Development of clinically validated artificial intelligence model for detecting ST-segment elevation myocardial infarction

Sang-Hyup Lee

The Graduate School
Yonsei University
Department of Medicine

Development of clinically validated artificial intelligence model for detecting ST-segment elevation myocardial infarction

A Dissertation Submitted
to the Department of Medicine
and the Graduate School of Yonsei University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Medical Science

Sang-Hyup Lee

December 2024

This certifies that the Dissertation of Sang-Hyup Lee is approved

Thesis Supervisor Young-Guk Ko

Thesis Committee Member Jung-Sun Kim

Thesis Committee Member Chul-Min Ahn

Thesis Committee Member Seng Chan You

Thesis Committee Member Junbeom Park

The Graduate School
Yonsei University
December 2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Development of clinically validated artificial intelligence model for detecting ST-segment elevation myocardial infarction

Background and aim: Although the importance of primary percutaneous coronary intervention (PCI) has been emphasized for ST-segment elevation myocardial infarction (STEMI), the appropriateness of the cardiac catheterization laboratory (CCL) activation remains suboptimal. This study aimed to develop a precise artificial intelligence (AI) model for diagnosis of STEMI and accurate CCL activation.

Methods: We used electrocardiography (ECG) waveform data from a prospective PCI registry in Korea in this study. Two independent board-certified cardiologists confirmed the true label of each ECG based on corresponding coronary angiography data. A deep ensemble model was developed by combining five convolutional neural networks. Clinical validation based on a symptom-based ECG dataset, comparisons with clinical physicians, and external validation were performed. Additive benefit on top of the critical pathway for detection of STEMI was evaluated. ECGs were visualized by the Gradient-weighted Class Activation Mapping (Grad-CAM) for assessment of model explainability.

Results: A total of 18,697 ECGs were used for the model development dataset and 1,745 (9.3 %) were STEMI. The AI model achieved an accuracy of 92.1 %, sensitivity of 95.4 % and specificity of 91.8 %. The performances of the AI model were well-balanced and outstanding in the clinical validation, comparison with clinical physicians, and the external validation. The AI model correctly re-classified 31.6 % of patients who incorrectly diagnosed as STEMI.

Conclusions: The deep ensemble AI model showed a well-balanced and outstanding performance. As visualized with the Grad-CAM, the AI model has a reasonable explainability.

Key words : STEMI, electrocardiography, deep ensemble model

1. INTRODUCTION

ST-segment elevation myocardial infarction (STEMI) is a fatal cardiovascular condition usually caused by the obstruction of the coronary artery, characteristically presenting as an elevation of the ST-segment on electrocardiography (ECG). With the advent of reperfusion therapy, particularly primary percutaneous coronary intervention (PCI), mortality rates associated with STEMI have markedly improved. Consequently, timely activation of the cardiac catheterization laboratory (CCL) for primary PCI is of paramount importance, as emphasized by current STEMI management guidelines. [1-3]

However, the diagnostic landscape is fraught with challenges. Various medical conditions, distinct from STEMI, can also manifest as an elevation of the ST-segment on ECG, causing unnecessary activation of the CCL. [4-6] The inappropriate CCL activation is further exacerbated by the misinterpretation of both ECG machines and clinicians. Alarmingly, the misinterpretation rate of STEMI ranges from 14% to 36% and patients with CCL cancellation were reported to be associated with a higher rate of comorbidities and mortality. [7-11] In addition, considering the complications associated with emergent invasive coronary angiography, such as access site bleeding or vascular injury, an accurate differential diagnosis is important, which may affect clinical outcomes. [12] In this context, several artificial intelligence (AI)-based algorithms for detecting STEMI have been developed; however, these studies have several limitations, including a limited number of ECGs, exclusive inclusion of ECGs with normal sinus rhythm, and absence of corresponding coronary angiography results. [13-20]

Hence, we aimed to develop an AI-based model for accurate STEMI diagnosis with clinical relevance in this study.

2. METHOD

2.1 Study Design and Setting

We identified all patients older than 19 years old who underwent PCI at the Severance Hospital (Seoul, South Korea) between 2006 and 2020 from a prospective multicenter PCI registry (Korean

Multicenter Angioplasty Team [KOMATE] registry, NCT03908463). The major exclusion criteria were as follow: 1) absence of coronary angiography data or 2) absence of an adequate ECG performed within 24 hours of PCI. Eligible ECGs were classified into two groups (the STEMI or Not-STEMI groups) at the discretion of two independent board-certified cardiologists.

According to the current guidelines for the management of STEMI, STEMI was identified by at least two contiguous leads with an ST-segment elevation of ≥ 2.5 mm in men aged < 40 years, ≥ 2.0 mm in men aged ≥ 40 years, or ≥ 1.5 mm in women in leads V2–3 and ≥ 1 mm in the other leads. [2,3] ECGs meeting the criteria were determined to be STEMI when a significant coronary artery stenosis that was compatible with the location of ST-segment elevation was detected in coronary angiography. Furthermore, patients in whom an abnormal finding, such as spontaneous coronary artery dissection, resulting flow limitation and coronary stent implantation was identified in coronary angiography were also determined to be STEMI. In case that the decisions from the two cardiologists were not identical, the group allocation was made after further discussion with a third investigator. All ECGs not identified as STEMI were allocated to the Not-STEMI group.

This study was approved by the Institutional Review Board of Yonsei University, which waived the requirement for informed consent owing to the retrospective nature of this study. The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement was followed in this study for ensuring the appropriateness of study design.

2.2 Data Collection

Standard 12-lead ECG waveform records, which had 10 seconds record with a 500 Hz sampling, and the corresponding computer-based interpretation information were extracted from the MUSE Cardiology Information System (GE HealthCare, Chicago, IL, USA) of Yonsei University Health System. We used data from eight leads (Leads I, II, and V1–6), as the data of remaining leads were calculated using a linear combination of those leads according to the nature of the ECG. [21] The detailed process for data specification is presented in **APPENDICES (Eligible Data Specification)**.

The model development dataset consisting of eligible ECGs were divided into three separate datasets (training, internal validation, and test sets). The training and internal validation sets were

randomly constructed from ECGs collected between 2006 and 2019 in a 9:1 ratio, while the test set consisted of ECGs collected in 2020 (**Fig. 1**).

