# An analysis of one-shot screening methods of ECG with different types of 2-D CNN

**Sanshiro Ishihara[1], Katsuhiko Fujiu[2,3], Takeshi Imai[1]**

[1]*Center for Disease Biology and Integrative Medicine, the Graduate School of Medicine, the University of Tokyo, Japan*
[2]*Department of Cardiovascular Medicine, the Graduate School of Medicine, the University of Tokyo, Japan*
[3]*Department of Advanced Cardiology, the Graduate School of Medicine, the University of Tokyo, Japan*

**Background**: The automatic classification of electrocardiogram (ECG) data using a convolutional neural network (CNN) model has been practiced earlier, but there are only a few studies on a 12-lead ECG dataset with various class labels. A large amount of ECG data is stored in hospital information systems in Japan, and this data can be used for machine learning. However, each sample in the data is mostly recorded for ten seconds and labelled with the corresponding abnormal classes, not for each lead or waveform, but for the entire 12-lead dataset. Therefore, the one-shot screening method using 2-D images of superimposed PQRST waveforms can be a solution in the given condition that all waveforms in a sample within a certain duration must be processed simultaneously.

**Objective**: We propose the one-shot screening method with different types of 2-D images of superimposed PQRST waveforms using CNN.

**Methods**: CNN and ensemble learning were applied to the ECG dataset, which contains over 9,000 samples with two classes, normal and abnormal, consolidated from 130 abnormal class labels for binary classification. We prepared three types of ECG images that were different in the manner in which they superimposed the PQRST waveforms of a single heartbeat: left-aligned, right-aligned, and centered. We compared the results of the three different images and analyzed false negative patterns to ascertain the characteristics of different types of 2D-CNN.

**Results**: The accuracy of all the frameworks were found to be above 0.867. The framework with the centered ECG images achieved the highest accuracy of 0.938 among the three. The listed abnormal classes with a high false negative ratio differed on the basis of the type of model.

**Conclusion**: The model with centered images showed the best score with the application of the one-shot method; however, the error analysis demonstrated that the characteristics of these models are varied.

**Keywords:** Electrocardiogram (ECG), Convolutional Neural Network (CNN), machine learning, arrhythmia, superimposed images, one-shot screening method.

## 1. INTRODUCTION

### 1.1 Background

The history of computerized Electrocardiogram (ECG) dates back to 1957 with Hubert Pipberger [1]. ECG classifications by automatic analyzers with rule-based algorithms have been developed for a long time. Automatic analyzers have been used in many hospitals; however, there are difficulties in the classification of abnormal findings [2]. In order to improve the performances of ECG classification, researches for the purpose of replacing rule-based algorithms with machine learning methods such as linear discriminant [3], SVM [4-6], Random Forest [7], and Neural Network [8] have been conducted. Convolutional Neural Network (CNN), a machine learning method, has been demonstrating successful results in other fields [9] and receiving attention in the medical field as well [10]. Several researches that apply CNN to ECG classification have already showed results that their models classified ECG datasets, such as the MIT-BIH Arrhythmia database and INCART, with high accuracy scores [11-12]. However, most of the datasets used in previous studies are very limited in their comprehensiveness, because the number of patients and types of labels given to abnormal findings in the datasets are very small [13]. Thus, in order to go a step ahead of previous studies in terms of comprehensiveness of the dataset, it is important to retrieve as many data samples as possible from the hospital information system, although most data samples are annotated by automatic analyzers in hospitals and not by experts.

With regard to researches using CNN models, most of them employed 1-D ECG data sequences for input of a CNN. However, a previous study reported that 2-D ECG data images were superior to 1-D ECG data sequences classified by a CNN [14]. Given that medical experts usually diagnose cardiac diseases using 2-D ECG data images and not 1-D ECG data sequences, it is assumed that the morphological features of ECG are significant. Moreover, applying 2-D ECG data to CNN models, contrary to 1-D ECG data, possibly makes its outcomes clinically interpretable. For these reasons, 2-D ECG data images were employed in this study.

### 1.2 One-shot screening

We have already proposed a method to use images of superimposed waveforms as inputs of CNN models, and termed it the one-shot screening method in a previous study [15]. The method can enhance the efficiency of computing, aggregating sparse parts of ECG data. ECG waveforms of a sample within the normal range can be transformed into a single heartbeat-like waveform image, as the waveforms are divided into single heartbeats and superimposed, because it is a repetition of almost the same shape of heartbeat waveforms at regular intervals. Meanwhile, in the case of abnormal samples, features of its abnormal findings are significantly expressed on superimposed images because of differences between the intervals and the shapes of each heartbeat. In particular, abnormal findings like premature irregular contractions can be easily emphasized without relying on the duration of ECG measurements.

### 1.3 Objectives

In the classification of ECG data, it is expected that preprocessing the images strongly affects the performance of the model. Our objective is to investigate the optimal setting of the one-shot screening method by comparing the results of classifications among three different types of preprocessing images.

