Development of Artificial Intelligence to Predict Coronary Revascularization using Exercise ECG

Boo, Dachung

Department of Biomedical Systems Informatics

Graduate School

Yonsei University

Development of artificial intelligence to predict coronary revascularization using exercise ECG

Advisor You, Seng Chan

A Master's Thesis Submitted
to the Department of Biomedical Systems Informatics
and the Committee on Graduate School
of Yonsei University in Partial Fulfillment of the
Requirements for the Degree of
Master of Biomedical Science

Boo, Dachung

June 2025

Development of artificial intelligence to predict coronary revascularization using exercise ECG

This Certifies that the Master's Thesis of Boo, Dachung is Approved

Committee Chair	**Cho, Iksung**
Committee Member	**Hong, Namki**
Committee Member	**You, Seng Chan**

Department of Biomedical Systems Informatics
Graduate School
Yonsei University
June 2025

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Prediction of coronary revascularization by exercise stress electrocardiography using an explainable artificial intelligence

Background: Exercise stress electrocardiography (ExECG) is widely used for coronary artery disease evaluation, but its interpretation remains challenging due to variable diagnostic accuracy. I aimed to develop and validate an explainable artificial intelligence (AI) model to enhance the prediction of coronary revascularization need based on ExECG findings.

Methods: The study included 20,534 patients who underwent ExECG using the modified Bruce protocol. I developed an explainable AI framework that first extracted clinically relevant ECG features using variational autoencoders and then trained a prediction model for coronary revascularization within 90 days after ExECG. Model performance was compared against clinicians and Duke Treadmill Score.

Results: The pre-trained VAE model extracted clinically relevant ECG features by representing high-dimensional ExECG data with a small number of latent variables across exercise Stages (13–17 per Stage). The AI model demonstrated superior performance with an area under the receiver operating characteristic curve of 0.84 (95% CI: 0.80–0.88) compared to clinicians (AUROC, 0.75; 95% CI 0.71-0.80), and the Duke Treadmill Score (AUROC, 0.78; 95% CI 0.73-0.82). The odds ratio of coronary revascularization cases defined by AI was 12.37 (8.43-18.49), whereas the odds ratios determined by Duke Treadmill Score and physician diagnosis were 5.65 (3.02-9.40) and 19.65 (13.56-28.65). The model identified ST-segment depression in the mid-recovery phase as the most significant predictor of coronary revascularization need.

Conclusion: I developed and validated an explainable artificial intelligence to predict coronary revascularization by using large-scale ExECG. By integrating advanced AI predictions with interpretable ECG feature analysis, my model may improve the diagnostic utility of ExECG in clinical practice.

Key words: Exercise stress electrocardiography, Artificial Intelligence, Variational autoencoder, Coronary revascularization

1. Introduction

1.1. Research background

Cardiovascular diseases (CVDs) are the leading cause of mortality worldwide, accounting for nearly 18.6 million deaths in 2019 (95% uncertainty interval: 17.1–19.7 million)[1,2]. Most commonly, coronary artery disease (CAD) involves the development of atherosclerotic plaques within epicardial coronary arteries, which can result in myocardial ischemia[3]. Coronary revascularization, such as percutaneous coronary intervention (PCI) or coronary artery bypass grafting (CABG), is recommended when these plaques significantly restrict coronary blood flow to prevent adverse outcomes, including myocardial infarction and death[4].

Exercise stress electrocardiography (ExECG) is a widely used noninvasive diagnostic tool for detecting CAD[5,6]. ExECG evaluates cardiac function by gradually increasing physical workload while monitoring using a 12-lead ECG. ExECG has the advantage of detecting abnormal patterns that may not be apparent under normal conditions[7]. According to recent cardiology guidelines, ExECG should not be used exclusively because of low diagnostic accuracy and unacceptable false positive and negative rate[8-10]. Due to the high false-positive rate, unnecessary invasive coronary angiography procedures are recommended, which raise healthcare costs and can even cause risk[11]. However, ExECG remains widely used as the initial diagnostic test for CAD in routine clinical practice due to its widespread availability, ease of use, and low cost. It is therefore necessary to develop methods for utilizing ExECG more effectively. Several traditional scoring systems, such as the Duke Treadmill Score (DTS)[12,13], have been developed to enhance the prognostic utility of ExECG. While such scoring models offer structured frameworks for the risk stratification of patients, it only reflects a limited number of parameters, not the full spectrum of ECG morphological changes that occur during exercise. Consequently, the need for more advanced analytical methods capable of integrating diverse ECG features (such as subtle morphological changes over time) across different phases of ExECG is increasingly recognized.

Recent advances in artificial intelligence (AI) and machine learning have enabled the prediction of cardiovascular disease, aging, and mortality from only resting 12-lead ECGs

by detecting abnormal patterns[14-19]. In the detection of different cardiac arrhythmias, DL systems have reached levels of accuracy comparable to cardiologist diagnoses[20]. It is possible that AI-based analysis could enhance the detection of clinically significant abnormalities in ExECG by utilizing comprehensive waveform data from multiple leads throughout different exercise Stages.

Developing AI models based on ExECG presents unique and formidable challenges that distinguish it from standard resting ECG analysis. First, ExECG datasets are substantially smaller than those available for standard 12-lead resting ECGs, as exercise stress testing is performed only on patients with specific clinical indications, resulting in limited training data for robust model development[21]. Second, ExECG signals are temporally extensive, typically spanning 10-15 minutes across multiple exercise Stages compared to the standard 10-second resting ECG recordings. This creates a classic high-dimensional, small-sample-size problem ($p >> n$)[22], where the feature space dramatically exceeds the number of available samples. Third, the interpretability of ExECG analysis remains limited, as conventional approaches focus on isolated parameters such as ST-segment changes or exercise duration, failing to capture the complex temporal dynamics and morphological variations that occur throughout the exercise protocol.

In the medical domain, where labeled data are inherently scarce, transfer learning has emerged as a powerful solution, enabling models to leverage knowledge from large, related datasets through feature pre-training and subsequent fine-tuning[23]. Furthermore, the development of transparent and interpretable AI models is crucial to address the "black box" nature of deep learning[24,25], which fundamentally hinders clinical trust and adoption in medical decision-making. Previous groundbreaking studies by van de Leur et al. (2022) and Wouters et al. (2023) demonstrated the potential of explainable AI approaches using variational autoencoders (VAE)[26,27]. By pre-training VAE models to capture fundamental factors governing median beat ECG morphology, they successfully fine-tuned these models to predict clinical outcomes including reduced ejection fraction, all-cause mortality, and cardiac resynchronization therapy (CRT) response. Notably, their approach proved effective even for CRT outcome prediction, where labeled datasets are inherently limited due to the relatively small population of patients eligible for this specialized therapy. Critically, their latent representation framework enabled quantitative interpretation of morphological features associated with each predictive task, bridging the gap between AI performance and clinical understanding.

1.2. Related literatures

Previously, Lee et al. (2022)[28] developed a machine learning (ML) model to improve ExECG diagnostic performance for detecting obstructive CAD. The study consisted of 2,325 patients who underwent both treadmill exercise testing (TET) and coronary angiography (CAG) within a 1-year interval. A significant obstructive CAD was defined as 70% narrowing in the left anterior descending (LAD), left circumflex arteries (LCA), or right coronary arteries (RCA) or their major branches; or > 50% narrowing in the left main coronary artery (LMCA). The dataset was randomly divided into a training set (70%) and a testing set (30%) after exclusion criteria (e.g., prior PCI, inadequate heart rate achievement, missing data). Two feature groups were used for model development: (1) TET30, which was composed of 30 features selected from an initial set of 93 TET features by recursive elimination of features; and (2) TET30+Clinical, which included the TET30 along with hypertension, diabetes, dyslipidemia, smoking status, height, weight, and the Framingham risk score. For the development of models, five machine learning algorithms were used: logistic regression, support vector machine (SVM), k-nearest neighbor (KNN), random forest (RF), and extreme gradient boosting (XGBoost). The model thresholds were selected according to 85% sensitivity, in accordance with the conventional interpretation of the TET. Based on TET30 and clinical features, RF achieved the best performance with an AUROC of 0.74, reducing false positives from 76.3% (conventional TET) to 55%. There was only marginal improvement with clinical features, showing that most of the predictive information can be extracted directly from the TET. The subgroup analysis revealed better performance in men, and higher AUCs in patients under 60 years of age. Compared to the DTS, the machine learning model demonstrated higher sensitivity (0.85 ± 0.06 vs. 0.27 ± 0.05) but lower specificity (0.43 ± 0.05 vs 0.86 ± 0.03). The study demonstrates that machine learning models based on treadmill exercise test data can be used to detect obstructive coronary artery disease. However, it emphasizes the need for external validation to ensure generalizability. It was also shown that the model relied heavily on signal-derived features from the waveform, which may have overlooked important morphological characteristics of the ECG. Additionally, "work-up bias" can lead to an overestimated sensitivity and underestimated specificity since ExECG results affect the operation of the CAG.