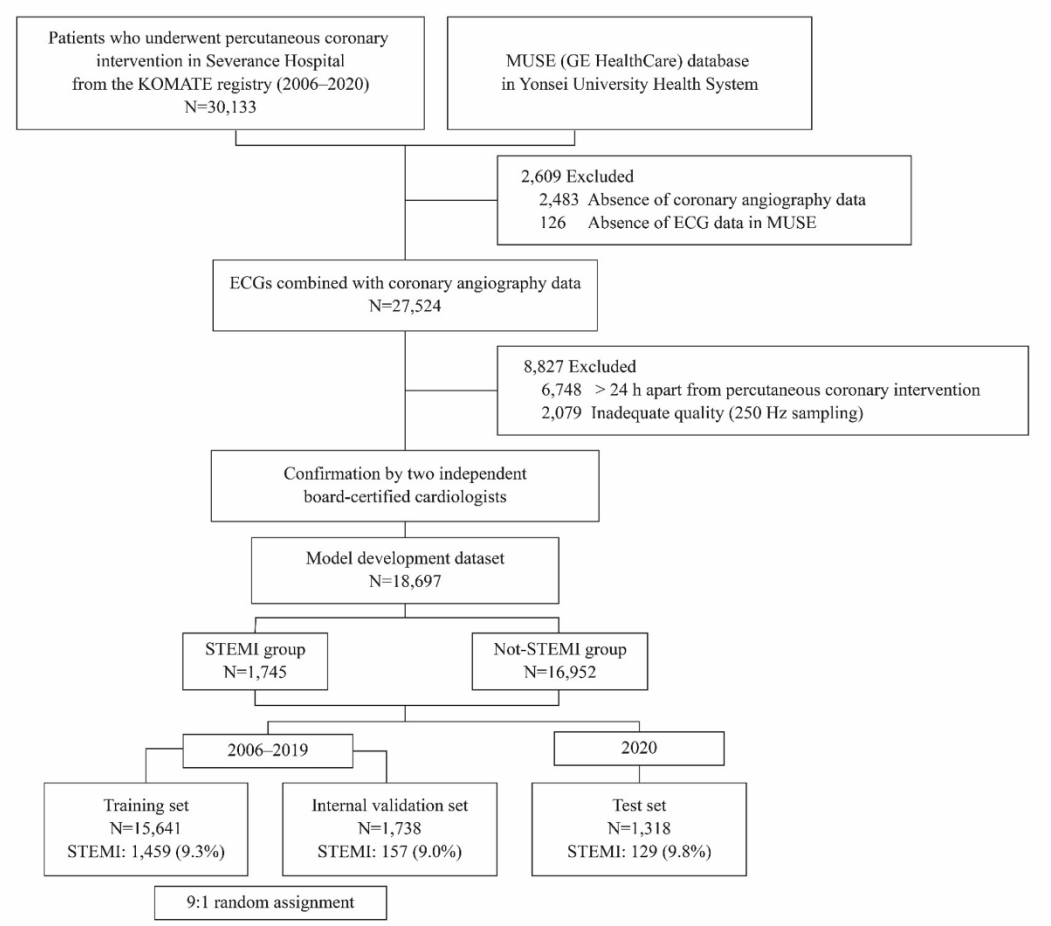

Figure 1. The model development dataset

Note: ECGs from patients who underwent percutaneous coronary intervention in Severance Hospital from the KOMATE registry were included for the model development dataset and were divided into three different datasets (training, internal validation, and test sets). **Abbreviation:** ECG, electrocardiography; KOMATE, Korean Multicenter Angioplasty Team; STEMI, ST-segment elevation myocardial infarction.

2.3 Development and Evaluation of the AI Model

For developing an AI model to classify ECG data into the STEMI or Not-STEMI groups without additional preprocessing, we employed a convolutional neural network (CNN)-based ensemble algorithm. The structure of the CNN was based on a previous model for identifying age and sex based on 12-lead ECG with additional hyperparameters and simplified layers. [22-24] During training, the model minimized Binary Cross Entropy Loss to align its classified outputs with actual labels, adjusting the neural network to learn ECG features for each STEMI class. To address class imbalance, we assigned class-specific weights to the loss function based on STEMI proportion in the training set. The simplest CNN architecture with the highest area under the precision-recall curve (AUPRC) value on the internal validation set was selected. An ensemble model was generated by combining five CNNs, averaging their outputs and determining a cutoff value maximizing Youden's index, which is calculated as (Sensitivity + Specificity - 1). [25,26] This ensemble model produced binary outcomes (1 for STEMI, 0 for Not-STEMI). The detailed process is illustrated in **APPENDICES (Details of Model Development)**.

In this study, we evaluated the model performance in terms of accuracy, sensitivity, and specificity. Furthermore, to evaluate the reliability of the AI models, calibration plots were generated, visually confirming whether the model output scores, ranging from 0 to 1, accurately reflect the proportion of ECGs having true STEMI labels. [27] Alignment with the diagonal reference line on the calibration plot implies the perfect agreement between the model output scores and proportion of true labels, suggesting high reliability of the AI model.

To compensate the black-box phenomenon, which is an inevitable limitation of deep neural network, we generated the Gradient-weighted Class Activation Mapping (Grad-CAM) plots to visualize the explainability of the developed AI model. [28,29] A localization map highlighting the ECG segments of interest to the AI model was presented in the Grad-CAM using a gradient of the final convolutional layers.

2.4 Statistical Analyses

Continuous variables are presented as medians with interquartile ranges, and categorical variables are presented as numbers with percentages. The variables were compared using the Mann–Whitney

U test for continuous variables, and the chi-square or Fisher's exact test was used for categorical variables, as appropriate. For evaluating the interrater reliability of development data, we calculated the Cohen's kappa score based on the decisions of the two cardiologists. [30] The bootstrap resampling method was used to calculate the 95% confidence intervals (CIs) of the area under the receiver operating characteristic curve (AUROC) and AUPRC. [31] The two-tailed P-value was computed and P-value < 0.05 was considered statistically significant. Python (Python Software Foundation) was used for all analyses in this study.

2.5 Performance of Clinical Physicians

In the real-world practice, even though the current commercial ECG machine provides a diagnostic result of STEMI, clinical physicians eventually make decision for diagnosis of STEMI. Therefore, the assessment of the AI model should include a comparison with clinical physicians. In this study, we recruited three second-year residents in training for internal medicine in order to evaluate their STEMI diagnosis performance. We randomly selected a total of 300 different ECGs from the test set while maintaining the prevalence of STEMI. The physicians were blinded to the prediction results from either the AI model or the ECG machine and independently determined whether provided ECGs were STEMI or not.

2.6 Clinical Validation

To evaluate the real-world efficacy of the AI model, we performed a clinical validation using ECGs from patients with chest pain who visited the emergency department of Severance Hospital in 2020, regardless of PCI implementation (**Fig. 2**). The clinical validation set consists of adequate ECGs within 48 hours of emergency department arrival. The true labels for ECGs in the clinical validation set were determined in the same way to the model development dataset. In case that coronary angiography data was not available, the decision was made based on the clinical decision considering the result of other exams, such as computed tomography or echocardiography.