## 2. MATERIALS AND METHOD

### 2.1 Materials

In this study, we used 12-lead ECG data measured in the University of Tokyo Hospital in 2016, which contains 9,190 samples from 6,281 adult patients. Each sample has 12-lead ECG data sequences sampled at 500Hz for 10 seconds and is annotated with one or more labels by the current automatic analyzer used in the hospital. Therefore, it is assumed that the dataset contains a few samples that are annotated with wrong labels. The number of types of abnormal labels is 131, and certain types of abnormal findings are subdivided on the basis of seriousness. For example, Inferior Infarction is subdivided into three classes: Inferior Infarction, Possible Inferior Infarction, and Possible Inferior Infarction (Suspect). If a sample does not correspond to any abnormal labels, the sample is annotated as normal class (i.e., Within Normal Limits).

### 2.2 Preprocessing

It is known that accurately detecting R peaks is the most important aspect of preprocessing for ECG data analysis and affects the performance of the model. Thus, we first developed an algorithm to detect R peaks in the following manner:

1) Determine whether each QRS complex of the waveform in 12 leads is upward or downward.
2) Extract R peaks in two patterns: tachycardia pattern, which finds peaks in a short window, and normal pattern, which finds peaks in a long window (normal pattern can also extract the R peaks of bradycardia patients).
3) Calculate standard deviation of R-R intervals in each extracting pattern.
4) Determine a pattern with the lowest standard deviation within a variable range, based on the number of extracted R peaks in its pattern.
5) Reflect the determined pattern in ECG data in the other 11 leads.

In the third step of the algorithm, if a sample has bradycardia, or slow heart rate, then very high or low standard deviations are computed from tachycardia patterns and vice versa. With regard to the fifth step, there is a concern that a reflected peak point moves slightly away from the true peak point because R peak points are slightly different among each lead, although ECG data is measured simultaneously in 12 leads. However, the benefit that the reflection can cover other leads that cannot be measured well due to noise or any abnormal situations overwhelms the concern. This algorithm is capable of detecting R peaks of normal samples with no errors, and intentionally regard premature contractions as exceptions in order to emphasize its feature on superimposed images.

Subsequently, ECG data sequences are divided into a single heartbeat each and plotted with superimposition on images of 150×150 pixels. Each sample has 12 superimposed images from 12 leads. We prepared three types of ECG images that differed in the manner in which PQRST waveforms of a single heartbeat were superimposed—left-aligned, right-aligned, and centered (Fig. **1**)—in order to identify the effects of the differences in the superimposed images. Left-aligned and right-aligned images are generated from the waveforms divided at R peak points, and centered images are generated from the waveforms divided at middle points between R-R intervals. Thus, with regard to left-aligned images, the portion between R wave and T wave tends to be aligned well, but the portion between P wave and R wave tends to be blurred; for right-aligned images, the opposite portion tends to be blurred. Centered images are different from these two types of images, and PQRST waveforms are clearly plotted, except for the edges of both sides. Given that each type of image has different blurred portions, it is expected that each type of image is unable to accurately classify the abnormal findings expressed in the blurred portions.

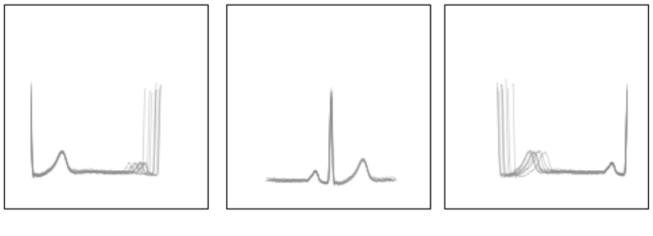

**Fig. 1.** Superimposition images: right-aligned, centered, and left-aligned.

With regard to generating superimposition images, we set the permeability of plotted waveforms to make the superimposed portions darker in correspondence with the number of superimpositions. All ECG data in this study were measured for 10 seconds; however, even if the duration of measurement is not 10 seconds, it is possible to reveal the number of superimposed waveforms in images within a certain period of time by changing permeability. The vertical and horizontal axes indicate electric potential and time course of heartbeats, respectively. The vertical axis is fixed, with a range between maximum and minimum potential, and the horizontal axis is also fixed with a duration of R-R intervals, which 95% of the samples in the datasets can fit in. The ranges of vertical axis are different among 12 leads. Because of this, waveforms of 5% of the samples exceeded the ranges of each axis, and the exceeding portions of waveforms are not plotted on the images.

### 2.3 Data augmentation

The number of abnormal samples was approximately two times more than normal samples in the datasets, since the ECG data was sampled in the hospital. An imbalance between classes might negatively influence classification performance; thus, we implemented data augmentation to ease this concern. There are several methods of data augmentation for image pattern recognition, such as rotation of images and using intermediate values between samples in the same class. However, we applied the vertical slide method for data augmentation, taking into consideration that we use only images that highly likely exist in clinical settings, even though the method may degrade the features of electric potential as an adverse effect.

### 2.4 Learning models

Fig. **2** illustrates the flowchart of the proposed framework. After the preprocessing phase, CNN and Random Forest were used for classification models, and the classification process was divided into two phases: (1) classification of waveforms in each lead by CNN and (2) classification of samples by Random Forest using the CNN outputs of 12 leads. Random Forest output predicted final values, thereby indicating normal or abnormal classes. For the first phase, we built 11-layer CNN models for each lead, so that the input of a CNN model is 22,500 values (image of 150 × 150 pixels). Each CNN model is independent of each other, so that a loss value computed from one particular CNN model is not backpropagated to other CNN models. The CNN models were optimized by tuning hyperparameters, such as the number of convolutional layers, fully connected layers, kernel size, dropout rate, and learning rate (Table **1**). All the CNN models were saved at the tenth epoch, because of the overfitting of the validation dataset after the epoch. The Random Forest model used in the second phase is an ensemble learning of multiple decision trees as small estimators. This model was applied to compute the importance of 12 leads. The input of the Random Forest model is 24 values computed from 12 CNN models with identity function. The Random Forest model was optimized by tuning hyperparameters such as max depth, max features, min samples split, and the number of estimators (Table **1**). It is easy to classify abnormal samples as Bradycardia and Tachycardia only by heart rate measured in the preprocessing; however, heat rate was not applied to the models that help to evaluate performance of our proposed system in image recognition. For the same reason, any other metadata, such as gender and age, was not applied to the models either.