Yilmaz et al. (2023)[29] developed a ML model that improved the diagnostic accuracy of the ExECG for predicting obstructive CAD using signal features such as P, QRS, and T waves in the TET report. In this study, 294 TET report of patients with DTS of -10 less

and underwent invasive CAG were reviewed, and 23 ECG features were manually extracted from the V5 lead. According to the RCA, LMCA, LAD, and LCA, an obstruction of 70% or more was an obstruction CAD. A total of 94 patients (31.9%) in this dataset had obstruction CAD. There were five ML models (XGBoost, KNN, SVM, MLP, and gaussian process classifier) trained and tested using a 75:25 data split and five-fold cross-validation. XGBoost demonstrated the best accuracy, specificity, sensitivity, and area under the ROC curve (AUC) with an accuracy of 80.9%, 84.6%, 67.2%, and 0.78. The performance of 17 cardiologists evaluated for the V5 signals was lower (accuracy: 41.8%, specificity: 32.4%, sensitivity: 73.3%). This study shows that ML models based on ECG waveform features can detect obstructive CAD. The study was conducted at a single center with a relatively small sample size, which may limit the generalizability of the model. It was also shown that the model relied heavily on signal-derived features from the waveform, which may have overlooked important morphological characteristics of the ECG.

Bock et al. (2024)[30] developed ML models to improve the identification of functionally relevant coronary artery disease (fCAD) using clinical variables and ExECGs. The study included 3,522 patients who underwent stress myocardial perfusion imaging (MPI-SPECT). fCAD was defined as stress-induced ischemia determined from MPI-SPECT, and the final diagnosis was based on MPI-SPECT and angiography or FFR findings. There were three models developed to predict fCAD and compared cardiologists' reading. First, CARPEclin was a random forest model that was trained using eight clinical variables (age, sex, height, weight, blood pressure, resting heart rate, and CAD history). Second, CARPEecg was a deep learning model trained on the same clinical variables and raw 12-lead ECG time-series data. Lastly, CARPEcoll is a logistic regression model based on the ensemble model and deep learning approach with the cardiologist's post-test judgement. In this study, held-out temporal tests and external validations from two Israeli centers (n = 906) were used to assess generalizability. CARPEecg achieved AUROCs of 0.71 on internal testing and 0.80 on external validation, outperforming both post-test judgement by a cardiologist (AUROC: 0.64). CARPEcoll. improved diagnostic performance (mean AUROC: 0.74) than CARPEecg and CARPEclin. CARPEclin reduced unnecessary MPI by 15–17% at a 15% risk threshold without increasing false negatives, providing greater net benefits than the cardiologist. This study developed the deep learning model by combining ExECG morphology with clinical variables to predict fCAD. Deep learning models were trained on segmented ECG sequences using a 2-6-2 slicing method. Specifically, 2 seconds from the pre-stress phase,

6 seconds from the stress phase, and 2 seconds from the recovery phase were sampled and concatenated multiple times per patient. Because of the slicing strategy, the model may have difficulty capturing the full temporal dynamics and morphological changes over the entire course of the stress ECG test. The explainability approach highlights temporal regions without specifying the precise morphological features (e.g., R-wave amplitude vs. QRS shape) that drive predictions[31]. Consequently, they provide only limited, case-specific insights and lack general interpretability.

1.3. Objective

Using a transfer learning framework, this study aims to develop an explainable AI model for predicting coronary revascularization. The overall study design is presented in Figure 1. The first step is to construct a pre-trained VAE model base on a large dataset of ExECGs to capture informative latent representations of ECG morphology. Secondly, latent features are used to fine-tune a predictive model for patients who may require coronary revascularization. The morphological features of coronary revascularization are interpreted based on latent representation. To ensure generalizability, external validation is conducted using datasets from independent institutions. Clinical validation is performed by providing the visual explanations that highlight key morphological features contributing to individual predictions.

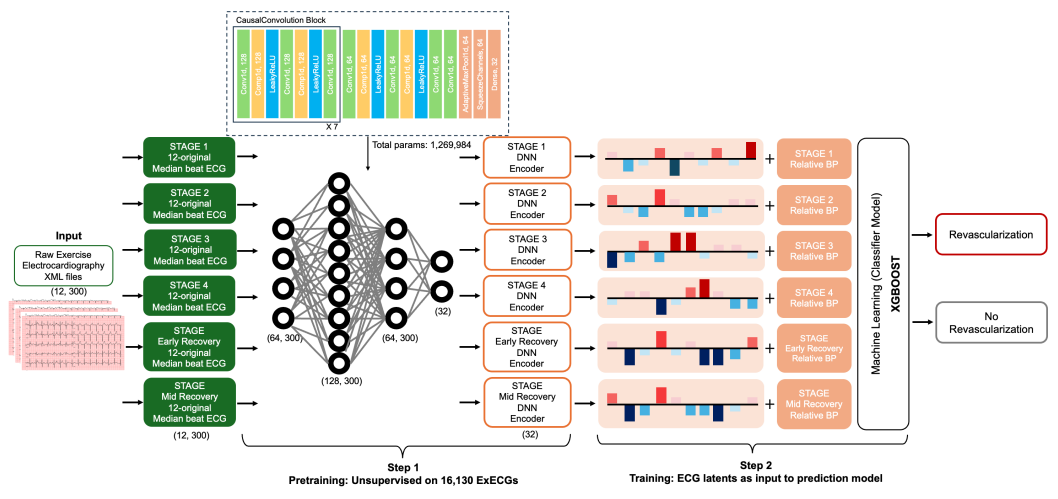


Figure 1. Schematic representation of the series of algorithms and processes

Abbreviations: Conv, Convolution; DNN, deep learning network; ECG, electrocardiogram; ExECG, exercise electrocardiogram; BP, blood pressure; XGBOOST, extreme gradient boosting decision tree; VAE, variational autoencoder.

2. Materials and Methods

2.1. Data sources and preprocessing

This study included 23,033 ExECG records from 18,998 patients who underwent at least one ExECG test using the Bruce protocol at Severance Hospital, a large tertiary referral center in South Korea, between June 23, 2020, and February 10, 2024. All raw 12-lead ExECGs were exported from GE Healthcare's MUSE Cardiology Information System[32]. Median beat ECGs of ExECGs were derived by aligning all QRS complexes during the 10 to 30 second ECG and then generating a representative QRS complex by taking the median voltage. Approximately 250 to 310 beats were measured in the median beat ECG. Afterwards, the median beat ECGs were preprocessed by padding or trimming to 300 beats. According to the Bruce protocol, median beat ECGs were mapped according to exercise times according to pretest, stress phases (Stage 1 through Stage 4), and recovery phases. Specifically, the recovery phase was subdivided into three intervals: the first 2 minutes were defined as the early-recovery phase, the subsequent 2 minutes as the mid-recovery phase, and the remaining duration as the late-recovery phase.

2.1.1. Data for pre-trained model development and testing

For the training and validation of the VAE, 633,340 median beat ECGs were obtained from 16,150 ExECGs from 13,997 patients who underwent ExECG between June 23, 2020, and February 9, 2023, were used (Figure 3). The pre-trained models were developed separately for each Stage based on the median beat ECG measurements corresponding to Stages 1, Stage 2, Stage 3, Stage 4, early recovery, and mid-recovery phases. The dataset was randomly divided into three sets, training (80%), validation (10%), and hold-out (10%) splits by participants. Labeling was not used for training the VAE model.

2.1.2. Data for coronary artery revascularization prediction model

To ensure robust model development and validation, the dataset was divided into subsets based on specific time periods. The coronary revascularization prediction model

was trained using 16,132 ExECGs obtained between June 23, 2020, and February 9, 2023. Testing and clinical validation were conducted using a separate subset of 6,431 ExECGs collected between February 10, 2023, and February 9, 2024 (Figure 4).

The performance of the prediction model was compared with that of both physician interpretation and the DTS. For DTS calculation[12], exercise duration, maximal net ST-segment deviation, and the angina index were extracted from the ExECG raw file. DTS was calculated as Duke Treadmill Score = duration of exercise, minutes – (5 × maximal net ST-segment deviation during or after exercise*, millimeters) – (4 × treadmill angina index).

For development and testing of prediction model, only ExECGs in which patients reached at least Stage 1—ensuring interpretable results—were included in the analysis. The ExECGs with insufficient information to calculate the DTS or without physician reading were excluded.

The coronary revascularization was defined by the percutaneous transluminal coronary angioplasty (PTCA) report and CABG surgery records within 90 days following the ExECG, which were retrieved from electronic medical record (EMR) databases. Each ExECG was subsequently labeled according to the occurrence of coronary revascularization.

2.1.3. Data for external validation

To evaluate the generalizability of the coronary revascularization prediction model, external validation was performed using an independent dataset collected from Yongin Severance Hospital (YSH), a secondary care hospital in South Korea. A total of 1,889 12-lead ExECG records from YSH were exported by the MUSE Cardiology Information System of GE Healthcare from January 2023 to June 2024 (Figure 5). The occurrence of coronary revascularization was labeled with ExECG utilizing the same criteria as applied to the Severance hospital.

