Classification of Prehospital-Electrocardiograms taken in Ambulance According to Severity using Deep Learning Neural Network

Ryo Oikawa
Graduate School of Software and Information Science, Iwate Prefectural University
152-52 Sugo, Takizawa, Iwate 020-0693, JAPAN
Email: g231t006@s.iwate-pu.ac.jp

Akio Doi
Faculty of Software and Information Science, Iwate Prefectural University152-52 Sugo, Takizawa, Iwate 020-0693, JAPAN
Email: doia@iwate-pu.ac.jp

Basabi Chakraborty
Faculty of Software and Information Science, Iwate Prefectural University
152-52 Sugo, Takizawa, Iwate 020-0693, JAPAN
Email: basabi@iwate-pu.ac.jp

Tomonori Itoh
Division of Cardiology, Department of Internal Medicine, /Division of Community Medicine, Department of Medical Education,Iwate Medical University
2-1-1 Idaidori, Yahaba, Shiwa, Iwate, 028-3695, JAPAN
Email: tomoitoh@iwate-med.ac.jp

Osamu Nishiyama
Iwate Prefectural Ninohe Hospital
38-2 Okawarage, Horino, Ninohe, Iwate, 028-6193, JAPAN
Email: yamari@mopera.net

Abstract— **The prehospital-electrocardiogram (PH-ECG) is an electrocardiogram (ECG) measurement performed by paramedics on a patient suspected of having a myocardial infarction, for example, in an ambulance and the data are transmitted to the hospital. A physician at the hospital can diagnose the condition of a patient based on the transmitted ECG, thus making efficient use of the time before the patient arrives and enabling an early start to treatment. The PH-ECG is particularly useful for patients who require immediate medical attention, such as those with ST-elevation myocardial infarction (STEMI). Multiple studies have shown that PH-ECG improves door-to-balloon time and in-hospital mortality. However, it is necessary to understand the various patterns of abnormal waveforms when analyzing PH-ECG, and it is difficult to make an accurate diagnosis quickly without a cardiologist. In areas where there is a shortage of hospitals and physicians, diagnosis is performed by non-cardiologists, and there is a need for an automated diagnosis system with performance similar to that of cardiologists.**

Recent studies on automated ECG diagnosis have focused on diagnosing specific abnormal findings, especially the classification and discrimination of myocardial infarction and arrhythmias. On the other hand, there are very few studies on the classification of disease severity, independent of the types of abnormal findings. In this work, we analyzed a 12-lead ECG measured in an ambulance using deep learning neural network to classify and evaluate the abnormal waveforms according to degrees of severity. For 88 cases of 12-lead ECG image data measured in the ambulance, each 12-lead waveform was divided into three parts, and 36 one-lead ECGs were extracted. An expert cardiologist annotated each image. The images were labeled in three classes according to the degree of severity, "normal," "mild or moderate," and "severe." Each image was thinned and binarized. Of 3,168 final images, 1,590 were normal waveforms, and 1,578 were abnormal waveforms. 80% of the images were used as training data and 20% of the images were used as test data. A total of 20% of the training data were used as validation data, five-fold cross validation was performed. EfficientnetB0 was used. The model was defined using the network designer in MATLAB. The input image size was set to 224 × 224 pixels, and resizing was performed when no match was found. The optimization method was Adam, and the hyperparameters were set to α = 0.0001, β_1 = 0.9, and β_2 = 0.999. We set the mini-batch size to 64 and the epoch to 100. We could achieve as a kappa coefficient of 0.810 and maximum classification accuracy of 86.6% for the test data. The result indicates the feasibility of an automatic diagnosis system using noisy ECGs measured in ambulances and is expected to provide a new research direction.

Keywords—Prehospital Electrocardiograms, Deep Learning

I. INTRODUCTION

The prehospital electrocardiogram (PH-ECG) is an ECG measurement performed by paramedics on a patient suspected of having a myocardial infarction, etc., and the data is transmitted to the hospital. A physician at the hospital can diagnose a patient's condition based on the ECG transmitted to him or her, thus making the best use of the time before the patient arrives and enabling an early start of treatment. PH-ECG is particularly useful for patients who require immediate medical attention, such as ST-elevation myocardial infarction (STEMI) [1]. Multiple studies have shown that PH-ECG improves door-to-balloon time and in-hospital mortality [1], [2], [3], [4]. However, it is necessary to understand the various patterns of abnormal waveforms when analyzing PH-ECG. It is difficult to make an accurate diagnosis quickly without a cardiologist. In areas where there is a shortage of hospitals and physicians, diagnosis is performed by non-cardiologists, and there is a need for an automated diagnosis system with performance equal to or better than that of cardiologists [5].