### 2.5 Experimental settings

We consolidated 131 abnormal labels into one abnormal class to implement binary classifications of 2,984 normal samples and 5,747 abnormal samples, excluding samples annotated as Arm Leads Reversed or Unsatisfactory Recorded, and samples measured in eight leads from the dataset. We finally applied data augmentation to the dataset and created a training dataset with 5,000 normal samples and 5,000 abnormal samples, and a test dataset with 748 normal samples and 748 abnormal samples. Three cases of the binary classification models based on different types of superimposition images were implemented. Apart from the three experiments with different types of superimposed images, we conducted another experiment using the Random Forest model with 72 integrated outputs (2 outputs × 12 leads × 3 types of images) from CNN models and predicted. Performances of the proposed frameworks were evaluated in each phase by measurements such as Accuracy, Precision, Recall, and F-1 Score. Subsequently, the results in the second phase were analyzed by sorting abnormal findings with high false negative rates in order to evaluate per-

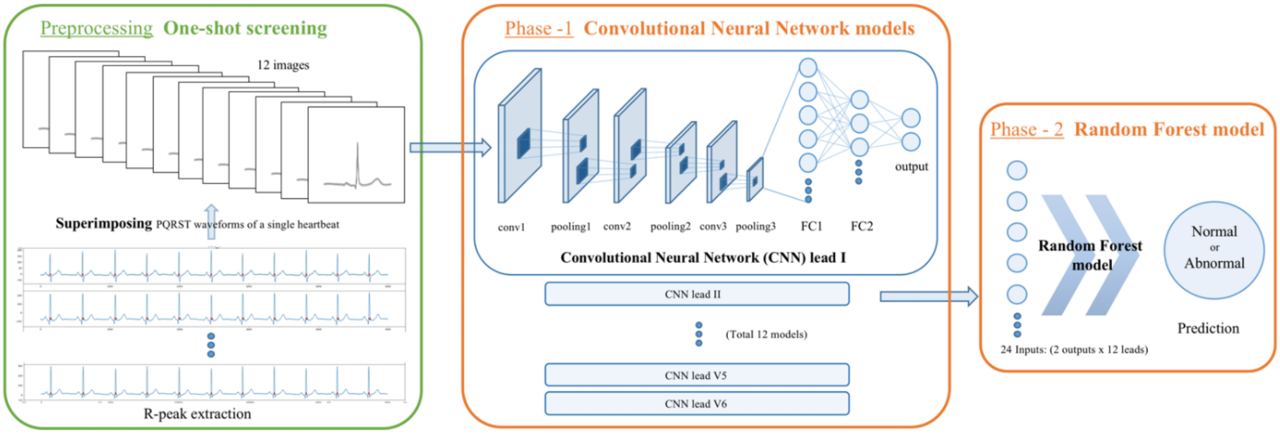
Preprocessing One-shot screening
12 images
Superimposing PQRST waveforms of a single heartbeat
R-peak extraction
Phase -1 Convolutional Neural Network models
conv1
pooling1
conv2
pooling2
conv3
pooling3
FC1
FC2
output
Convolutional Neural Network (CNN) lead I
CNN lead II
(Total 12 models)
CNN lead V5
CNN lead V6
Phase - 2 Random Forest model
Random Forest model
Normal or Abnormal
Prediction
24 Inputs: (2 outputs x 12 leads)

**Fig. 2.** Flowchart of the framework**.**

**Table 1. Optimized setting and search values of hyperparameter tuning for each model.**

| Hyperparameters | Optimized Setting | Search Values |
|---|---|---|
| *CNN model* | | |
| number of convolutional layers | 6 | [4, 5, 6] |
| kernel size of filters | 3 | [2, 3] |
| number of fully connected layers | 2 | [2, 3] |
| dropout rate | 0.3 | [0.1, 0.2, 0.3, 0.4, 0.5] |
| learning rate | 0.0005 | [0.1, 0.01, 0.001, 0.0005] |
| *Random Forest model* | (Left, Center, Right, Integrated) | |
| max depth | (20, 20, 20, 5) | [3, 5, 20] |
| max features | (3, 3, 3, 3) | [1, 3, 10] |
| min samples split | (20, 5, 3, 20) | [3, 5, 20] |
| number of estimators | (500, 50, 100, 50) | [50, 100, 500, 1000] |

formances with regard to specific abnormal findings and compare the characteristics of the three models. Furthermore, the importance of each lead for the classification in the second phase were computed using average reduction of Gini coefficient of small estimators in the Random Forest model.