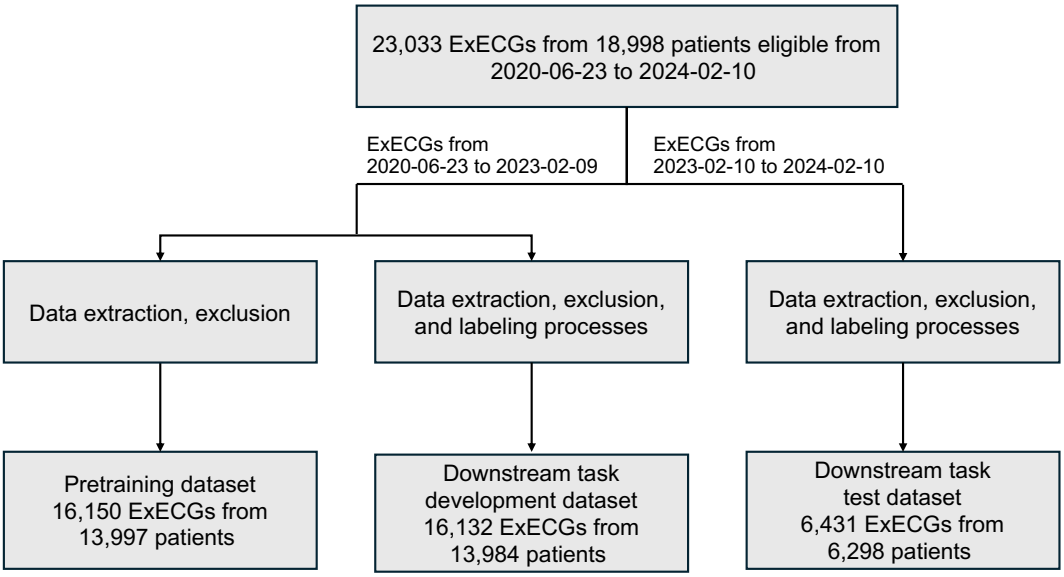


Figure 2. Data flow diagram (overview)

Abbreviations: ExECG, exercise electrocardiogram

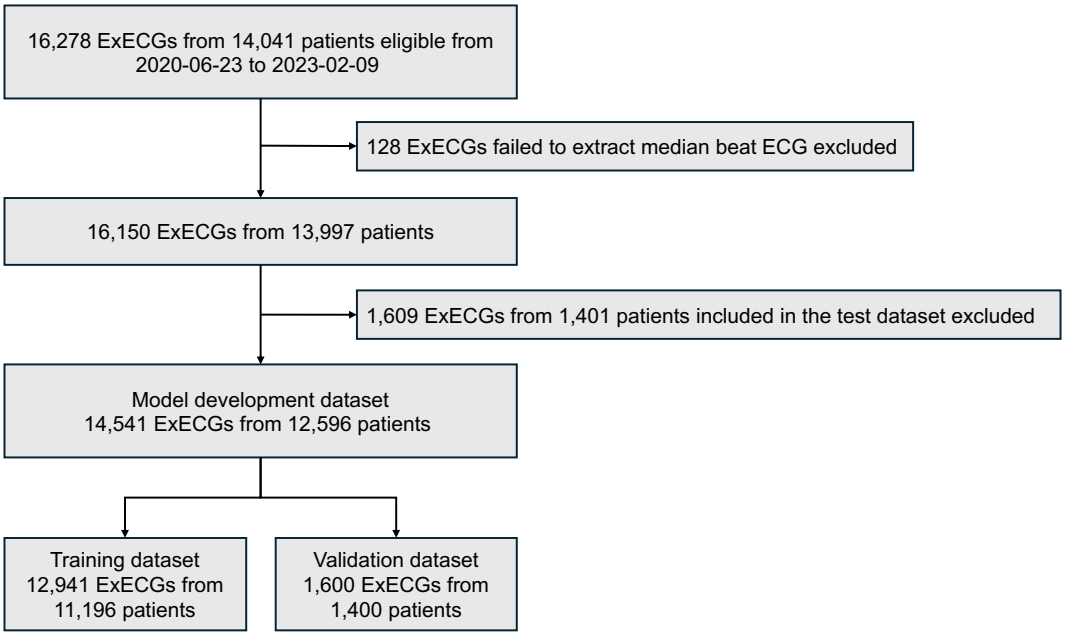


Figure 3. Data flow diagram for the development of pre-trained model

Abbreviations: ExECG, exercise electrocardiogram

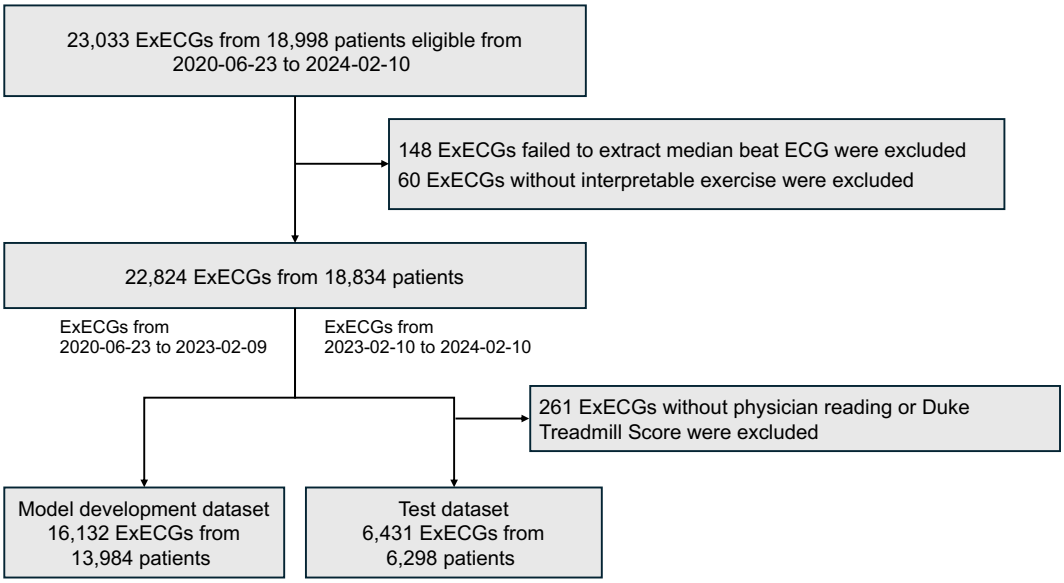


Figure 4. Data flow diagram for the development of prediction model for coronary revascularization

Abbreviations: ExECG, exercise electrocardiogram

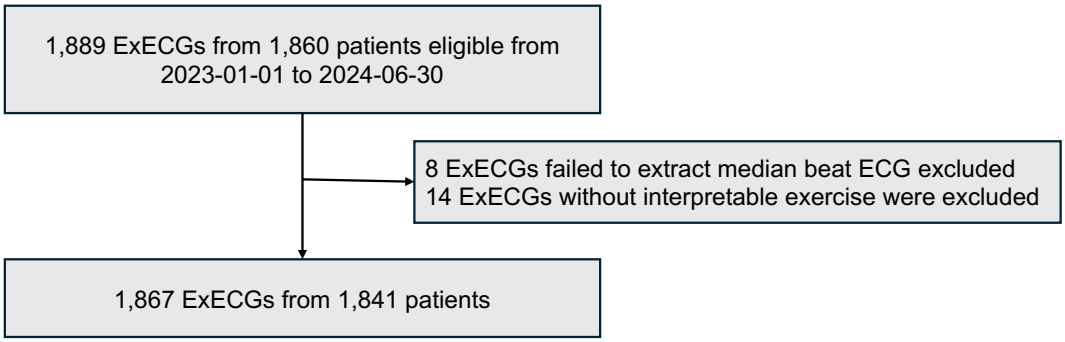


Figure 5. Data flow diagram for the external validation of prediction model for coronary revascularization

Abbreviations: ExECG, exercise electrocardiogram

2.2. Development and Validation of pre-trained model

The beta-variational autoencoder (β-VAE)[33], which used a weighted Kullback-Leibler Divergence (KLD) term in the loss function to enforce disentanglement and encode ECG data into a low-dimensional latent space, has been demonstrated to be effective in previous studies[26].

Figure 6 demonstrates the overview of model architecture. The encoder receives 300 median beat ECG data points (12x300) and structures eight 1D causal convolution blocks containing 1D causal convolution, weight normalization, leaky ReLU activations, and residual connections to transform the input into a 64x300-dimensional feature map. Adaptive max pooling reduces the temporal dimension, creating a 64-dimensional feature vector. Finally, two parallel 64-to-32 linear layers map the feature vector to mean and standard deviation parameters for a latent Gaussian distribution, with SoftPlus activation and a small constant ($\epsilon = 0.001$) applied to the standard deviation. The decoder mirrors the encoder to reconstruct lower-dimensional representations of the original electrocardiograms with continuous outputs. An initial 32-dimension vector z is transformed into a 64-dimension vector using a linear layer and subsequently reshaped into a 64x300 matrix using a second linear transformation. The output (12x300) is flattened into a vector, and two parallel linear layers map it to the mean and standard deviation for a Gaussian distribution, with SoftPlus activation and a small constant ($\epsilon = 0.001$) applied to the standard deviation. The final ECG reconstruction is reshaped back to 12x300.