Recent studies on automated ECG diagnosis have focused on diagnosing specific abnormal findings, especially the classification and discrimination of myocardial infarction and arrhythmias [6], [7]. On the other hand, there are not enough studies on the classification of disease severity independent of the types of abnormal findings [8]. In this study, we analyzed the 12-lead ECG measured in the ambulance using deep learning to classify and evaluate the three classes of severity. In addition, the ambulance environment is prone to noise contamination due to patient motion and electrode dislodgement. Our main focus is to train and test without excluding data containing such artefacts. An example of an artefact is shown in Fig. 1. In the next section, we briefly describe the related studies, followed by a description of our proposed approach in the next section. In Section IV, we present the experimental results and conclude in the last section.

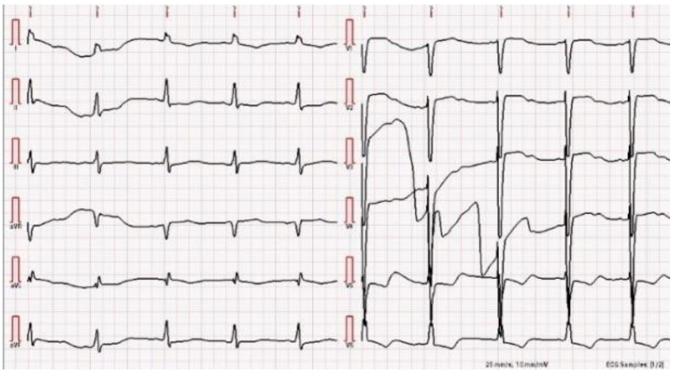


Fig. 1 Wandering artefact due to shaking of electrodes

II. RELATED WORKS

Although there are few studies on the detection of abnormal findings by machine learning using only PH-ECG, Al-Zaiti et al. used a study population of Americans to predict acute coronary syndromes [9]. However, since Simonson [10] reported a statistically quite large difference between Japanese and American ECGs, we used Japanese as the study population in the experiments in the next section.

III. METHODS

Dataset: Large amount of labeled data are needed to train a neural network for the classification of disease severity independent of the type of abnormal finding. For 88 cases of 12-lead ECG image data measured in the ambulance, each 12-lead waveform was divided into three parts, and 36 one-lead ECGs were extracted from one 12-lead ECG. An expert cardiologist annotated each image. They were labeled in three classes according to severity: "normal," "mild or moderate," and "severe." Each image was thinning and binarized. Of the 3,168 final data, 1,590 were normal waveforms, and 1,578 were abnormal waveforms. 80% of the samples were used as training data and 20% as test data for classification experiment In the next section, 80% of these data were used as training data and 20% as test data. A total of 20% of the training data were used as validation data, and cross-validation was performed in five parts. Data were provided by Iwate Prefectural Ninohe Hospital.

Model: EfficientnetB0 [11] was used. The model was defined using the network designer in MATLAB. The network structure of the model is shown in Fig. 2. The input image size was set to 224 × 224 pixels, and resizing was performed when no match was found. The optimization method is Adam, and the hyperparameters are set to α=0.0001, β_1=0.9, β_2=0.999. We set the mini-batch size to 64 and the epoch to 100. As shown in Fig. 3, Loss converged before epoch 100, and there was no improvement in learning accuracy when the epoch was increased above 100. We performed a horizontal and vertical shift of 30 pixels on the training data and used slightly different datasets for each epoch. However, considering the effect of vertical shift on training, we performed the same experiment with the only horizontal shift.