We conducted another survey on the case that the framework yielded the best score in binary classification. This is essential for analyzing marked errors in order to improve our proposed method. Apart from the errors due to the framework, it is possible that the employed dataset annotations from the automatic analyzers contradict the diagnosis by clinical experts. We extracted samples with values computed by one or more CNN models predicting the opposite class of the samples; two experts diagnosed the samples. If the diagnosis of the two experts was conflicting, a final diagnosis was determined through an additional discussion. A prediction value exceeding a threshold set by interquartile ranges was defined as an ***outlier*** (Fig. **3**).

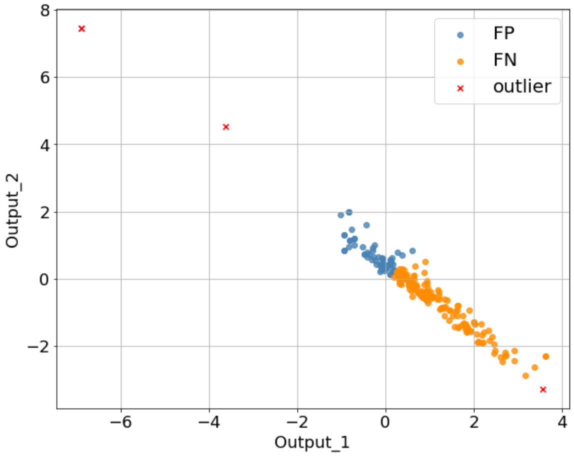

**Fig. 3.** A scatter plot of output values computed by the CNN model for lead aVR in the case of centered images. The plotted samples were false positives or false negatives in the classification by the CNN model. In this case, three outliers were recognized.

## 3. RESULTS

### 3.1 Results of CNN models (Phase 1)

Mean scores of 12-lead accuracy yielded by CNN models in all the cases were above 0.796; the highest score was 0.838 and was found in the case of centered images (Table **2**). Leads that yielded the highest and lowest accuracy were different in all the cases. In the cases of left-aligned and right-aligned images, each CNN model yielded almost the same number of false positive (FP) and false negative (FN), and the scores of Precision and Recall were close. On the other hand, in the case of centered images, there was significant imbalance between the number of FP and the number of FN in certain leads, which showed that CNN models in certain leads had gaps in the scores for Precision and Recall. For example, the CNN model in the V4 lead showed a Precision of 0.924 and Recall of 0.680.

### 3.2 Results of the Random Forest models (Phase 2)

In all the cases, accuracy scores yielded by the Random Forest model were above 0.863 and higher than results in the first phase (Table **3**). The Random Forest model with centered images achieved the highest accuracy of 0.937 and the greatest improvement from the result in the first phase. Accuracy yielded by the Random Forest model with integrated output from the CNN models was 0.904, which was higher than that of the models in the cases of left-aligned and right-aligned images, and lower than that of the model in the case of centered images. To avoid a misleading statement, hereafter, a result yielded by a Random Forest model is referred to as a result of a framework.

### 3.3 Abnormal findings with false negatives

Table **4** (on the sixth page) presents lists of abnormal findings contained in the test dataset, which contains over 10 samples and FN rates. We sorted them by FN rates and categorized them into three groups: Group A has abnormal findings with an FN rate of 0.1≦, Group B has abnormal findings with an FN rate of 0＜0.1, and Group C has abnormal findings with an FN rate of 0. We ascertained features of each type of images from differences in the comparison. As the framework with centered images yielded the highest accuracy score, the

**Table 2. Accuracy, Precision, and Recall scores of CNN models and mean values of scores in 12 leads.**

| | I | II | III | aVR | aVL | aVF | V1 | V2 | V3 | V4 | V5 | V6 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Left* | | | | | | | | | | | | | |
| Accuracy | 0.789 | 0.826 | 0.777 | 0.805 | 0.785 | 0.805 | 0.781 | 0.799 | 0.782 | 0.787 | 0.806 | 0.816 | 0.797 |
| Precision | 0.803 | 0.857 | 0.767 | 0.811 | 0.781 | 0.821 | 0.783 | 0.796 | 0.787 | 0.796 | 0.846 | 0.845 | 0.808 |
| Recall | 0.766 | 0.783 | 0.797 | 0.795 | 0.791 | 0.779 | 0.778 | 0.805 | 0.774 | 0.773 | 0.749 | 0.773 | 0.780 |
| *Center* | | | | | | | | | | | | | |
| Accuracy | 0.832 | 0.847 | 0.846 | 0.851 | 0.838 | 0.870 | 0.834 | 0.859 | 0.783 | 0.812 | 0.842 | 0.840 | 0.838 |
| Precision | 0.941 | 0.905 | 0.862 | 0.896 | 0.825 | 0.916 | 0.934 | 0.894 | 0.735 | 0.924 | 0.874 | 0.906 | 0.884 |
| Recall | 0.707 | 0.775 | 0.825 | 0.794 | 0.857 | 0.816 | 0.718 | 0.814 | 0.886 | 0.680 | 0.798 | 0.758 | 0.786 |
| *Right* | | | | | | | | | | | | | |
| Accuracy | 0.784 | 0.799 | 0.781 | 0.803 | 0.782 | 0.797 | 0.775 | 0.820 | 0.811 | 0.797 | 0.799 | 0.798 | 0.796 |
| Precision | 0.792 | 0.818 | 0.779 | 0.823 | 0.790 | 0.802 | 0.772 | 0.810 | 0.835 | 0.817 | 0.816 | 0.778 | 0.803 |
| Recall | 0.770 | 0.769 | 0.786 | 0.771 | 0.769 | 0.789 | 0.781 | 0.834 | 0.777 | 0.767 | 0.773 | 0.834 | 0.785 |

number of abnormal findings in Group A was three and the smallest among the three cases, and the number of abnormal findings in Group C was 19 and the largest among the three cases.