The number of latent variables and the β-value, identified as the two most important hyperparameters in the β-VAE, were derived from a previous study[26]. The VAE model was trained on the entire VAE train set, using the Adam optimizer with a learning rate of 0.001, and batch size was set at 128[34]. Each model was trained over 200 epochs, and the model achieving the lowest evaluation loss was selected as the final model.

To identify the latent dimensions essential for ECG signal reconstruction, latent traversals were performed. Each individual ECG latent value was varied between -3 (represented in blue) and 3 (represented in red), while the other latent values remained constant, allowing visualization of a distinct median beat ECG morphology for each latent.

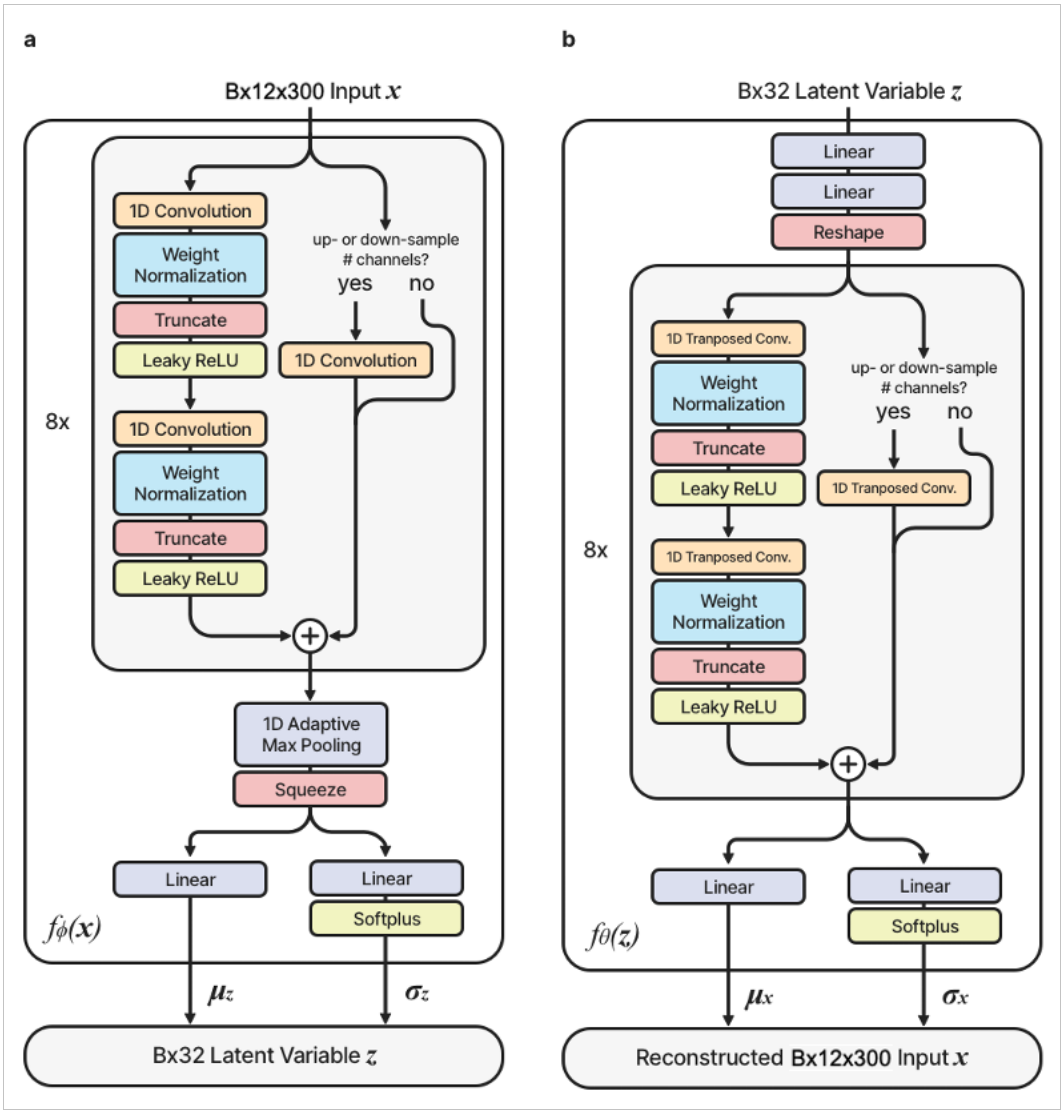


Figure 6. Overview of the VAE architecture: (a) Encoder and (b) Decoder

Abbreviations: VAE, variational autoencoder

2.3. Development and validation of coronary revascularization prediction model

For the prediction of coronary revascularization, XGBoost models[35] was trained using the significant latent values derived from the encoder of each VAE model, along with relative systolic and diastolic blood pressure measurements at each Stage compared with the resting blood pressures. The predictive performance for coronary revascularization was compared between cases where the physician diagnosis was positive and cases where the DTS was less than -10[12].

To ensure model-level interpretability, Shapley Additive Explanations (SHAP) were employed to elucidate the contribution of individual variables to specific predictions and to identify the most influential features across all variables[36]. Based on the SHAP value, predictors and ECG latent values were obtained at the patient level. The selected ECG latent variables were then visualized in relation to the median beat ECG morphology.

2.4. Clinical validation

As part of the diagnostic evaluation, a randomized cross-over trial was conducted to compare the diagnostic performance of physicians with and without assistance from the AI model. In the test set, 100 ExECG records were randomly selected with non-diagnostic or borderline findings that were considered indeterminate. All reports were independently interpreted by four cardiologists. The reports were blinded to patient information, including age, sex, and clinical history. A cross-over design was employed, in which the physicians were randomly and evenly divided into two groups. Physicians were instructed to assign a binary label for the negative or positive of coronary revascularization. The model assistant indicated whether coronary revascularization was negative or positive, along with a probability and visual representation. Initially, group 1 read the examinations without a model assistant, and group 2 read the examinations with a model assistant. After a 7-day washout period, the test sets were randomly reordered. Subsequently, Group 1 interpreted the reports with model assistance, whereas Group 2 did so without.

2.5. Statistical analysis

The baseline characteristics were presented as mean + SD or median with interquartile range for continuous variables, and as number with corresponding percentages for categorical variables. Continuous variables were compared using one-way analysis of variance or the Kruskal–Wallis test, and categorical variables using Pearson's chi-square test or Fisher's exact test. The discriminatory performance of the models was evaluated in the test sets using the area under the receiver operating curve (AUROC), sensitivity, specificity, negative predictive value (NPV), positive predictive value (PPV), balanced accuracy and area under the precision-recall curve (AUPRC). To compare the AUROCs statistically, DeLong's test was performed. To estimate 95% confidence interval (CI), 1,000 bootstrap resamples were generated.

The risk stratification for coronary revascularization using AI-based models was assessed by comparing models using physician diagnosis and DTS. Based on a probability cutoff value that was optimized during internal validation, the AI model assigned patients to high- or low-risk groups. Physician assessed the ExECG into three levels: positive, equivocal, and negative. For risk stratification, "positive" readings were categorized as high risk, while "equivocal" and "negative" readings were categorized as low risk. For DTS, under –10 was defined as high risk, and scores above –10 were considered low risk. In each method, odds ratios with 95% CI were calculated by comparing the high-risk group with the low-risk group.

Additionally, subgroup analyses were conducted within the test datasets. There were subgroups according to sex, age (under 60 years and 60 and older), and final exercise Stage (Stage 1, Stage 2, Stage 3, Stage 4), as well as previous coronary artery revascularizations. The prediction performance of coronary artery revascularization and risk stratification performance were compared for each subgroup of AI, physician, and DTS.

All analyses were performed using Python version 3.8.12 (Python Software Foundation, http://www.python.org) and R version 4.3.1 (the R Foundation, www.R-project.org).

3. Results

3.1. Baseline characteristics

3.1.1 Severance hospital

The development dataset for the coronary revascularization prediction model included 16,132 ExECGs from 13,894 patients and the test dataset included 6,431 ExECGs from 6,298 patients (Figures 4). Baseline characteristics of the development and test datasets are summarized in Table 1. In both datasets, 1.9% of cases underwent coronary revascularization.

Patients with coronary revascularization were significantly older than those without (development dataset: 64 [59–69] vs. 61 [50–68] years, p<0.001; test dataset: 64 [58–70] vs. 61 [50–68] years, p<0.001). More male patients underwent coronary revascularization than female patients (development: 84.2% vs. 64.1%, p<0.001; test: 94.3% vs. 64.3%, p<0.001). Neither the development nor the test dataset found a significant difference between patients with histories of coronary revascularization and those without (development: 2.5% vs. 0.6%, p=0.220; test: 16.5% vs. 14.6%, p=0.373). In the development dataset, patients undergoing coronary revascularization were less likely to reach higher exercise Stages. Only 42.6% of revascularized patients reached Stage 4 with 66.5% of non-revascularized patients did Stage 4 (p<0.001). It was also observed in the test dataset that the proportion of patient who achieved Stage 4 was significantly lower in the coronary revascularization group (43.9 vs 73.1%, p<0.001).