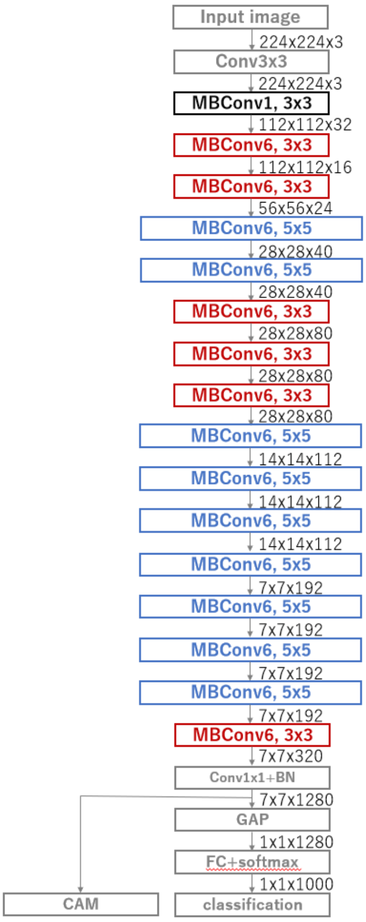


Fig. 2 Structure of convolutional neural network used

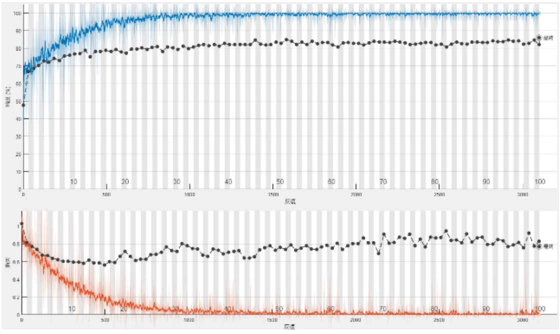


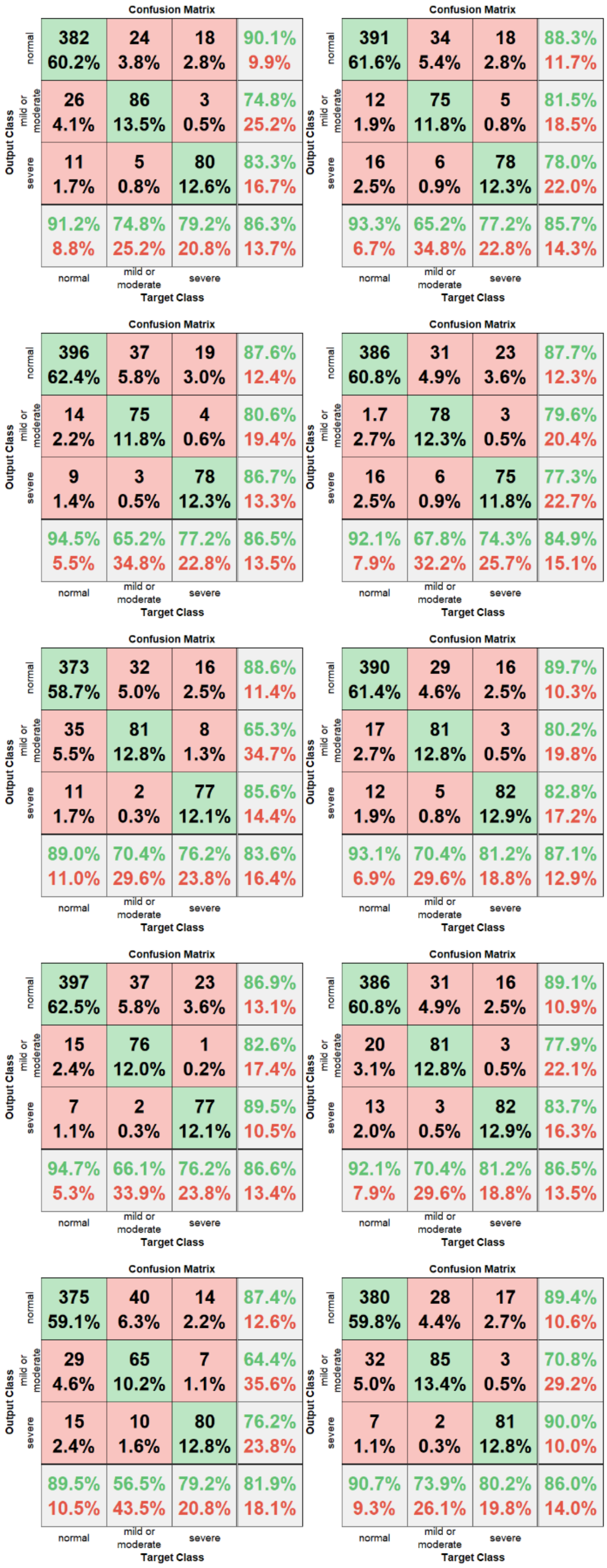
Fig. 3 Learning curve for best fold, blue means accuracy and red means loss.

IV. RESULTS

The accuracy and Kappa coefficients are shown in Table 1. The confusion matrix resulting from the inference on the test data is shown in Fig. 4. Class Activation Map (CAM) is shown in Fig. 5.