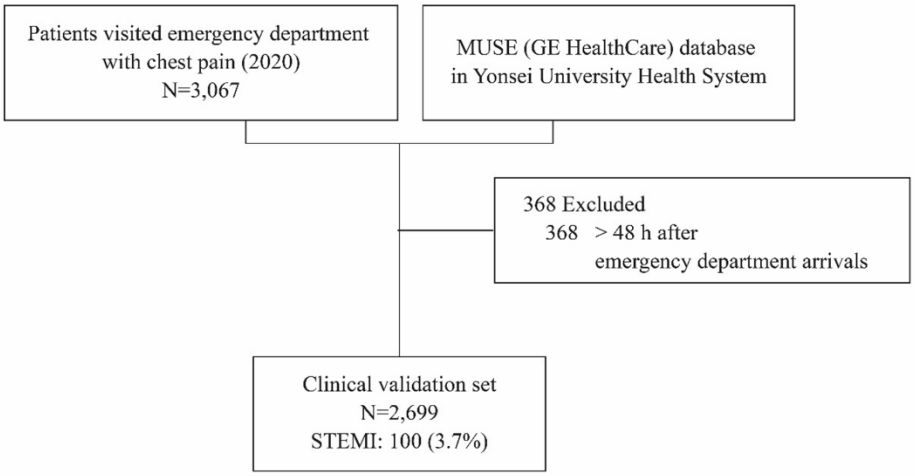

Figure 2. The clinical validation set

Note: Electrocardiography from patients who visited the emergency department with chest pain was included for the clinical validation set. **Abbreviation:** STEMI, ST-segment elevation myocardial infarction.

2.7 Additive Benefit of the AI Model

The eventual aim of the AI model is to support clinical physicians for precise diagnosis of STEMI. For evaluating the additive benefit of the AI model, we assessed the expected changes in clinical decision resulted from the AI model using a critical pathway (CP) cohort. In Severance Hospital, there is a CP for the rapid diagnosis of STEMI and subsequent activation of CCL. The CP is activated by an agreement of two physicians in the emergency department. For this analysis, we analyzed patients in whom the CP was activated from 2007 to 2020 (**Fig. 3**). The true STEMI labels were reviewed in the same way to the clinical validation set. Then, the proportions of patients in whom the decisions were changed by the AI model were evaluated.

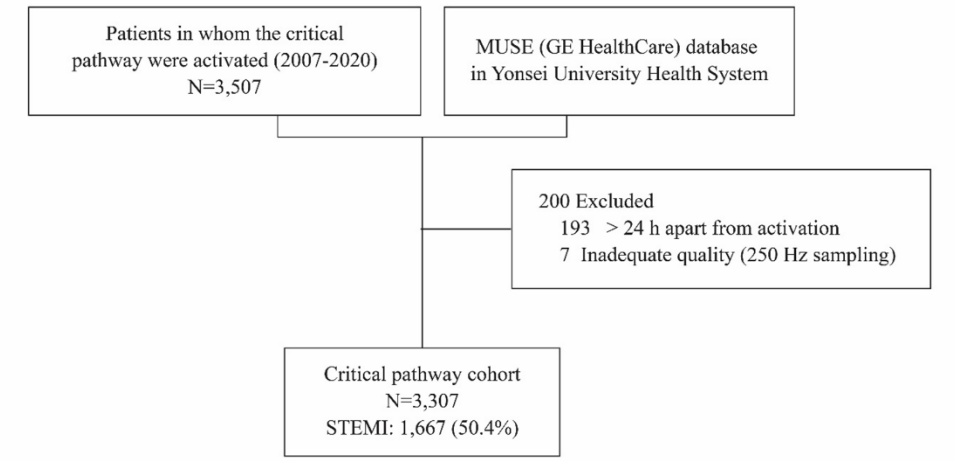

Figure 3. The critical pathway cohort

Note: Electrocardiography from patients in whom the critical pathway was activated were included for the critical pathway cohort. **Abbreviation:** STEMI, ST-segment elevation myocardial infarction.

2.8 External Validation

For external validation of the AI model, we used a publicly available ECG dataset in this study, which is the PTB-XL. [32,33] As validated in the previous study using the PTB-XL, ECGs annotated to acute myocardial infarction in the PTB-XL were reviewed by the two cardiologists and those that were considered STEMI were included in the external validation set. [34]

3. RESULTS

3.1 Baseline Characteristics of Study Population

The CONSORT diagram for the model development dataset is shown in **Fig. 1**. Among the 30,133 PCIs performed in Severance Hospital from the KOMATE registry, 2,609 ECGs were excluded because of the absence of either coronary angiography or ECG data. In addition, 8,827 ECGs were excluded because they were performed > 24 h apart from the PCI or had inadequate quality for model development. Finally, 18,697 ECGs were eligible, of which 1,745 (9.3 %) were classified as STEMI (**Fig. 1**). For the model development dataset, the Cohen's kappa score was 0.848 (**Table 1**).

Table 1. Interrater agreement for the model development dataset

		Rater 1		
		STEMI	**Not-STEMI**	**Total**
Rater 2	**STEMI**	1,655	222	1,877
	Not-STEMI	299	16,521	16,820
	Total	1,954	16,743	18,697

Note: The decisions by the two cardiologists for the model development dataset are presented. The Cohen's kappa score is 0.848 accordingly. **Abbreviation:** STEMI, ST-segment elevation myocardial infarction.

The baseline characteristics for the model development dataset is presented in **Table 2**. ECGs included in the STEMI group were associated with patients who were more likely to be younger (64 years vs. 66 years), male (78.3 % vs. 72.4 %), have a lower body mass index (24.2 kg/m^2 vs. 24.4 kg/m^2), and be current smokers (35.0 % vs. 19.4 %) than those included in the Not-STEMI group. The STEMI group was associated with lower proportions of comorbidities, such as hypertension (51.9 % vs. 66.0 %), diabetes (29.2 % vs. 37.0 %), and dyslipidemia (54.6 % vs. 76.0 %), than the Not-STEMI group. The proportions of patients with prior PCI (15.5 % vs. 27.6 %) and coronary artery bypass graft surgery (0.6 % vs. 3.4 %) were lower in the STEMI group than in the Not-STEMI group. Plasma hemoglobin (14.4 mg/dL vs. 13.7 mg/dL), platelet count (244,000 /μL vs. 225,000 /μL), and serum creatinine (0.99 mg/dL vs. 0.91 mg/dL) levels were higher in the STEMI group than in the Not-STEMI group.

Table 2. Baseline characteristics for the model development dataset

	STEMI (N=1,745)	Not-STEMI (N=16,952)	P-value
Age, yr.	64 (53–72)	66 (58–73)	<0.001
Male	1,367 (78.3)	12,273 (72.4)	<0.001
Body mass index, kg/m^2	24.2 (22.3–26.3)	24.4 (22.6–26.4)	0.002
Hypertension	905 (51.9)	11,177 (66.0)	<0.001

Diabetes	510 (29.2)	6,266 (37.0)	<0.001
Dyslipidemia	952 (54.6)	12,870 (76.0)	<0.001
Atrial fibrillation	89 (5.1)	700 (4.1)	0.063
Current smoker	611 (35.0)	3,297 (19.4)	<0.001
Prior PCI	270 (15.5)	4,680 (27.6)	<0.001
Prior CABG	11 (0.6)	583 (3.4)	<0.001
Prior MI	181 (10.4)	1,386 (8.2)	0.002
Pacemaker	1 (0.1)	64 (0.4)	0.051
Bundle branch block			0.002
LBBB	32 (1.8)	162 (1.0)	
RBBB	85 (4.9)	740 (4.4)	
None	1,627 (93.3)	15,986 (94.7)	
Hemoglobin, g/dL	14.4 (13.0–15.6)	13.7 (12.4–14.9)	<0.001
Platelet count, /μL	244,000 (197,000–297,000)	225,000 (188,000–268,000)	<0.001
Serum creatinine, mg/dL	0.99 (0.84–1.20)	0.91 (0.78–1.09)	<0.001

Note: Data are presented as medians (interquartile ranges) or numbers (%). **Abbreviation:** CABG, coronary artery bypass graft; ECG, electrocardiography; LBBB, left bundle branch block; MI, myocardial infarction; PCI, percutaneous coronary intervention; RBBB, right bundle branch block; STEMI, ST-segment elevation myocardial infarction.