Among comorbidities, hypertension and diabetes were more prevalent in the coronary revascularization group in the development dataset (hypertension: 47.1% vs. 40.6%, p=0.025; diabetes: 33.2% vs. 21.4%, p<0.001). In the test dataset, only diabetes showed a significant difference between group (28.5% vs. 19.2%, p=0.014). Other comorbidities, including dyslipidemia, myocardial infarction, heart failure, peripheral artery disease, ischemic stroke, and atrial fibrillation, did not show significant differences between the groups in either dataset.

3.1.2 Yongin Severance hospital

The external validation dataset from YSH included 1,867 ExECGs from 1,841 patients (Figures 5). Among the 1,867 patients in the external validation dataset, 38 patients (2%) underwent coronary revascularization following ExECG.

Table 2 shows the characteristics of the external validation dataset. Patients with coronary revascularization were significantly older (59 [55-67] years vs. 51 [37-62] years, p=0.001), and the majority were males (84.2% vs. 58.2%, p=0.002). Coronary revascularization patients were less likely to reach higher exercise Stages, although the differences did not reach statistical significance (Stage 4: 60.5% vs. 74.4%, p=0.081; Stage 3: 36.8% vs. 22.1%, p=0.050). None of the coronary revascularization patients had undergone prior coronary revascularization.

The coronary revascularization group was more likely to suffer from comorbid conditions such as hypertension (55.3% vs. 29.6%, p0.001), diabetes (36.8% vs. 13.7%, p0.001), and dyslipidemia (81.6% vs. 35.4%, p0.001). Atrial fibrillation was also significantly more frequent among patients with coronary revascularization (13.2% vs. 4.2%, p=0.011).

Table 2. Clinical characteristics of development and test dataset

	Development dataset			Test dataset		
	Revascularization			Revascularization		
	Negative	Positive	*P*-value*	Negative	Positive	*P*-value*
No. of cases	15822	310		6308	123	
Age, years	61 [50-68]	64 [59-69]	<0.001	61 [50-68]	64 [58-70]	<0.001
Sex						
Male, *n (%)*	10136 (64.1)	261 (84.2)	<0.001	4054 (64.3)	116 (94.3)	<0.001
History of coronary revascularization						
At least one time, *n (%)*	401 (2.5)	2 (0.6)	0.220	1043 (16.5)	18 (14.6)	0.373
Prior Cardiac Imaging						
Angiography, *n (%)*	738 (4.6)	31 (10.0)	<0.001	220 (3.5)	16 (13.0)	<0.001
Heart CT, *n (%)*	834 (5.3)	29 (9.4)	0.002	421 (6.7)	13 (10.5)	0.128
Post Cardiac Imaging						
Angiography, *n (%)*	717 (4.5)	296 (95.5)	<0.001	171 (2.7)	123 (100)	<0.001
Heart CT, *n (%)*	1268 (8.0)	85 (27.4)	<0.001	445 (7.1)	39 (31.7)	<0.001

Achieved exercise Stage						
Stage 1	138 (0.8)	3 (0.9)	1.000	16 (0.3)	1 (0.8)	0.757
Stage 2	526 (3.3)	44 (14.2)	<0.001	63 (1.0)	19 (15.4)	<0.001
Stage 3	4634 (29.3)	131 (42.3)	<0.001	1619 (25.7)	49 (39.8)	0.001
Stage 4	10524 (66.5)	132 (42.6)	<0.001	4610 (73.1)	54 (43.9)	<0.001
Comorbidities						
Hypertension, *n (%)*	6426 (40.6)	146 (47.1)	0.025	2483 (39.4)	54 (43.1)	0.457
Diabetes, *n (%)*	3383 (21.4)	103 (33.2)	<0.001	1209 (19.2)	35 (28.5)	0.014
Dyslipidemia, *n (%)*	7450 (47.1)	158 (51.0)	0.194	2816 (44.6)	62 (50.4)	0.237
Previous myocardial infarction, *n (%)*	1233 (7.8)	21 (6.7)	0.578	396 (6.3)	11 (8.9)	0.310
Heart failure, *n (%)*	1351 (8.5)	24 (7.7)	0.693	557 (8.8)	12 (9.8)	0.843
Peripheral arterial disease, *n (%)*	506 (3.2)	10 (3.2)	1.000	230 (3.6)	6 (4.9)	0.633
Ischemic stroke, *n (%)*	302 (1.9)	4	0.562	93	1	0.832

		(1.3)		(1.5)	(1.0)	
Atrial fibrillation, *n (%)*	1173 (7.4)	16 (5.2)	0.164	475 (7.5)	9 (7.3)	1.000

The comorbidities were identified based on ICD-10 codes, including hypertension (I10, I11, I12, I13, I15), diabetes mellitus (E10–E14), dyslipidemia (E78), previous myocardial infarction (I21, I22, I25.2), heart failure (I11.0, I50, I97.1), peripheral arterial disease (I70, I71), ischemic stroke (I63, I64), and atrial fibrillation (I48). Continuous variables are presented as median [Q1-Q3] and categorical variables are presented as number (percentage).

* P-values were derived using Pearson's chi-squared test for categorical variables and the Wilcoxon rank-sum test for continuous numeric variables.

Table 2. Clinical characteristics of external validation dataset

	Negative for Revascularization	Positive for Revascularization	*P*-value*
No. of cases	1,829	38	
Age, years	51 [37-62]	59 [55-67]	<0.001
Sex			
Male, *n (%)*	1,064 (58.2)	32 (84.2)	0.002
History of coronary revascularization			
At least one time, *n (%)*	74 (4.0)	0 (0.0)	0.398
Achieved exercise Stage			
Stage 1	11 (0.6)	0 (0.0)	1.000
Stage 2	53 (2.9)	1 (2.6)	1.000
Stage 3	404 (22.1)	14 (36.8)	0.050
Stage 4	1361 (74.4)	23 (60.5)	0.081
Comorbidities			
Hypertension, *n (%)*	541 (29.6)	21 (55.3)	<0.001
Diabetes, *n (%)*	250 (13.7)	14 (36.8)	<0.001
Dyslipidemia, *n (%)*	647 (35.4)	31 (81.6)	<0.001
Previous myocardial infarction, *n (%)*	69 (3.8)	4 (10.5)	0.054
Heart failure, *n (%)*	236 (12.9)	4 (10.5)	1.000
Peripheral arterial disease, *n (%)*	54 (3.0)	0 (0.0)	0.614
Ischemic stroke, *n (%)*	15 (0.8)	0 (0.0)	1.000
Atrial fibrillation, *n (%)*	76 (4.2)	5 (13.2)	0.011

The comorbidities was identified based on ICD-10 codes, including hypertension (I10, I11, I12, I13, I15), diabetes mellitus (E10–E14), dyslipidemia (E78), previous myocardial infarction (I21, I22, I25.2), heart failure (I11.0, I50, I97.1), peripheral arterial disease (I70, I71), ischemic stroke (I63, I64), and atrial fibrillation (I48). Continuous variables are presented as median [Q1-Q3] and categorical variables are presented as number (percentage).

* P-values were derived using Pearson's chi-squared test for categorical variables and the Wilcoxon rank-sum test for continuous numeric variables.

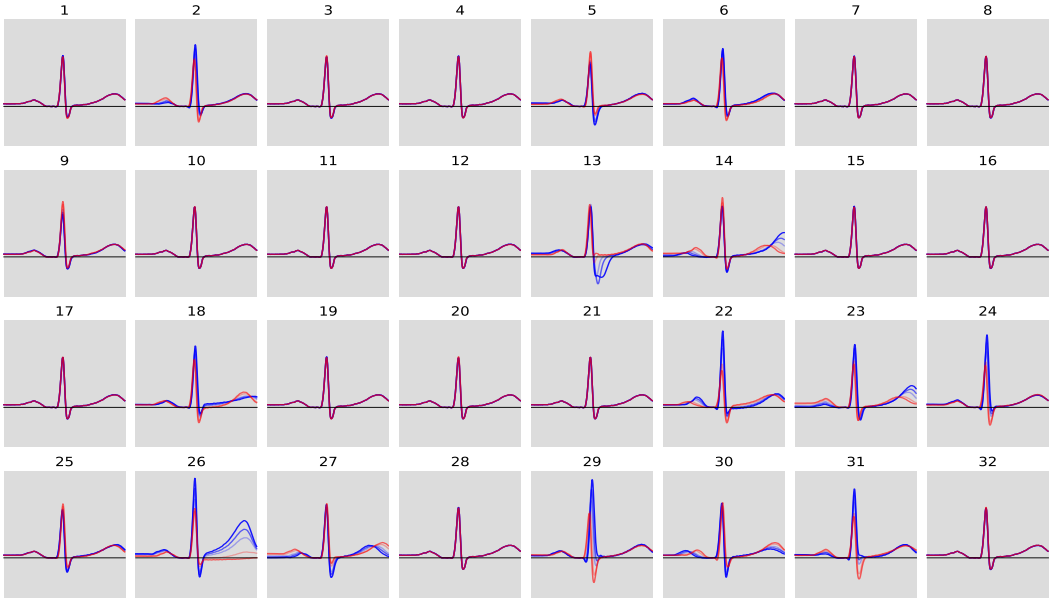
3.2. Performance of the variational auto-encoder

The training dataset consisted of 12-lead median beat ECGs, which were distributed as follows: 65,808 samples for the Stage 1 VAE model, 65,185 for the Stage 2 VAE model, 59,130 for the Stage 3 VAE model, 31,467 for the Stage 4 VAE model, and 105,706 for the recovery phase VAE model.