Table 1 Accuracy and Kappa coefficient

Fold	Accuracy of experiment (a) (%)	Kappa of Experiment (a)	Accuracy of Experiment (b) (%)	Kappa of Experiment (b)
1	86.3	0.810	85.7	0.780
2	86.5	0.814	84.9	0.746
3	83.6	0.816	87.1	0.820
4	86.6	0.801	86.5	0.814
5	81.9	0.810	86.0	0.748
Avg.	85.0	0.810	86.0	0.782



Confusion Matrix

Output Class \ Target Class	normal	mild or moderate	severe	
normal	382 60.2%	24 3.8%	18 2.8%	90.1% 9.9%
mild or moderate	26 4.1%	86 13.5%	3 0.5%	74.8% 25.2%
severe	11 1.7%	5 0.8%	80 12.6%	83.3% 16.7%
	91.2% 8.8%	74.8% 25.2%	79.2% 20.8%	86.3% 13.7%

Confusion Matrix

Output Class \ Target Class	normal	mild or moderate	severe	
normal	391 61.6%	34 5.4%	18 2.8%	88.3% 11.7%
mild or moderate	12 1.9%	75 11.8%	5 0.8%	81.5% 18.5%
severe	16 2.5%	6 0.9%	78 12.3%	78.0% 22.0%
	93.3% 6.7%	65.2% 34.8%	77.2% 22.8%	85.7% 14.3%

Confusion Matrix

Output Class \ Target Class	normal	mild or moderate	severe	
normal	396 62.4%	37 5.8%	19 3.0%	87.6% 12.4%
mild or moderate	14 2.2%	75 11.8%	4 0.6%	80.6% 19.4%
severe	9 1.4%	3 0.5%	78 12.3%	86.7% 13.3%
	94.5% 5.5%	65.2% 34.8%	77.2% 22.8%	86.5% 13.5%

Confusion Matrix

Output Class \ Target Class	normal	mild or moderate	severe	
normal	386 60.8%	31 4.9%	23 3.6%	87.7% 12.3%
mild or moderate	1.7 2.7%	78 12.3%	3 0.5%	79.6% 20.4%
severe	16 2.5%	6 0.9%	75 11.8%	77.3% 22.7%
	92.1% 7.9%	67.8% 32.2%	74.3% 25.7%	84.9% 15.1%

Confusion Matrix

Output Class \ Target Class	normal	mild or moderate	severe	
normal	373 58.7%	32 5.0%	16 2.5%	88.6% 11.4%
mild or moderate	35 5.5%	81 12.8%	8 1.3%	65.3% 34.7%
severe	11 1.7%	2 0.3%	77 12.1%	85.6% 14.4%
	89.0% 11.0%	70.4% 29.6%	76.2% 23.8%	83.6% 16.4%

Confusion Matrix

Output Class \ Target Class	normal	mild or moderate	severe	
normal	390 61.4%	29 4.6%	16 2.5%	89.7% 10.3%
mild or moderate	17 2.7%	81 12.8%	3 0.5%	80.2% 19.8%
severe	12 1.9%	5 0.8%	82 12.9%	82.8% 17.2%
	93.1% 6.9%	70.4% 29.6%	81.2% 18.8%	87.1% 12.9%

Confusion Matrix

Output Class \ Target Class	normal	mild or moderate	severe	
normal	397 62.5%	37 5.8%	23 3.6%	86.9% 13.1%
mild or moderate	15 2.4%	76 12.0%	1 0.2%	82.6% 17.4%
severe	7 1.1%	2 0.3%	77 12.1%	89.5% 10.5%
	94.7% 5.3%	66.1% 33.9%	76.2% 23.8%	86.6% 13.4%

Confusion Matrix

Output Class \ Target Class	normal	mild or moderate	severe	
normal	386 60.8%	31 4.9%	16 2.5%	89.1% 10.9%
mild or moderate	20 3.1%	81 12.8%	3 0.5%	77.9% 22.1%
severe	13 2.0%	3 0.5%	82 12.9%	83.7% 16.3%
	92.1% 7.9%	70.4% 29.6%	81.2% 18.8%	86.5% 13.5%