**Table 3. The performance of Random Forest for the three types of superimposing ECG images and the integrated model.**

| | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Left | 0.863 | 0.844 | 0.890 | 0.867 |
| Center | **0.937** | **0.931** | **0.944** | **0.938** |
| Right | 0.878 | 0.853 | 0.913 | 0.882 |
| Integrated | 0.904 | 0.898 | 0.910 | 0.904 |

### 3.4 Importance of 12 leads for the second phase

Table **5** presents the importance of 12 leads for the classification in the second phase in all the cases. Lead aVR was most important for Random Forest models with centered and right-aligned images. Furthermore, the orders of the important leads in both the cases (center and right) were similar to each other.

**Table 5. Importance of features for Random Forest classification in the case of center-aligned images. (This table is placed before Table 4 due to layout consideration.)**

| Left | | Center | | Right | |
|---|---|---|---|---|---|
| **Leads** | **Importance** | **Leads** | **Importance** | **Leads** | **Importance** |
| V5 | 0.276 | aVR | 0.1609 | aVR | 0.0874 |
| V1 | 0.1233 | V2 | 0.0739 | aVL | 0.0845 |
| V3 | 0.088 | aVL | 0.0603 | V3 | 0.0807 |
| aVL | 0.0633 | V1 | 0.0563 | V2 | 0.0678 |
| V2 | 0.0538 | V3 | 0.0487 | III | 0.064 |
| aVR | 0.0445 | III | 0.0355 | aVF | 0.0304 |
| III | 0.0354 | II | 0.0311 | V1 | 0.0286 |
| V4 | 0.0205 | aVF | 0.0252 | II | 0.0219 |
| II | 0.0133 | V6 | 0.0022 | V4 | 0.0148 |
| aVF | 0.0113 | I | 0.0021 | I | 0.014 |
| V6 | 0.0025 | V4 | 0.0019 | V5 | 0.0052 |
| I | 0.001 | V5 | 0.0019 | V6 | 0.0007 |

### 3.5 Outlier analysis

Several samples were TP-outliers or TN-outliers: certain CNN models predicted the wrong class with outliers, but the framework predicted this correctly. Such a sample was not significant for the classification. It was confirmed by experts that both the predictions were correct. For example, in the case of an abnormal sample with abnormal findings that its features appear only in certain leads, waveforms in several leads are normal. The number of FP-outliers and FN-outliers that both the CNN models and the framework predicted wrongly were 13 and 23, respectively, which add up to a total of 36. Of these, 3 out of 23 FP-outliers and 3 out of 13 FN-outliers were the cases in which the experts confirmed the existence of annotation errors by the automatic analyzer.

## 4. DISCCUSION

### 4.1 Overall evaluation

With regard to the results in the first phase, CNN models with centered images achieved comprehensively better scores than other models. Interestingly, imbalances (i.e., high precision and low recall and vice versa) occurred only in CNN models with centered images. The imbalances may positively affect classification in the second phase. However, the causal relationship is not clarified yet and needs further analysis. With regard to the result of importance, focusing on the case of centered images, the order of the important leads for the classification was different from the perspectives of clinical experts. The results may also be related to the breakdown of abnormal findings in the dataset; thus, it is required that this be investigated in a future study.

The highest accuracy of the proposed framework was 0.937 using centered superimposed images. The more serious the abnormal findings were, the more accurately they were classified. Morphological differences of serious abnormal findings with respect to normal samples could be greater. Overall, the result supported our presumption that applying 2-D ECG data to classifications using the CNN model was a valid method. Our highest score was lower than the score of 0.999 obtained in a previous study that used 2-D ECG data; however, it was successful, given that our task targeted a larger number of abnormal findings and that the abnormal

**Table 4. Lists of abnormal findings in which the test dataset contains over 10 samples and FN rates. The numbers in parentheses are the total number of abnormal findings in the test dataset. The names of abnormal findings that were classified with no errors in the only case of center-aligned images are presented in bold.**