3.2 Performance Assessment of the AI Model

For the deep ensemble, the cut-off value of the model output scores was optimized at 0.0768 by which the Youden's index was maximized (**Fig. 4**). While the deep ensemble outperformed the single neural network in terms of AUROC (0.979 [0.969–0.988] vs. 0.973 [0.960–0.984]; P=0.007), the deep ensemble showed an AUPRC value comparable to that of the single neural network (0.870 [0.817–0.914] vs. 0.850 [0.796–0.898]; P=0.062) (**Table 3**). The calibration plots of the single neural network and deep ensemble models are presented in **Fig. 5**.

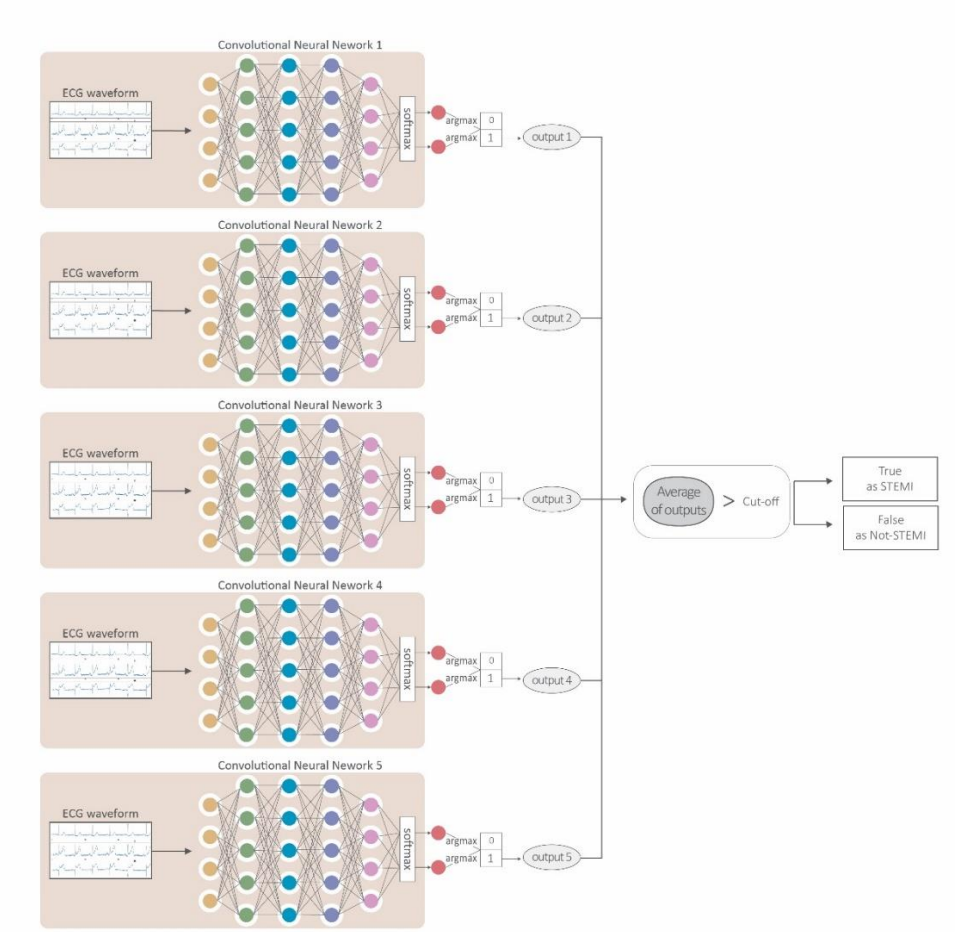

Figure 4. Model architecture of the deep ensemble model

Note: The architecture of the deep ensemble model consisted of five single convolutional neural networks, each trained using the same architecture with different random seeds and dropout. Each individual network used ECG waveforms as inputs, recorded at 500 Hz for 10 s with eight leads. After passing through the convolutional neural network, the softmax function generates outputs, ranging from 0 to 1, for the two labels, 1 for 'STEMI', 0 for 'Not-STEMI'. The ensemble averaged the outputs across five networks as model output score, and then compared them with a predetermined cut-off to classify the presence of STEMI. **Abbreviation:** ECG, electrocardiography; STEMI, ST-segment elevation myocardial infarction.

Table 3. Comparison between the single neural network and deep ensemble model for the internal validation set

	Single neural network	Deep ensemble	P-value
AUROC	0.973 (0.960–0.984)	0.979 (0.969–0.988)	0.007
AUPRC	0.850 (0.796–0.898)	0.870 (0.817–0.914)	0.062

Note: The values of AUROC and AUPRC with 95% confidence intervals are presented. P-value indicates comparisons between the single neural network and deep ensemble models. **Abbreviation:** AUPRC, area under the precision-recall curve; AUROC, area under the receiver operating characteristic curve.

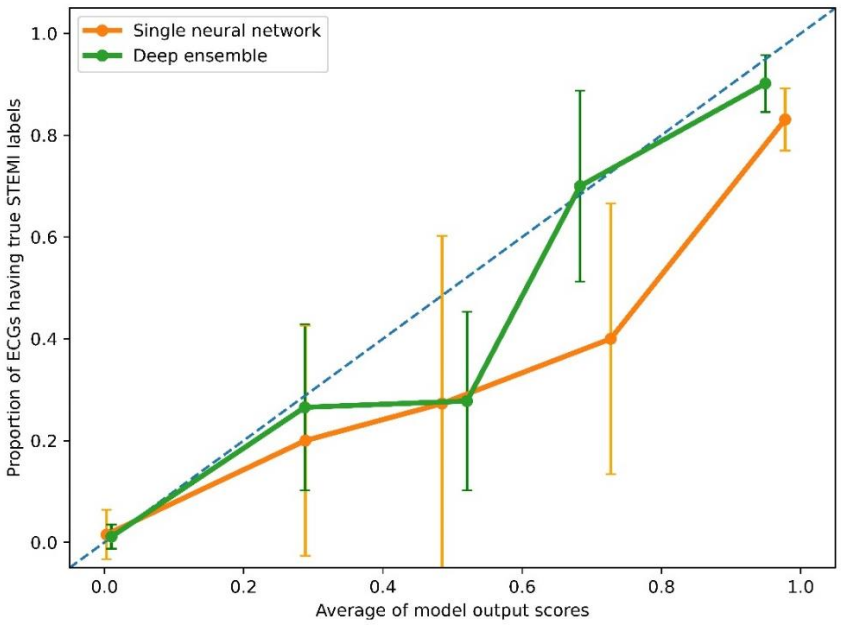

Figure 5. Calibration plots for the single neural network and deep ensemble models

Note: For the internal validation set, the proportions of ECGs having true STEMI labels are presented with 95% confidence intervals according to the stratified model output scores. The I bars indicate the 95% confidence intervals of the proportions. The deep ensemble model had better alignment with the diagonal reference line (dashed line) than did the single neural network model. **Abbreviation:** ECG, electrocardiography; STEMI, ST-segment elevation myocardial infarction.

