVD- or LVH-positive) otherwise. However, the threshold value for this binary classification should be determined on the basis of existing medical knowledge or by considering its combined effects on sensitivity, specificity, and PPV. This study did not examine the effects of such varying threshold values. Second, we had no data describing the patients’ backgrounds. Therefore, we could not obtain detailed information, such as a history of hypertension, cardiac surgery, or pacemaker implantation. Additionally, we could not obtain information regarding whether any of the subjects had acute heart failure, which could significantly affect LVD values. This was a limitation when we adjusted for confounding background factors.

Conclusion

Deep learning is a useful method for the detection of LVD and LVH from 12-lead ECGs. Deep learning may contribute to the large-scale screening of heart failure and help alleviate the global pandemic of heart failure.

Acknowledgments

We thank Ellen Knapp, PhD, from Edanz (https://jp.edanz.com/ac) for editing a draft of this manuscript.

Disclosure

Conflicts of interest: The authors declare that there is no conflict of interest.

References

1. Ponikowski P, Anker SD, AlHabib KF, et al. Heart failure: preventing disease and death worldwide. ESC Heart Fail 2014; 1: 4-25.
2. Savarese G, Lund LH. Global public health burden of heart failure. Card Fail Rev 2017; 3: 7-11.
3. Tsutsui H, Ide T, Ito H, et al. JCS/JHFS 2021 guideline focused update on diagnosis and treatment of acute and chronic heart failure. Circ J 2021; 85: 2252-91.
4. Reed BN, Suet A. Stage B: what is the evidence for treatment of asymptomatic left ventricular dysfunction? Card Rev 2015; 11: 18-22.
5. Echouffo-Tcheugui JB, Eroqu S, Butler J, et al. Assessing the risk of progression from asymptomatic left ventricular dysfunction to overt heart failure: a systematic overview and meta-analysis. JACC Heart Fail 2016; 4: 237-48.
6. Gottdiener JS, Arnold AM, Aurigemma GP, et al. Predictors of congestive heart failure in the elderly: the Cardiovascular Health Study. J Am Coll Cardiol 2000; 35: 1628-37.
7. Devereux RB, Alonso DR, Lutas EM, et al. Echocardiographic assessment of left ventricular hypertrophy: comparison to necropsy findings. Am J Cardiol 1986; 57: 450-8.
8. Sokolow M, Lyon TP. The ventricular complex in left ventricular hypertrophy as obtained by unipolar precordial and limb leads. Am Heart J 1949; 37: 161-86.
9. Casale PN, Devereux RB, Kligfield P, et al. Electrocardiographic detection of left ventricular hypertrophy: development and prospective validation of improved criteria. J Am Coll Cardiol 1985; 6: 572-80.
10. Romhilt DW, Estes EH Jr. A point-score system for the ECG diagnosis of left ventricular hypertrophy. Am Heart J 1968; 75: 752-8.
11. Khurshid S, Friedman S, Pirruccello JP, et al. Deep learning to predict cardiac magnetic resonance-derived left ventricular mass and hypertrophy from 12-Lead ECGs. Circ Cardiovasc Imaging 2021; 14: e012281.
12. Cho Y, Kwon JM, Kim KH, et al. Artificial intelligence algorithm for detecting myocardial infarction using six-lead electrocardiography. Sci Rep 2020; 10: 20495.
13. Sawano S, Kodera S, Katsushika S, et al. Deep learning model to detect significant aortic regurgitation using electrocardiography: detection model for aortic regurgitation. J Cardiol 2022; 79: 334-41.
14. Katsushika S, Kodera S, Nakamoto M, et al. The Effectiveness of a Deep Learning Model to Detect Left Ventricular Systolic Dysfunction from Electrocardiograms. Int Heart J 2021; 62: 1332-41.
15. 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.

16. Asch FM, Miyoshi T, Addetia K, et al. WASE Investigators. Similarities and Differences in Left Ventricular Size and Function among Races and Nationalities: Results of the World Alliance Societies of Echocardiography Normal Values Study. J Am Soc Echocardiogr 2019; 32: 1396-406.e2.

17. Van Rossum G, Drake FL Jr. Python Reference Manual. Amsterdam: Centrum voor Wiskunde en Informatica; 1995.

18. Rahman QA, Tereshchenko LG, Kongkatong M, et al. Utilizing ECG-based heartbeat classification for hypertrophic cardiomyopathy identification. IEEE Trans Nanobioscience 2015; 14: 505-12.

19. Kwon JM, Kim KH, Jeon KH, et al. Development and validation of deep-learning algorithm for electrocardiography-based heart failure identification. Korean Circ J 2019; 49: 629-39.

20. Umemura S, Arima H, Arima S, et al. The Japanese Society of Hypertension Guidelines for the Management of Hypertension (JSH 2019). Hypertens Res 2019; 42: 1235-481.

21. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143: 29-36.

22. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988; 44: 837-45.

23. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int J Comput Vis 2020; 128: 336-59.

24. Loomba RS, Shah PH, Nijhawan K, et al. Cardiothoracic ratio for prediction of left ventricular dilation: a systematic review and pooled analysis. Future Cardiol 2015; 11: 171-5.

25. Pewnsner D, Jüni P, Egger M, et al. Accuracy of electrocardiography in diagnosis of left ventricular hypertrophy in arterial hypertension: systematic review. BMJ 2007; 335: 711.

26. Lim DY, Sng G, Ho WH, et al. Machine learning versus classical electrocardiographic criteria for echocardiographic left ventricular hypertrophy in a pre-participation cohort. Kardiol Pol 2021; 79: 654-61.

27. Kwon JM, Jeon KH, Kim HM, et al. Comparing the performance of artificial intelligence and conventional diagnosis criteria for detecting left ventricular hypertrophy using electrocardiography. Europace 2020; 22: 412-9.

28. McKelvie RS, Benedict CR, Yusuf S. Prevention of congestive heart failure and treatment of asymptomatic left ventricular dysfunction. Evidence based cardiology. London: BMJ Publishing Group; 1998: 703-21.

29. Rautaharju PM, Zhou SH, Calhoun HP. Ethnic differences in ECG amplitudes in North American white, black, and Hispanic men and women. Effect of obesity and age. J Electrocardiol 1994; 27 Suppl 1: 20-31.

30. Wang D, Xu JZ, Zhang W, et al. Performance of electrocardiographic criteria for echocardiographically diagnosed left ventricular hypertrophy in Chinese hypertensive patients. Am J Hypertens 2020; 33: 831-6.

31. Siontis KC, Liu K, Bos JM, et al. Detection of hypertrophic cardiomyopathy by an artificial intelligence electrocardiogram in children and adolescents. Int J Cardiol 2021; 340: 42-7.

32. Maron BJ, Friedman RA, Kligfield P, et al. Assessment of the 12-lead ECG as a screening test for detection of cardiovascular disease in healthy general populations of young people (12-25 Years of Age): a scientific statement from the American Heart Association and the American College of Cardiology. Circulation 2014; 130: 1303-34.