| Left | Rate | Center | Rate | Right | Rate |
|---|---|---|---|---|---|
| Short PR | 0.455 (11) | RSR' Pattern | 0.167 (30) | RSR' Pattern | 0.300 (30) |
| A-V Junctional Rhythm | 0.364 (11) | Sinus Arrhythmia | 0.154 (26) | Short PR | 0.182 (11) |
| RSR' Pattern | 0.233 (30) | High Voltage (Left Ventricle) | 0.129 (31) | Counterclockwise Rotation | 0.170 (94) |
| High Voltage (Left Ventricle) | 0.226 (31) | Counterclockwise Rotation | 0.096 (94) | Abnormal Q | 0.154 (13) |
| Counterclockwise Rotation | 0.213 (94) | Short PR | 0.091 (11) | High Voltage (Left Ventricle) | 0.129 (31) |
| Slight Right Ventricular Hypertrophy | 0.200 (15) | Possible Inferior Infarction | 0.091 (11) | Borderline Abnormal Q | 0.125 (24) |
| Clockwise Rotation | 0.133 (30) | Left Atrial Enlargement | 0.091 (22) | IRBBB | 0.122 (41) |
| Possible Inferior Infarction (Suspect) | 0.118 (17) | Poor R Progression | 0.086 (35) | Sinus Arrhythmia | 0.115 (26) |
| Poor R Progression | 0.114 (35) | Borderline Abnormal Q | 0.083 (24) | A-V Junctional Rhythm | 0.091 (11) |
| Right Axis Deviation | 0.094 (32) | Slight ST-T Abnormality (Suspect) | 0.067 (15) | Poor R Progression | 0.086 (35) |
| Flat T | 0.088 (91) | Clockwise Rotation | 0.067 (30) | Slight ST-T Abnormality (Suspect) | 0.067 (15) |
| Borderline Abnormal Q | 0.083 (24) | Possible Inferior Infarction (Suspect) | 0.059 (17) | Low Voltage (Limb Leads) | 0.061 (33) |
| Abnormal Q | 0.077 (13) | Sinus Bradycardia | 0.048 (42) | Possible Inferior Infarction (Suspect) | 0.059 (17) |
| Slight ST-T Abnormality | 0.075 (40) | Flat T | 0.033 (91) | Slight ST-T Abnormality | 0.050 (40) |
| Slight ST-T Abnormality (Suspect) | 0.067 (15) | Low Voltage (Limb Leads) | 0.030 (33) | Negative T | 0.048 (62) |
| Negative T | 0.065 (62) | Premature Atrial Contraction | 0.026 (38) | Left Atrial Enlargement | 0.045 (22) |
| Low Voltage (Limb Leads) | 0.061 (33) | Slight ST-T Abnormality | 0.025 (40) | Flat T | 0.044 (91) |
| PR Prolongation | 0.050 (40) | Slight Left Axis Deviation | 0.019 (104) | Slight QT Prolongation | 0.038 (26) |
| Slight Left Axis Deviation | 0.048 (104) | Negative T | 0.016 (62) | Slight Left Axis Deviation | 0.038 (104) |
| Sinus Bradycardia | 0.048 (42) | PR Prolongation | 0.000 (40) | Clockwise Rotation | 0.033 (30) |
| Left Atrial Enlargement | 0.045 (22) | QT Prolongation | 0.000 (29) | Right Axis Deviation | 0.031 (32) |
| A-V Block 1 | 0.045 (22) | ST -T Abnormality | 0.000 (79) | Sinus Bradycardia | 0.024 (42) |
| Sinus Arrhythmia | 0.038 (26) | **Abnormal Q** | 0.000 (13) | Left Ventricular Hypertrophy | 0.018 (57) |
| Left Ventricular Hypertrophy | 0.035 (57) | **Right Axis Deviation** | 0.000 (32) | CRBBB | 0.015 (66) |
| Left Axis Deviation | 0.027 (37) | Inferior Infarction | 0.000 (16) | ST -T Abnormality | 0.013 (79) |
| IRBBB | 0.024 (41) | CRBBB | 0.000 (66) | PR Prolongation | 0.000 (40) |
| QT Prolongation | 0.000 (29) | CLBBB | 0.000 (12) | QT Prolongation | 0.000 (29) |
| ST -T Abnormality | 0.000 (79) | Slight QT Prolongation | 0.000 (26) | Inferior Infarction | 0.000 (16) |
| Inferior Infarction | 0.000 (16) | Slight Right Ventricular Hypertrophy | 0.000 (15) | Possible Inferior Infarction | 0.000 (11) |
| Possible Inferior Infarction | 0.000 (11) | Left Axis Deviation | 0.000 (37) | CLBBB | 0.000 (12) |
| CRBBB | 0.000 (66) | **Left Ventricular Hypertrophy** | 0.000 (57) | Slight Right Ventricular Hypertrophy | 0.000 (15) |
| CLBBB | 0.000 (12) | Premature Ventricular Contraction | 0.000 (38) | Left Axis Deviation | 0.000 (37) |
| Slight QT Prolongation | 0.000 (26) | Atrial Fibrillation | 0.000 (67) | Premature Atrial Contraction | 0.000 (38) |
| Premature Atrial Contraction | 0.000 (38) | Artificial Pacemaker Rhythm | 0.000 (13) | Premature Ventricular Contraction | 0.000 (38) |
| Premature Ventricular Contraction | 0.000 (38) | Possible Anterior Infarction | 0.000 (12) | Atrial Fibrillation | 0.000 (67) |
| Atrial Fibrillation | 0.000 (67) | **IRBBB** | 0.000 (41) | Artificial Pacemaker Rhythm | 0.000 (13) |
| Artificial Pacemaker Rhythm | 0.000 (13) | A-V Block 1 | 0.000 (22) | Possible Anterior Infarction | 0.000 (12) |
| Possible Anterior Infarction | 0.000 (12) | A-V Junctional Rhythm | 0.000 (11) | A-V Block 1 | 0.000 (22) |

findings classified with many false negatives were not covered in the previous study.