3.3 Performance Comparison between the AI Model, ECG Machine, and Clinical Physicians

The AI model had AUROC of 0.981 and AUPRC of 0.913 in the test set while achieving an accuracy of 92.1 %, sensitivity of 95.4 %, and specificity of 91.8 % (Youden's index 0.872) (**Fig. 6** and **Table 4**). Meanwhile, the commercial ECG machine achieved an accuracy of 94.6 %, sensitivity of 60.5 %, and specificity of 98.3 % (Youden's index 0.588). Clinical physicians achieved an accuracy of 79.6 %, sensitivity of 81.1 %, and specificity of 77.4 % in average (Youden's index 0.585) (**Fig. 6** and **Table 4**).

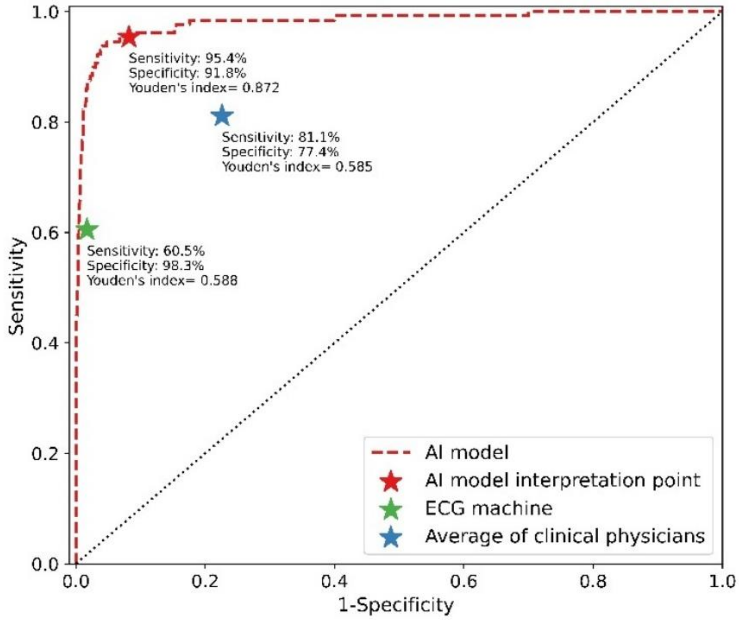

Figure 6. The performances of the AI model, ECG machine algorithm, and clinical physicians
Note: The receiver operating characteristic curve and the selected interpretation point of the AI model are presented (Red line and red star) with the performance of the ECG machine algorithm (Green star) and clinical physicians (Blue star). **Abbreviation:** AI, artificial intelligence; ECG, electrocardiography.

Table 4. Performances of the AI model, the ECG machine algorithm, and the clinical physicians

		Accuracy	Sensitivity	Specificity
Test set	AI model	92.1 %	95.4 %	91.8 %
	ECG machine	94.6 %	60.5 %	98.3 %
	Clinical physicians	79.6 %	81.1 %	77.4 %
Clinical validation	AI model	89.3 %	95.0 %	89.1 %
	ECG machine	96.9 %	60.0 %	98.3 %
External validation	AI model	97.6 %	83.3 %	97.9 %

Note: The performances of three different diagnostic systems were presented. The performances of clinical physicians were assessed on a dataset consisting of 300 ECGs randomly selected from the test set. The external validation was performed using the PTB-XL dataset. **Abbreviation:** AI, artificial intelligence; ECG, electrocardiography.

3.4 The Grad-CAM

Examples of Grad-CAM for ECGs accurately predicted as STEMI or Not-STEMI are presented in **Fig. 7**, highlighting ECG segments with higher contribution and relevance to the model's predictive performance.

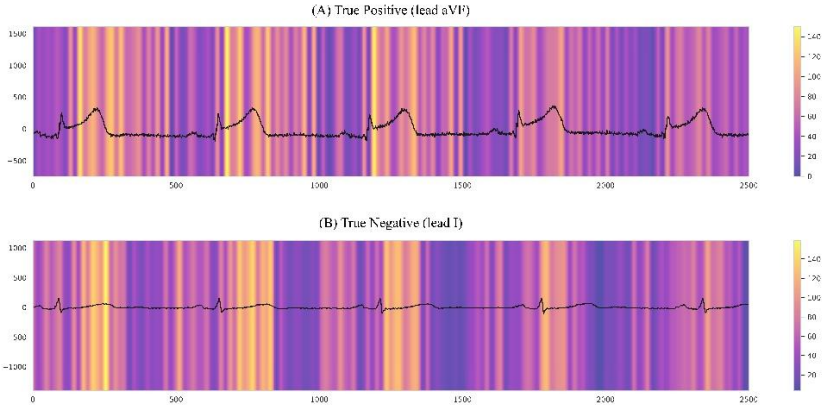

Figure 7. Examples of Grad-CAM

Note: (A) True positive and **(B)** true negative examples of Grad-CAM are presented. The degree of contribution of each ECG segment is highlighted in the Grad-CAM. ST and T segments are mainly considered a recognition feature as highlighted. **Abbreviation:** ECG, electrocardiography; Grad-CAM, gradient-weighted class activation mapping.

3.5 Clinical Validation

A total of 2,699 ECGs, which were 5.2 % of the total patients visited emergency department in 2020, were included in the clinical validation set and the prevalence of STEMI was 3.7 % among them (**Fig. 2**). The baseline characteristics of patients in the clinical validation set are presented in **Table 5**. Among 2,599 Not-STEMI patients in the clinical validation set, 283 (10.9 %) patients were identified as STEMI by the AI model. The final diagnoses for those patients are presented in **Table 6**.

Table 5. Baseline characteristics of the clinical validation set

	STEMI (N=100)	Not-STEMI (N=2,599)	P-value
Age, yr.	63 (54–70)	60 (44–72)	0.031
Male	79 (79.0)	1,458 (56.1)	<0.001
Body mass index, kg/m^2	25.1 (22.9–27.5)	23.9 (21.9–26.4)	0.005
Hypertension	21 (21.0)	908 (34.9)	0.006
Diabetes	13 (13.0)	489 (18.8)	0.182
Dyslipidemia	13 (13.0)	650 (25.0)	0.009
Atrial fibrillation	4 (4.0)	263 (10.1)	0.066
Current smoker	30 (30.0)	98 (3.8)	<0.001
Prior PCI	10 (10.0)	291 (11.2)	0.833
Prior CABG	0 (0.0)	7 (0.3)	>0.999
Prior MI	7 (7.0)	186 (7.2)	>0.999
Pacemaker	1 (1.0)	19 (0.7)	0.531
Bundle branch block			0.745
LBBB	2 (2.0)	35 (1.4)	
RBBB	3 (3.0)	101 (3.9)	

None	94 (94.9)	2,444 (94.7)	
Hemoglobin, g/dL	13.9 (12.7–14.9)	13.9 (12.8–15.0)	0.776
Platelet count, /μL	231,000 (191,000–263,000)	242,000 (202,000–284,000)	0.112
Serum creatinine, mg/dL	0.80 (0.72–1.04)	0.81 (0.68–0.95)	0.530

Note: Data are presented as medians (interquartile ranges) or numbers (%). **Abbreviation:** CABG, coronary artery bypass graft; ECG, electrocardiography; LBBB, left bundle branch block; MI, myocardial infarction; PCI, percutaneous coronary intervention; RBBB, right bundle branch block; STEMI, ST-segment elevation myocardial infarction.