Confusion Matrix

Output Class \ Target Class	normal	mild or moderate	severe	
normal	375 59.1%	40 6.3%	14 2.2%	87.4% 12.6%
mild or moderate	29 4.6%	65 10.2%	7 1.1%	64.4% 35.6%
severe	15 2.4%	10 1.6%	80 12.8%	76.2% 23.8%
	89.5% 10.5%	56.5% 43.5%	79.2% 20.8%	81.9% 18.1%

Confusion Matrix

Output Class \ Target Class	normal	mild or moderate	severe	
normal	380 59.8%	28 4.4%	17 2.7%	89.4% 10.6%
mild or moderate	32 5.0%	85 13.4%	3 0.5%	70.8% 29.2%
severe	7 1.1%	2 0.3%	81 12.8%	90.0% 10.0%
	90.7% 9.3%	73.9% 26.1%	80.2% 19.8%	86.0% 14.0%

Fig. 4 Confusion matrix

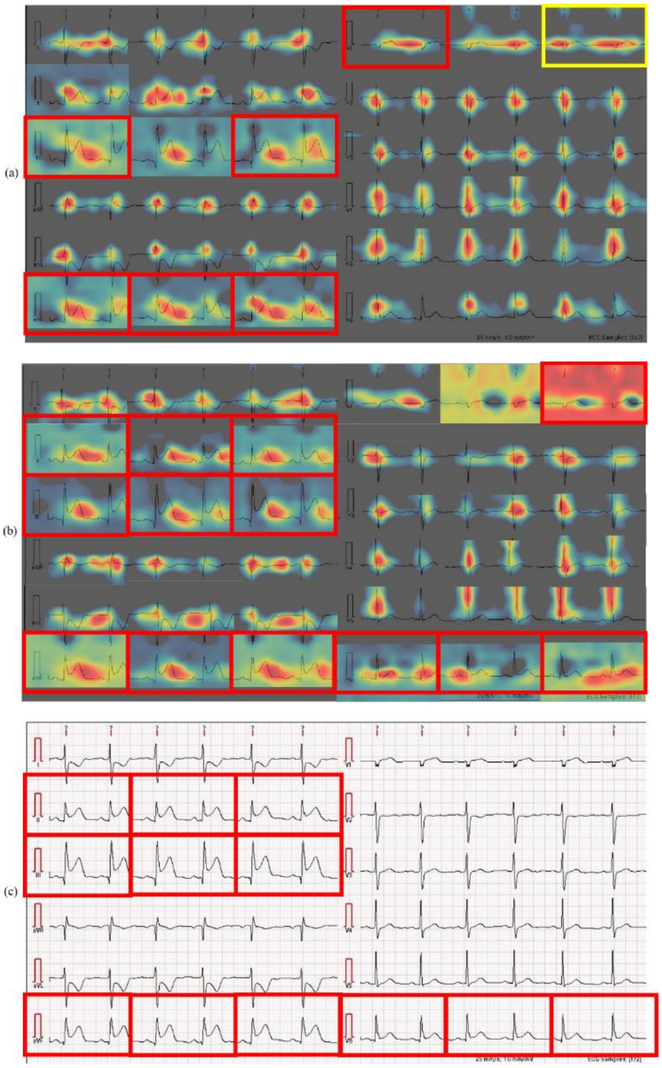


Fig. 5 Inference and CAM on test data: **a** model with vertical data extension, **b** model without vertical data extension, **c** ground-truth

Here, Accuracy means the percentage of the classification results that match the correct labels. The Kappa coefficient is a statistic that describes the degree of agreement between the results of two observers' observations of a phenomenon. As a result, at the best fold (Table 1 experiment (a) Fold 3), the Kappa coefficient between the model and the cardiologist was 0.816. This was in the range of 0.81-1.00, which is judged as "almost perfect agreement" in the guidelines of Landis [12]. At the worst fold, the Kappa coefficient in the external validation was 0.782, which is in the range of 0.61-0.80, which is considered a "fair agreement.

Fig. 5 shows that, in experiment (a), even when the model could focus on the correct location, they often gave wrong answers. The vertical shift in data expansion may have caused the loss of height characteristics, making it difficult to detect findings such as ST elevation and T-wave amplification in the waveform. On the other hand, the average Kappa coefficient was higher in experiment (a). The lack of vertical data expansion may have degraded the generalization performance of experiment (b).

V. Conclusion

In this study, 12-lead ECGs measured in an ambulance were analyzed using deep learning to classify and evaluate the severity of the disease. As a result, inference on the test data resulted in a kappa coefficient of 0.810. In the future, we plan to ensemble the model of this study to classify the severity of the disease in the entire 12-lead PH-ECG.