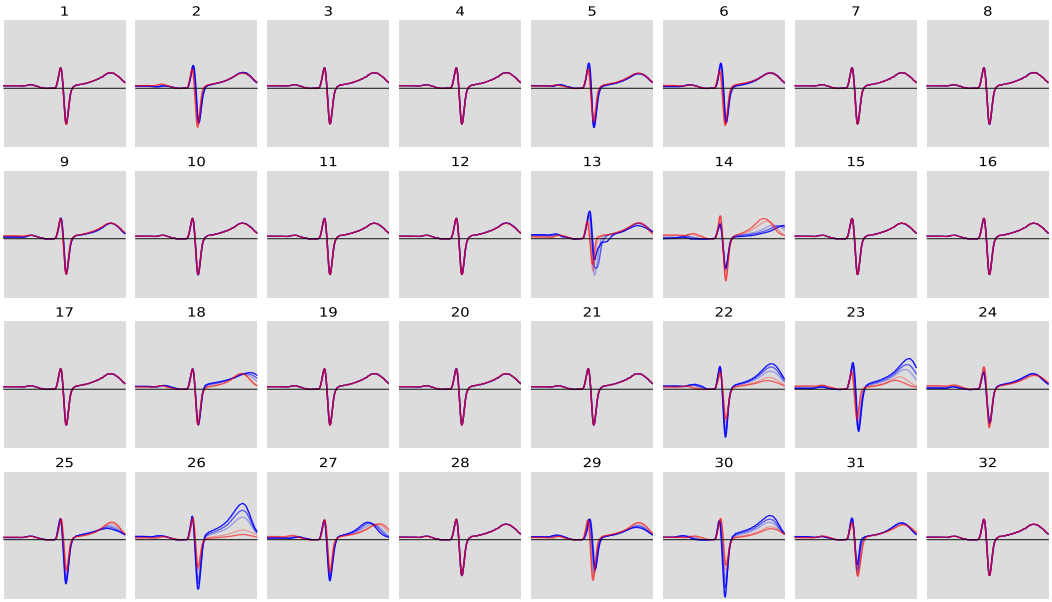
The performance of VAE models in reconstructing median beat ECGs was evaluated. Based on the Stage 1 VAE model, Pearson correlation coefficient was 0.935 ($P < 0.001$). Furthermore, the Stage 2 VAE model showed a correlation of 0.933 ($P < 0.001$), while the Stage 3 and Stage 4 VAE models both achieved a correlation of 0.934 ($P < 0.001$). The recovery phases VAE model had the highest reconstruction accuracy with a mean Pearson correlation coefficient of 0.940 ($P < 0.001$). These results indicate a high level of reconstruction performance across all VAE models.

The results indicate that only a subset of the 32 latent variables is actively utilized for reconstruction at each Stage VAE. Specifically, 16 latent variables were utilized in Stage 1, 14 latent variables in both Stage 2 and Stage 3, 13 latent variables in Stage 4, and 17 latent variables in the recovery Stage. Latent traversals for these variables are illustrated in Figures 2–8.

(A) Lead II



(B) Lead V3



(C) Lead V4

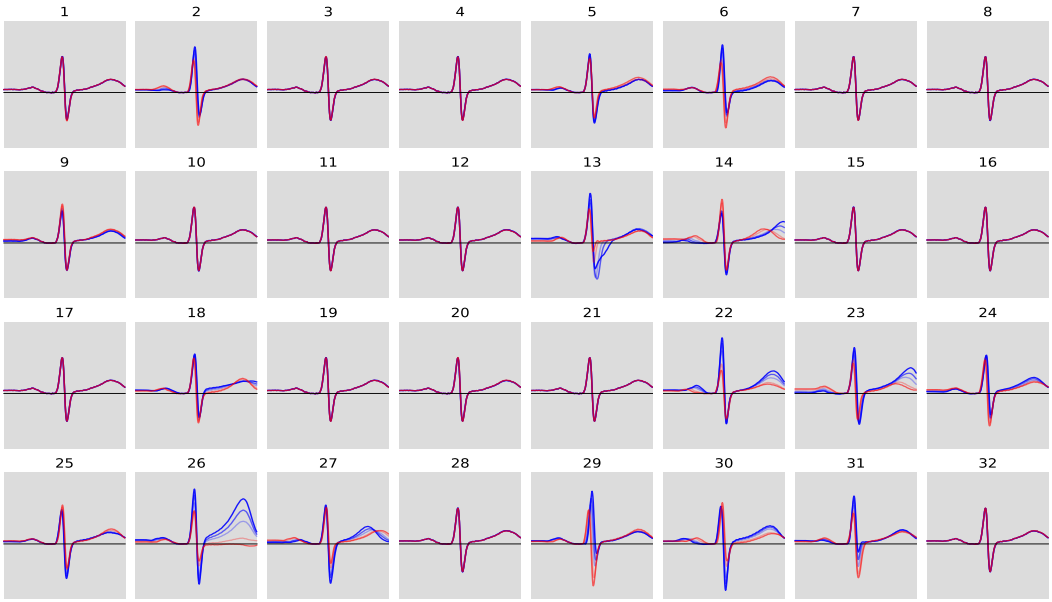
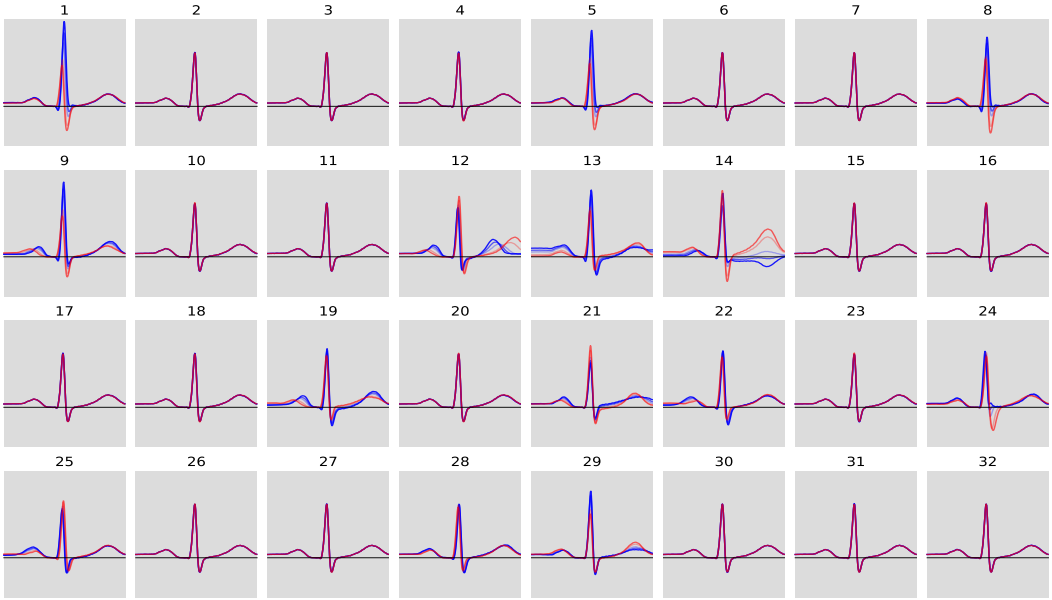


Figure 7. Latent traversals of all the ECG factors from Stage 1 pre-trained model (lead II, lead V3, and lead V4)

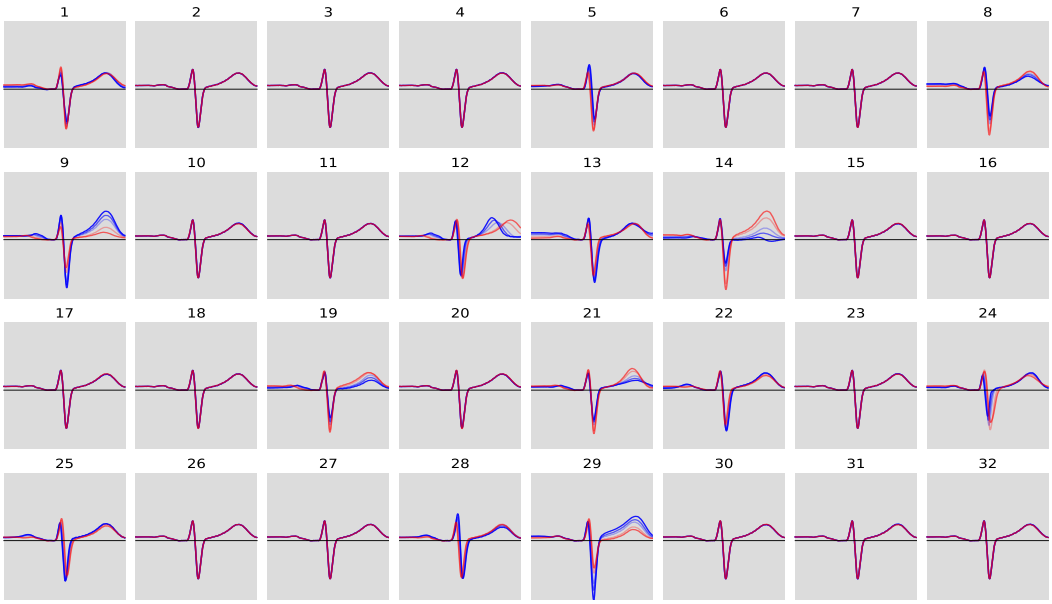
latent traversals of a subset of the 16 ECG factors (latent number 2, 5, 6, 9, 13, 14, 18, 22, 23, 24,25, 26, 27, 29, 30, 31) that hold significant information for correctly reconstructing electrocardiograms.