In terms of comparison of methods, the experiments showed various results in different types of superimposed images. In fact, there was an abnormal finding that was  better classified in the cases of left-aligned and right-aligned images than centered images, although the framework with centered images achieved the greatest performance. The Integrated Random Forest model was supposed to achieved the highest accuracy, but its result differed from our expectation. Thus, it is essential to design a model that is more suitable for integrated data in a future study. The effects of differences in the types of superimposed images were identified through an analysis of false negatives, which is discussed in the following section.

## 4.2 Comparison of superimposition images

### *4.2.1 Left-aligned and right-aligned:*

Short PR and A-V Junctional Rhythm were the abnormal findings that the framework with left-aligned images could not classify correctly, although the other frameworks could. The significant features of both the abnormal findings should be expressed over the portion of waveforms between the P-wave and R-wave, which is blurred in most of the left-aligned images, as discussed earlier in the paper; thus, it is considered that this is the reason why the two abnormal findings were classified at high FN rates. The framework with left-aligned images classified Incomplete Right Bundle Branch Block (IRBBB) more correctly than with right-aligned images. This result indicates that the significant features of IRBBB are more likely to be blurred on right-aligned images as compared to left-aligned images. With regard to Abnormal Q and Borderline Abnormal Q, they were found in Group A only in the case of right-aligned images. It is difficult to ascertain the cause of the errors because this result contradicts our presumption that the abnormal features of Q-wave would be clearly expressed on particularly right-aligned images.

### *4.2.2 Centered:*

Focusing on specific abnormal findings, Abnormal Q, Right Axis Deviation, Left Ventricular Hypertrophy, and IRBBB were found in Group C only in the case of centered images. However, Sinus Arrhythmia was not classified correctly by the framework with centered images, and its FN-rate was higher than that of the other frameworks. With regard to the superimposed images, centered images show the shape of the QRS complex most clearly, but this is not sufficiently adequate to reveal the features of Sinus Arrhythmia. A slightly high variance of R-R intervals regarded as the feature of Sinus Arrhythmia is expressed as dispersion of superimposed waveforms on the edges in the case of a centered image. Consequently, such a feature on a centered image was not sufficient for image pattern recognition; rather, left-aligned and right-aligned images were considered more suitable to emphasize Sinus Arrhythmia.

### *4.2.3 Common features:*

RSR Pattern and High Voltage (Left) were found in Group A in all the cases. With regard to the RSR Pattern, it is the least serious abnormal finding in the three abnormal labels related to the Right Bundle Branch Block; therefore, it is considered difficult for the frameworks to classify the waveforms of the RSR Pattern because of the similarity with waveforms of normal samples in images. It was assumed that High Voltage (Left) would be classified with errors, since each image was vertically moved for data augmentation.

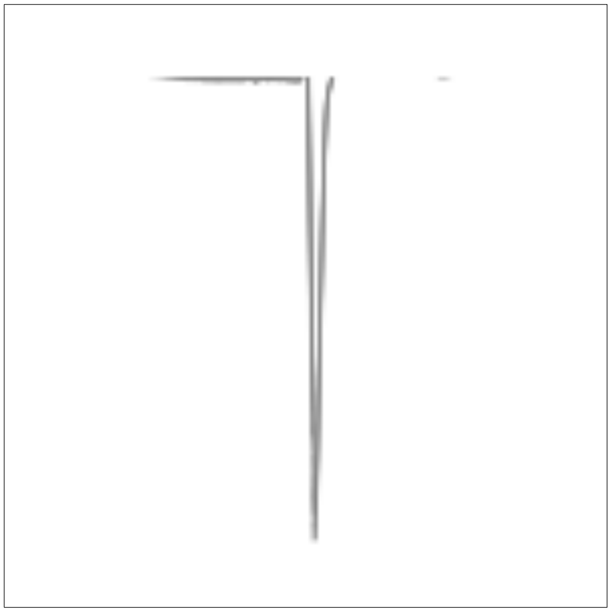

**Fig. 4.** A superimposition image of a waveform cut-off in the inappropriate direction.

Further, it was clarified that several abnormal findings remain as difficulties for classification, even if the framework with centered images is employed. However, High Voltage (Left) is supposed to be classified correctly using electric potential values in ECG data sequences. Sinus Arrhythmia is also supposed to be classified using the values of R-R intervals and its variance, which can be extracted in the preprocessing. Counterclockwise Rotation was not classified correctly in all the cases; however, it was recommended that the case annotated only with Counterclockwise Rotation should be regarded as having a normal status by medical experts. In the case of centered images, there were 12 false negative samples annotated with only one of the three abnormal labels such as High Voltage (Left), Sinus Arrhythmia, and Counterclockwise Rotation. Assuming that the three abnormal findings would be classified correctly, accuracy would be improved to 0.945. It is impossible to estimate how such an improvement will influence false positives, because it is difficult to ascertain which abnormal findings were predicted to be normal by the framework.

## 4.3 FP-outliers and FN-outliers

According to the analysis on the outliers of the prediction results, we found that at least 6/36 cases were wrongly annotated by the automatic analyzer. Even though the total number of wrong annotations by the automatic analyzer was not investigated, the accuracy with centered images would be improved at least to 0.947, counting the six cases as accurately predicted. Our proposed method needs to be improved in order to more efficiently identify automatic annotation errors of FP-outliers and FN-outliers.