Table 6. Final diagnosis in patients with false-positive prediction in the clinical validation set

Category of final diagnosis	Number of patients
Cardiac cause	**166**
NSTE-ACS	80
Arrhythmia	43
Cardiomyopathy	22
Perimyocarditis	11
Other cardiac	10
Extracardiac cause	**77**
Gastrointestinal disorder	37
Chest wall pain	23
Pleuritic pain	17
Systemic cause	**40**
Hyperventilation or panic disorder	20
Ethanol intoxication	12
Sepsis	5
Contrast allergy	2
Seizure	1
Total	**283**

Abbreviation: NSTE-ACS, non-ST-segment elevation acute coronary syndrome

For the clinical validation, the AI model achieved an accuracy of 89.3 %, a sensitivity of 95.0 %, and a specificity of 89.1 %, while AUROC was 0.978 (0.959–0.992), and AUPRC was 0.808 (0.718–0.893) (**Table 4** and **Fig. 8**). Meanwhile, the ECG machine achieved an accuracy of 96.9 %, sensitivity of 60.0 %, and specificity of 98.3 % in the clinical validation set (**Table 4**).

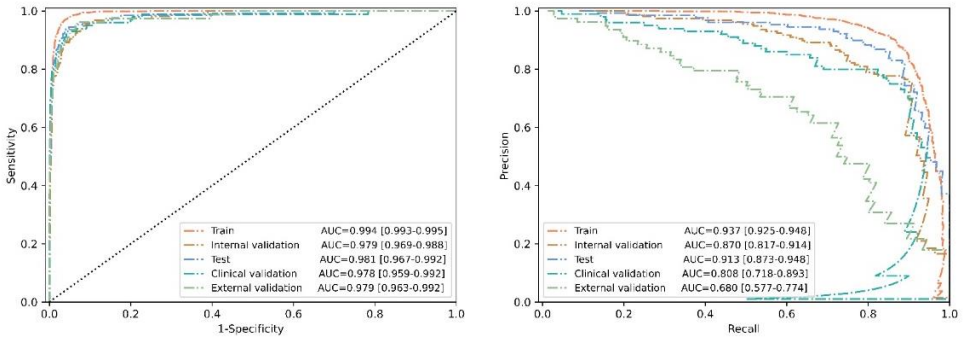

Figure 8. Receiver operating characteristic and precision-recall curves of the developed model
Note: (A) Receiver operating characteristic and **(B)** precision-recall curves are presented with the AUROC, AUPRC values and 95% confidence intervals. **Abbreviation:** AUPRC, area under the precision-recall curve; AUROC, area under the receiver operating characteristic curve

3.6 Additive Benefit in the Critical Pathway Cohort

A total of 3,307 ECGs from 2007 to 2020 were included in the CP cohort (**Fig. 3**). Among them, 1,640 (49.6 %) ECGs were identified to be Not-STEMI. Meanwhile, the AI model reclassified 518 patients, in whom the CP was inappropriately activated, to Not-STEMI, while 34 patients with true STEMI labels were reclassified to Not-STEMI (**Table 7**).

Table 7. The confusion matrix for the AI model for the CP cohort

		True label		
		STEMI	**Not-STEMI**	**Total**
AI model prediction	**STEMI**	1,633	1,122	2,755
	Not-STEMI	34	518	552
	Total	1,667	1,640	3,307

Note: The prediction of the AI model for the CP cohort is presented. Among 1,640 patients in whom the CP were inappropriately activated, 518 (31.6 %) patients were reclassified to Not-STEMI. **Abbreviation:** AI, artificial intelligence; CP, critical pathway; STEMI, ST-segment elevation myocardial infarction.

3.7 External Validation

For the external validation set consisted of 5,991 ECGs with "Normal" annotation and 79 STEMI ECGs, the AI model achieved an accuracy of 97.6 %, sensitivity of 83.3 %, and specificity of 97.9 % (**Table 4**). Meanwhile, AUROC of the AI model was 0.979 (0.963–0.992) and AUPRC was 0.680 (0.577–0.774) in the external validation set (**Fig. 8**).

4. DISCUSSION

The AI model provided in this study based on the 12-lead ECG has several advantages over previous models. First, the AI model was trained and validated using a set of ECGs combined with real-world coronary angiography information to guarantee the accuracy of the true values. Second, a deep ensemble model was developed for a higher performance than that of a single model. Third, the clinical validation using a symptom-based ECG data and the external validation were performed to minimize the selection bias and the overfitting issue of the AI model. Fourth, the Grad-CAM implied the explainability of the AI model and compensated the black box phenomenon. Last, the possible additional benefit from the AI model was presented in this study using the CP cohort.

To develop an accurate AI model and avoid verification bias, an accurate true diagnosis should be guaranteed. Therefore, previous AI models for STEMI diagnosis had their own confirmation processes for true diagnoses by cardiologists or trained experts. [13-20] Similarly in the current study, two independent attending cardiologists confirmed the diagnosis using the coronary angiography data corresponding to each ECG. Furthermore, as the Cohen's kappa score was over 0.8 and a senior cardiologist was involved in case of disagreement, the accuracy of the ECGs used for the development of the AI model was highly guaranteed.

The AUROC value is inappropriate for assessing model performance, as it often leads to the overestimation of model performance in an imbalanced dataset. [35] Previous studies have primarily evaluated model performance using only the AUROC, along with corresponding sensitivity and specificity values. However, to address this limitation, we used the AUPRC, which represents a different trade-off between precision and recall, as a primary metric for model development as the prevalence of STEMI was less than 10% in the dataset. [14-16,18,19] Furthermore, because an overfitted model can make incorrect decisions for data not represented in the training set, we expanded the AI model to an ensemble of five independently trained neural networks. This decision was made because the deep ensemble model has been reported to mitigate overfitting of the AI model compared to a single model. [21,36] Consequently, the calibration plot showed that the deep ensemble model had better alignment with the diagonal reference line compared to the single neural network. This improvement indicated enhanced calibration of the deep ensemble model compared to the single neural network.