Abbreviations: ECG, electrocardiogram.

(A) Lead II



(B) Lead V3



(C) Lead V4

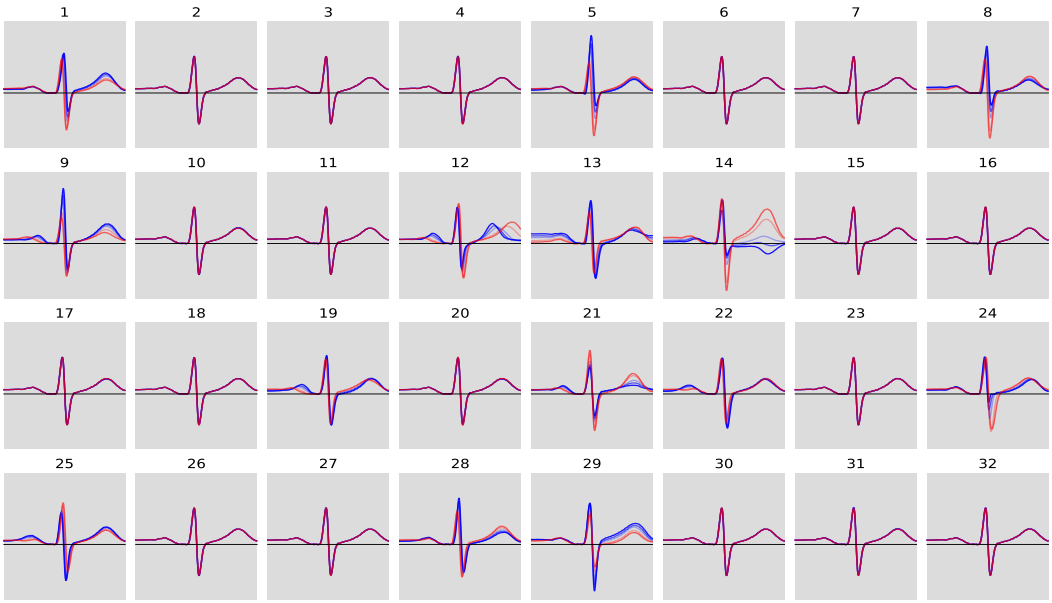
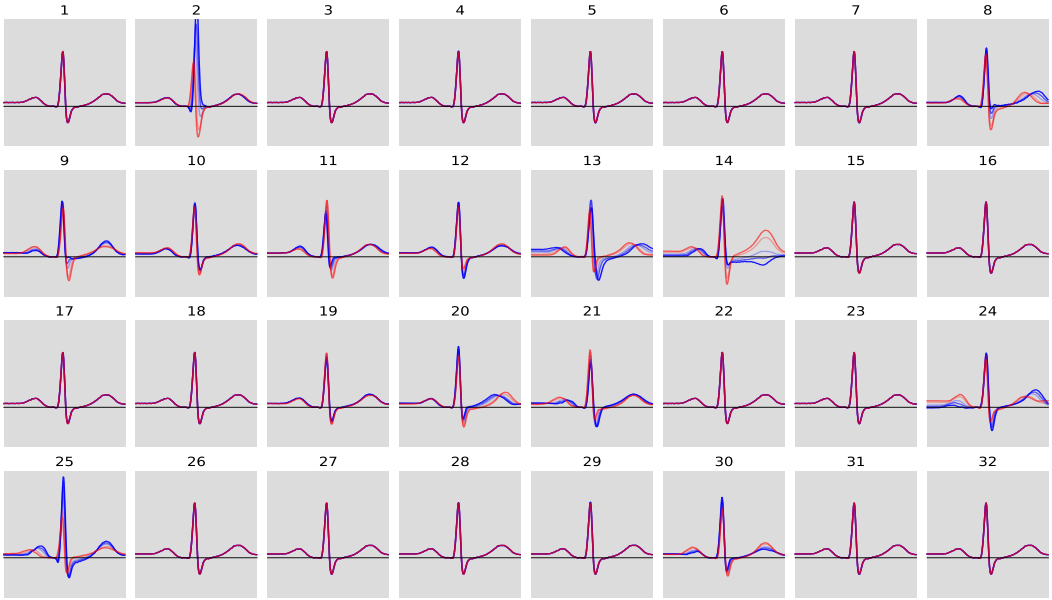


Figure 8. Latent traversals of all the ECG factors from Stage 2 pre-trained model (lead II, lead V3, and lead V4)

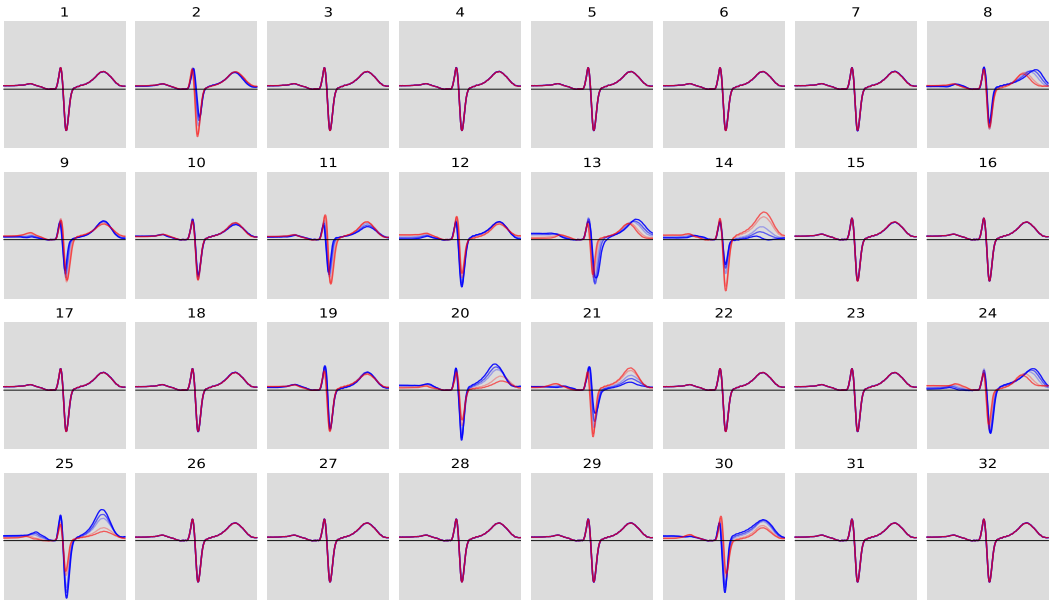
latent traversals of a subset of the 14 ECG factors (latent number 1, 5, 8, 9, 12, 13, 14, 19, 21, 22, 23, 24, 25, 28, 29) that hold significant information for correctly reconstructing electrocardiograms.

Abbreviations: ECG, electrocardiogram.