ACKNOWLEDGMENT

Acknowledgments This research was supported by the "FY2020 Iwate Strategic Research and Development Promotion Project (Feasibility Study Stage)"

REFERENCES

[1] A. Martinoni, S. De Servi, E. Boschetti, R. Zanini, T. Palmerini, A. Politi, G. Musumeci, G. Belli, M. De Paolis, F. Ettori, E. Piccaluga, D. Sangiorgi, A. Repetto, M. D'Urbano, B. Castiglioni, F. Fabbiocchi, M. Onofri, N. De Cesare, G. Sangiorgi, C. Lettieri, F. Poletti, S. Pirelli, and S. Klugmann, (on behalf of the Lombardima Study Group), "Importance and limits of pre-hospital electrocardiogram in patients with ST elevation myocardial infarction undergoing percutaneous coronary angioplasty", European journal of cardiovascular prevention and rehabilitation, Volume 18, Issue 3, 1, pp.526–532, 2011.

[2] J. Nam, K. Caners, J. M. Bowen, M. Welsford, and D. O'Reilly, "Systematic Review and Meta-analysis of the Benefits of Out-of-Hospital 12-Lead ECG and Advance Notification in ST-Segment Elevation Myocardial Infarction Patients", Annals of Emergency Medicine, Vol.64, Issue 2, pp.176-186.e9, 2014.

[3] N. D. Brunetti, G. Di Pietro, A. Aquilino, A. I Bruno, G. Dellegrottaglie, G. Di Giuseppe, C. Lopriore, L. De Gennaro, S. Lanzone, P. Caldarola, G. Antonelli, and M. Di Biase, "Pre-hospital electrocardiogram triage with tele-cardiology support is associated with shorter time-to-balloon and higher rates of timely reperfusion even in rural areas: data from the Bari- Barletta/Andria/Trani public emergency medical service 118 registry on primary angioplasty in ST-elevation myocardial infarction", European Heart Journal. Acute Cardiovascular Care, Vol.3, Issue 3, 1, pp.204–213, 2014.

[4] T. Quinn, S. Johnsen, C. P Gale, H. Snooks, S. McLean, M. Woollard, and C. Weston, (On behalf of the Myocardial Ischaemia National Audit Project (MINAP) Steering Group), "Effects of prehospital 12-lead ECG on processes of care and mortality in acute coronary syndrome: a linked cohort study from the Myocardial Ischaemia National Audit Project", Heart, Vol.100, pp.944-950, 2014.

[5] T. Sakai, O. Nishiyama, M. Onodera, S. Matsuda, S. Wakisawa, M. Nakamura, Y. Morino, and T. Itoh, "Predictive ability and efficacy for shortening door-to-balloon time of a new prehospital electrocardiogram-transmission flow chart in patients with ST-elevation myocardial infarction – Results of the CASSIOPEIA study", Journal of Cardiology, Vol.72, Issue 4, pp.335-342, 2018.

[6] U. Rajendra Acharya, Hamido Fujita, Shu LihOh, Yuki Hagiwara, Jen Hong Tan, and Muhammad Adam, "Application of deep convolutional neuralnetwork for automated detection of myocardial infarction using ECG signals", Information Sciences,Vol.415–416, pp.190-198, 2017.

[7] G. Sannino and G. De Pietro, "A deep learning approach for ECG-based heartbeat classification for arrhythmia detection", Future Generation Computer Systems, Vol.86, pp.446-455, 2018.

[8] S. Furubayashi, T. Imai, S. Ishihara, K. Fujiu, and K. Ohe, "A study on normal abnormal determination of electrocardiogram waveform using deep learning", JSAI Technical Report, Type 2 SIG, pp.05-01-05-05, 2018.

[9] S. Al-Zaiti, L. Besomi, Z. Bouzid, Z. Faramand, S. Frisch, C. Martin-Gill, R. Gregg, S. Saba, C. Callaway, and Ervin Sejdić, "Machine learning-based prediction of acute coronary syndrome using only the pre-hospital 12-lead electrocardiogram", Nat Commun 11, 3966, 2020.

[10] Simonson, E., "Differentiation Between Normal and Abnormal in Electrocardiography", CV St. Louis, Mosby, 1961.

[11] M. Tan and Quoc V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks", International conference on machine learning, pp.6105-6114, 2019.

[12] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data", Biometrics, Vol.33, No. 1, pp.159-174, 1977.