In this study, we compared the developed AI model with the ECG machine algorithm and the clinical physicians. Although the ECG machine algorithm showed an excellent accuracy and specificity, the sensitivity was not acceptable. Meanwhile, the clinical physicians had a higher sensitivity and a lower specificity compared to the ECG machine algorithm. In contrary, the AI model achieved well-balanced performance. As there is always a trade-off between sensitivity and specificity, the Youden's index is usually measured as an overall metric for estimation of the AI model. [26] In this study, the AI model achieved the Youden's index of 0.872, while the ECG machine algorithm achieved 0.588, and the clinical physicians achieved 0.585, which implies the AI model was evenly excellent algorithm for diagnosis of STEMI.

As the training data was limited in ECGs from patients who underwent PCI, the excellent performance from the data constructed for model development did not guarantee the performance from the real-world data. In this regard, we performed the clinical validation to evaluate predictive performance of the AI model in the data representing real-world practice. Even though the performance indices of the AI model were lower in the clinical validation set than in the test set, the values were still well-balanced compared to that from the ECG machine algorithm. Considering the difference in clinical characteristics between the training data and the clinical validation set, the clinical validation implies the probability for expanding the AI model to the real-world practice. Furthermore, the benefit of the AI model on top of previous clinical practice, which was represented by the CP, was elucidated in this study. As approximately a half of the decision by the CP was incorrect and induced inappropriate CCL activation, the AI model identified 31.6 % (518/1,640) of patients to be correctly reclassified to Not-STEMI, by which the AI model could prevent inappropriate CCL activation. Regarding the possibility of benefit reported in this study, further prospective validation should follow to elucidate real-world benefit of the AI model.

As the detailed logical process is not elucidated, one of the most critical hurdles is the explainability of the result from the AI model. As observed in the Grad-CAM, the AI model recognizes a feature including the ST and T segments for identifying ECG, which is similar to what clinical physicians concentrate on based on the current STEMI guidelines. The model explainability is essential for achievement of clinical adherence for the AI model. Without the explainable rationale of the AI model, the clinical physicians might hesitate to accept the result from the AI model, resulting limitation of application to the real-world practice. With this regard, this study could support the further application of the AI model by providing the model explainability with the Grad-CAM.

5. CONCLUSION

In conclusion, the developed deep ensemble model for the diagnosis of STEMI achieved outstanding and well-balanced performance in both a PCI registry and a symptom-based ECG set. The Grad-CAM also enhanced the explainability of the AI model and its alignment with real-world

practice. Further studies with prospective validation regarding clinical benefit in a real-world setting should be warranted.

REFERENCES

1. Puymirat E, Simon T, Steg PG, et al. Association of changes in clinical characteristics and management with improvement in survival among patients with ST-elevation myocardial infarction. *JAMA*. 2012;308:998–1006.
2. O'Gara PT, Kushner FG, Ascheim DD, et al. 2013 ACCF/AHA guideline for the management of ST-elevation myocardial infarction: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. *Circulation*. 2013;127:e362–425.
3. Ibanez B, James S, Agewall S, et al. 2017 ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation: The Task Force for the management of acute myocardial infarction in patients presenting with ST-segment elevation of the European Society of Cardiology (ESC). *Eur Heart J*. 2018;39:119–177.
4. Wang K, Asinger RW, Marriott HJ. ST-segment elevation in conditions other than acute myocardial infarction. *N Engl J Med*. 2003;349:2128–2135.
5. Jeong HC, Ahn Y. False Positive ST-Segment Elevation Myocardial Infarction. *Korean Circ J*. 2013;43:368–369.
6. Thygesen K, Alpert JS, Jaffe AS, et al. Fourth Universal Definition of Myocardial Infarction (2018). *Circulation*. 2018;138:e618–e651.
7. Larson DM, Menssen KM, Sharkey SW, et al. "False-positive" cardiac catheterization laboratory activation among patients with suspected ST-segment elevation myocardial infarction. *JAMA*. 2007;298:2754–2760.
8. Kontos MC, Kurz MC, Roberts CS, et al. An evaluation of the accuracy of emergency physician activation of the cardiac catheterization laboratory for patients with suspected ST-segment elevation myocardial infarction. *Ann Emerg Med*. 2010;55:423–430.
9. McCabe JM, Armstrong EJ, Kulkarni A, et al. Prevalence and factors associated with false-positive ST-segment elevation myocardial infarction diagnoses at primary percutaneous coronary intervention–capable centers: a report from the Activate-SF registry. *Arch Intern Med*. 2012;172:864–871.
10. McCabe JM, Armstrong EJ, Ku I, et al. Physician accuracy in interpreting potential ST-segment elevation myocardial infarction electrocardiograms. *J Am Heart Assoc*.

2013;2:e000268.

11. Lange DC, Conte S, Pappas-Block E, et al. Cancellation of the Cardiac Catheterization Lab After Activation for ST-Segment-Elevation Myocardial Infarction. *Circ Cardiovasc Qual Outcomes*. 2018;11:e004464.
12. Al-Hijji MA, Lennon RJ, Gulati R, et al. Safety and Risk of Major Complications With Diagnostic Cardiac Catheterization. *Circ Cardiovasc Interv*. 2019;12:e007791.
13. Kavak S, Chiu XD, Yen SJ, et al. Application of CNN for Detection and Localization of STEMI Using 12-Lead ECG Images. *IEEE Access*. 2022;10:38923–38930.
14. Al-Zaiti S, Besomi L, Bouzid Z, et al. Machine learning-based prediction of acute coronary syndrome using only the pre-hospital 12-lead electrocardiogram. *Nat Commun*. 2020;11:3966.
15. Zhao Y, Xiong J, Hou Y, et al. Early detection of ST-segment elevated myocardial infarction by artificial intelligence with 12-lead electrocardiogram. *Int J Cardiol*. 2020;317:223–230.
16. Chang KC, Hsieh PH, Wu MY, et al. Usefulness of multi-labelling artificial intelligence in detecting rhythm disorders and acute ST-elevation myocardial infarction on 12-lead electrocardiogram. *Eur Heart J Digit Health*. 2021;2:299–310.
17. Liu WC, Lin CS, Tsai CS, et al. A deep learning algorithm for detecting acute myocardial infarction. *EuroIntervention*. 2021;17:765–773.
18. Choi HY, Kim W, Kang GH, et al. Diagnostic Accuracy of the Deep Learning Model for the Detection of ST Elevation Myocardial Infarction on Electrocardiogram. *J Pers Med*. 2022;12:336.
19. Gibson CM, Mehta S, Ceschim MRS, et al. Evolution of single-lead ECG for STEMI detection using a deep learning approach. *Int J Cardiol*. 2022;346:47–52.
20. Wu L, Huang G, Yu X, et al. Deep Learning Networks Accurately Detect ST-Segment Elevation Myocardial Infarction and Culprit Vessel. *Front Cardiovasc Med*. 2022;9:797207.
21. Gustafsson S, Gedon D, Lampa E, et al. Development and validation of deep learning ECG-based prediction of myocardial infarction in emergency department patients. *Sci Rep*. 2022;12:19615.
22. Attia ZI, Friedman PA, Noseworthy PA, et al. Age and Sex Estimation Using Artificial Intelligence From Standard 12-Lead ECGs. *Circ Arrhythm Electrophysiol*. 2019;12:e007284.