With regard to the remaining prediction errors with outliers, 10 out of 13 FN-outliers should be considered as clinically critical errors; however, most of them were samples with abnormal findings such as Sinus Arrhythmia and Counterclockwise Rotation: they are usually considered as normal-state in a clinical setting. The remainder were annotated with significant abnormal findings and remain to be resolved.

On the other hand, 20 out of FP-outliers are not considered as clinically critical errors, given that they would pass the secondary screening by experts. The analysis identified that most of the 20 FP-outliers were caused by artifacts such as noise and baseline drift, which implies the necessity of a function that filters artifacts. In addition to artifacts, marked high voltage was also a cause of FP-outliers, because some waveforms exceeded the vertical range of superimposed images and the excessive portions were cut off. As discussed in 2.2, the vertical range of images is not set for the samples that have large differences in electric potential. If the vertical range is allowed to have the capacity for large gaps in electric potential, the vertical length of most waveforms plotted on images is shortened and the resolution of images is degraded. However, this technical difficulty is expected to be eased by appropriately plotting waveforms to ensure that baselines remain on images (Fig. **4**) .

### 4.4 Limitations

One of the limitations of this study is that the proposed method is not suitable for samples with large gaps in electric potential. Another limitation is that the samples in the dataset were not annotated by experts but by automatic analyzers. In this study, we conducted experiments using the dataset stored in the University of Tokyo Hospital instead of a small dataset annotated by experts, with the purpose of applying a certain number of samples to our proposed framework. We were unable to thoroughly investigate certain abnormal findings, because the number of samples with several abnormal findings in the dataset was not sufficiently large. However, we plan to make an arrangement for an alternative dataset with true annotations and a greater number of patients. It is considered that almost the same performance quality is yielded even if the annotations are changed. The development of materials is expected to resolve the two last limitations.

## 5. CONCLUSION

Apart from the results of evaluation scores, the analysis of false negatives led us to conclude that centered images were the most acceptable means of superimposing waveforms and yielded comprehensively great performance. At the same time, abnormal findings that were not classified correctly by our proposed methods and their causes were identified. The three types of methods differed in terms of their favorable and unfavorable abnormal findings. Thus, we could design specific solutions to the issues and show potential improvements in our methods in the future.

## CONFLICT OF INTEREST

The authors confirm that there is no conflict of interest with regard to the content of this article.

## ACKNOWLEDGMENTS

The authors are deeply grateful to Dr. Toshiya Kojima of Department of Cardiovascular Medicine, the Graduate School of Medicine, the University of Tokyo for his assistance and expert knowledge.

## REFERENCES

[1] Eyewitness to history: Landmarks in the development of computerized electrocardiography

[2] Yoshihiko W, Noboru O. An assessment of a diagnostic accuracy for a computerized interpretation of 12-lead electrocardiograms by using the newest version of two representative programs in Japan. JPN. J. ELECTROCARDIOLOGY 2006; 26 (5).

[3] Yeh YC, Wang WJ, Chiou CW. Cardiac arrhythmia diagnosis method using linear discriminant analysis on ECG signals. Measurement 2009; 42: 778-789.

[4] Melgani F, Bazi Y. Classification of Electrocardiogram Signals With Support Vector Machines and Particle Swarm Optimization. IEEE Trans Inf Technol Biomed Sep 2008; 12(5): 667-677.

[5] Asl BM, Setarehdan SK, Mohebbi M. Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal. Artif Intell Med 2008; 44: 51-64.

[6] Nasiri JA, Naghibzadeh M, Yazdi HS, Naghibzadeh B. ECG Arrhythmia Classification with Support Vector Machines and Genetic Algorithm. In Third UKSim European Symposium on Computer Modeling and Simulation; 2009 Nov 25-27; Athens, Greece.

[7] Ganeshkumar R, Kumaraswamy YS. Investigating cardiac arrhythmia in ECG using random forest classification. Int J Comput Appl Jan 2012; 37(4): 31-34

[8] Prasad GK, Sahambi JS. Classification of ECG arrhythmias using multi-resolution analysis and neural networks. IEEE Conf. on Convergent Technologies; 2003; Bangalore, India. (1): 227-231.

[9] Krizhenvsky A, Sutshever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems 2012 Dec 3-6; Lake Tahoe, Nevada. (1): 1097-1105.

[10] Rajpurkar P, Hannun AY, Haghpanahi M, Bourn C, Ng AY. Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. arXiv:1707.01836v1 2017.

[11] Kiranyaz S, Ince T, Gabbouj M. Real-time patient-specific ECG classification by 1-D convolutional neural networks. IEEE T Bio-Med Eng Mar 2016; 63(3): 664-675.

[12] Xiang Y, Luo J, Zhu T, Wang S, Xiang X, Meng J. ECG-Based

Heartbeat Classification Using Two-Level Convolutional Neural Network and RR Interval Difference. Ieice T Inf Syst Apr 2018; 101(4): 1189-1198.

[13] Moody GB, Mark RB. The impact of the MIT-MIH Arrhythmia Database. IEEE Eng Med Biol 2001; 20:45-50.

[14] Lu W, Hou H, Chu J. Feature fusion for imbalanced ECG data analysis. Biomed Signal Proces 2018; 41: 152-160.

[15] Senami F, Takeshi I, Sanshiro I, Katsuhiko F, Kazuhiko O. A study on normal abnormal determination of electrocardiogram waveform using deep learning. SIG-AIMED 2018; 5(5): 1-5.