23. Ko WY, Siontis KC, Attia ZI, et al. Detection of Hypertrophic Cardiomyopathy Using a Convolutional Neural Network-Enabled Electrocardiogram. *J Am Coll Cardiol*. 2020;75:722–733.
24. Ribeiro AH, Ribeiro MH, Paixão GMM, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat Commun*. 2020;11:1760.
25. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*. 2017;30:6405–6416.
26. Ruopp MD, Perkins NJ, Whitcomb BW, et al. Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biom J*. 2008;50:419–430.
27. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. *Proceedings of the 22nd international conference on Machine learning.* 2005:625–632.
28. Selvaraju RR, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision*. 2017:618–626.
29. Arrieta AB, Díaz-Rodríguez N, Del Ser J, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*. 2020;58:82–115.
30. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*. 2012;22:276–282.
31. DiCiccio TJ, Efron B. Bootstrap confidence intervals. *Statistical science*. 1996;11:189–228.
32. Wagner P, Strodthoff N, Bousseljot R-D, et al. PTB-XL, a large publicly available electrocardiography dataset. *Scientific data*. 2020;7:1–15.
33. Strodthoff N, Mehari T, Nagel C, et al. PTB-XL+, a comprehensive electrocardiographic feature dataset. *Scientific data*. 2023;10:279.
34. Gustafsson S, Gedon D, Lampa E, et al. Development and validation of deep learning ECG-based prediction of myocardial infarction in emergency department patients. *Scientific Reports*. 2022;12:19615.
35. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning.* 2006:233–240.

36. Mohammed A, Kora R. A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University-Computer and Information Sciences*. 2023;35:757–774.

APPENDICES

1. Eligible Data Specification

In this section, we specify the requirements for the electrocardiography (ECG) waveform data utilized in our model. The resting 12-lead ECG waveforms were collected from both emergency department, outpatient, and inpatient settings at Severance Hospital, Seoul, South Korea, using the MUSE Cardiology Information System (GE HealthCare, Chicago, IL, USA). ECG data points were subjected to both a high-pass filter (cut-off frequency, 0.16 Hz) and a low-pass filter (cut-off frequency, 150 Hz). An alternating current filter with a frequency of 60 Hz was applied. The standard 12-lead ECG data were recorded at a 500 Hz sampling rate for 10 seconds, resulting in 5,000 samples, with amplitude values expressed in microvolts per lead. Notably, no further preprocessing was conducted, and data containing noise were not excluded from the analysis.

2. Details of Model Development

2.1 Input of the Model

Each input datum was transformed into an 8 × 5000 matrix, where the first dimension represents spatial positions over leads, and the second represents temporal dimensions.

2.2 Rationale of the Convolutional Neural Network (CNN)

We utilized a CNN implemented using the Keras Framework with a TensorFlow backend. It includes one-dimensional convolutional layers with varying dilation rates, batch normalization, and activation layers to enhance receptive fields without increasing parameters, enabling the network to learn local and global patterns. A global average pooling layer followed by fully connected layers with dropout regularization and softmax activation was employed for class prediction. The model architecture was based on a previous CNN algorithm for ECG feature identification. [1] Unlike the reference model that diagnosed seven ECG abnormalities using residual blocks and complex parameters, we omitted residual blocks for the ST-segment elevation myocardial infarction (STEMI) identification task due to excessive complexity. Additionally, inspired by a model for estimating age and sex, we diversified the number of convolutional layers and kernel sizes, reduced max pooling,

and incorporated dilated convolutions to mitigate information loss while maintaining computational efficiency. [2]

2.3. Ensemble of CNNs

An ensemble model was then derived from five independent deep neural networks with the same architecture. Each model was initialized with a unique random seed and trained separately using a batch size of 128, a learning rate of 0.001, and 50 epochs with the Adam optimizer and binary cross-entropy loss function. Class weights of 5.3601 and 0.5514 were used to address class imbalance, and early stopping with a patience of 30 was applied. The model output score for the deep ensemble model was calculated as the average of the probabilities derived from the softmax function of the five neural networks. The cut-off for the model output score was then calculated to be 0.0768, maximizing Youden's index using internal validation set.

3. References for Appendices

1. Ribeiro AH, Ribeiro MH, Paixão GMM, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat Commun.* 2020;11:1760.

2. Attia ZI, Friedman PA, Noseworthy PA, et al. Age and Sex Estimation Using Artificial Intelligence From Standard 12-Lead ECGs. *Circ Arrhythm Electrophysiol.* 2019;12:e007284.

ABSTRACT IN KOREAN

ST-분절 상승 심근경색에 대한 임상적으로 검증된 인공지능 진단 모델의 개발

배경 및 목적: ST-분절 상승 심근경색 (ST-segment elevation myocardial infarction, STEMI)에 대해 일차적 경피적 관상동맥 중재술 (percutaneous coronary intervention, PCI)이 강조되고 있지만, 적절한 심도자실 활성화가 이루어지지 못하는 경우가 많다. 본 연구에서는 적절한 심도자실 활성화를 위해 인공지능 (artificial intelligence, AI)을 이용한 STEMI 진단 모델을 개발하고 성능을 평가하고자 한다.

방법: 본 연구에서는 전향적 PCI 레지스트리에 등록된 환자들의 심전도 파형 정보를 이용하였다. 두 명의 독립적인 심장내과 전문의가 환자의 관상동맥 조영술 결과를 기반으로 각 심전도의 STEMI 여부를 확인하였다. 5개의 개별 합성곱 신경망 (convolutional neural network, CNN)을 통합하여 딥 앙상블 모델 (deep ensemble model)을 개발하였다. 모델 성능 검증을 위해 증상-기반 데이터셋을 바탕으로 한 임상 검증, 임상의와의 성능 비교, 원내 진료지침 (critical pathway, CP)과의 성능 비교, 및 외부 검증을 시행하였다. 모델의 설명 가능성을 확인하기 위해 Grad-CAM 방법을 이용해 시각화하였다.

결과: 총 18,697 개의 심전도가 모델 개발에 사용되었으며, 이 중 1,745 (9.3 %) 개의 심전도가 STEMI로 확인되었다. 개발된 AI 모델은 92.1 %의 정확도, 95.4 %의 민감도 및 91.8 %의 특이도를 보였다. 임상 검증, 임상의와의 비교, 및 외부 검증에서 AI 모델은 민감도 및 특이도 측면에서 균형 잡힌 성능을 보였다. AI 모델은 원내 CP를 통해 STEMI로 잘못 확인된 환자 중 31.6 %를 비-STEMI로 재분류하였다. Grad-CAM을 통한 시각화에서는 심전도의 ST 분절이 강조되는 것이 확인되었다.

결론: 본 연구를 통해 개발된 딥 앙상블 AI 모델은 균형 잡힌 높은 성능을 보였다. Grad-CAM에서의 ST 분절의 강조를 통해 본 AI 모델의 합리적인 설명 가능성을

확인하였다.

핵심되는 말: ST-분절 상승 심근경색, 심전도, 딥 앙상블 모델