(A) Lead II



(B) Lead V3



(C) Lead V4

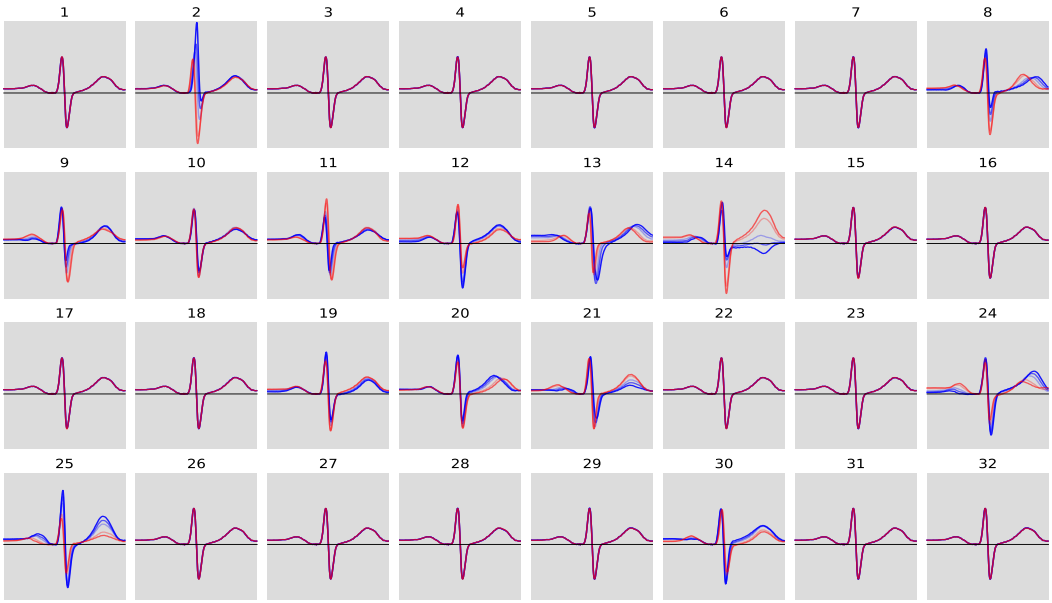
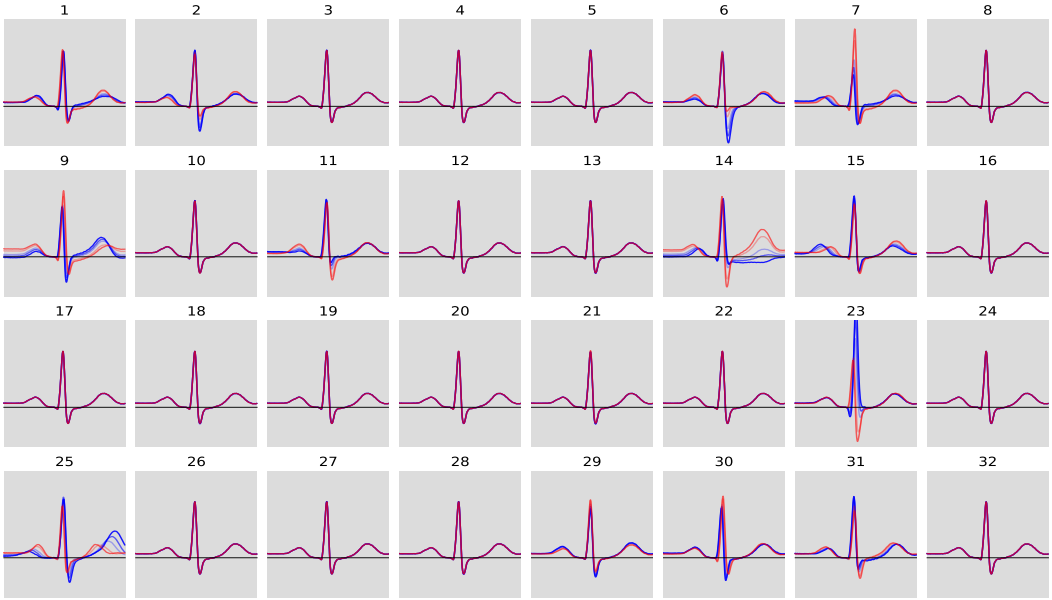


Figure 9. Latent traversals of all the ECG factors from Stage 3 pre-trained model (lead II, lead V3, and lead V4)

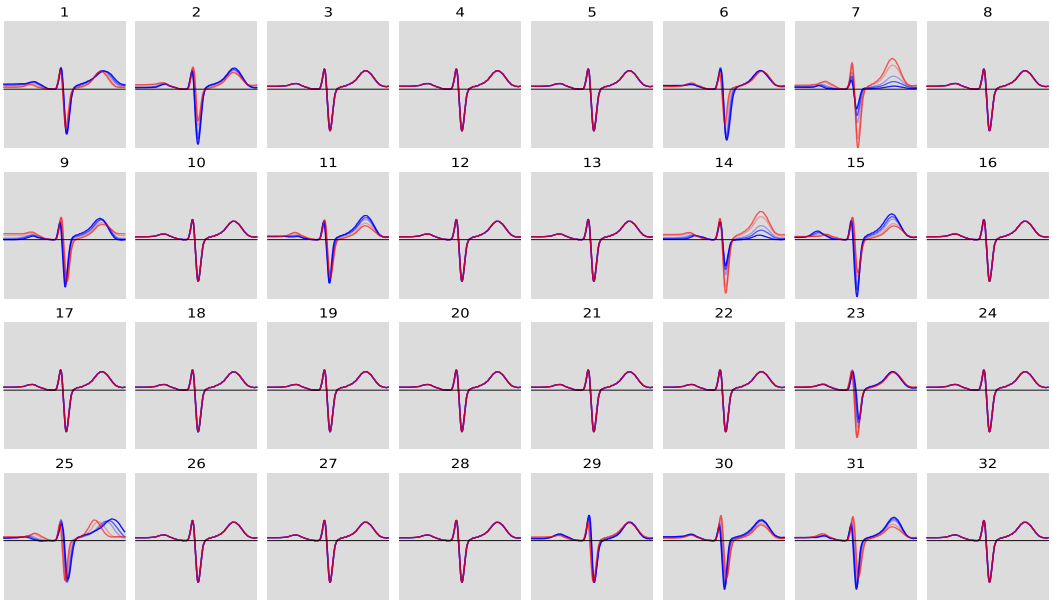
latent traversals of a subset of the 14 ECG factors (latent number 2, 8, 9, 10, 11, 12, 13, 14, 19, 20, 21, 24, 25, 30) that hold significant information for correctly reconstructing electrocardiograms.

Abbreviations: ECG, electrocardiogram.

(A) Lead II



(B) Lead V3



(C) Lead V4

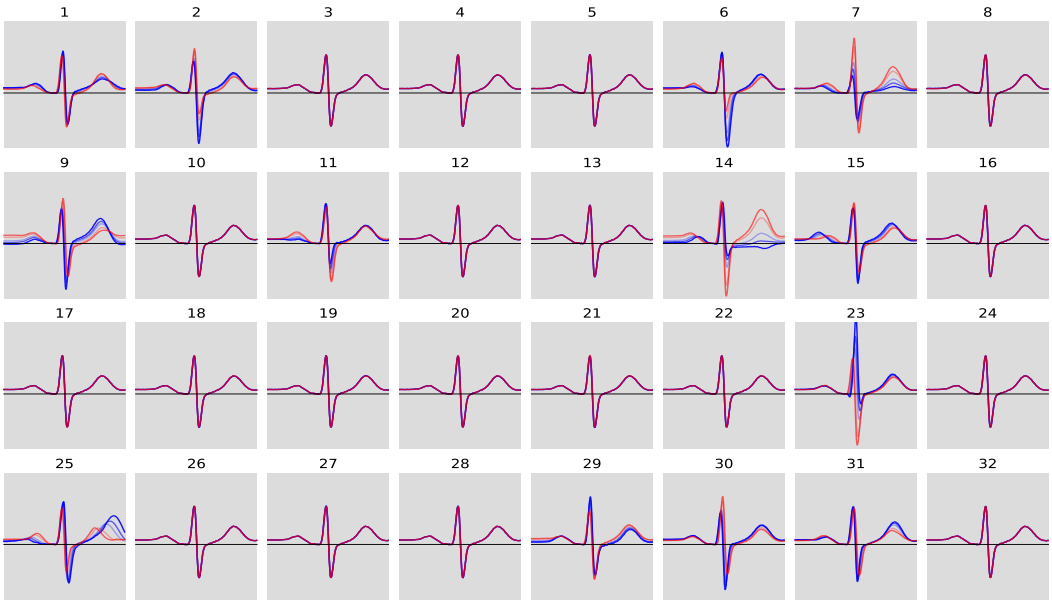
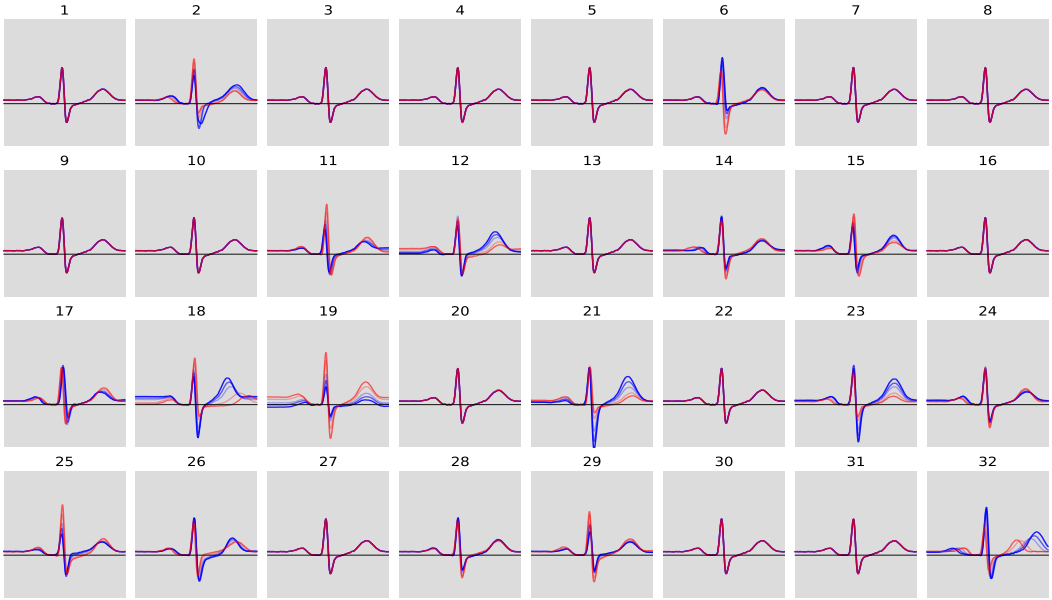


Figure 10. Latent traversals of all the ECG factors from Stage 4 pre-trained model (lead II, lead V3, and lead V4)

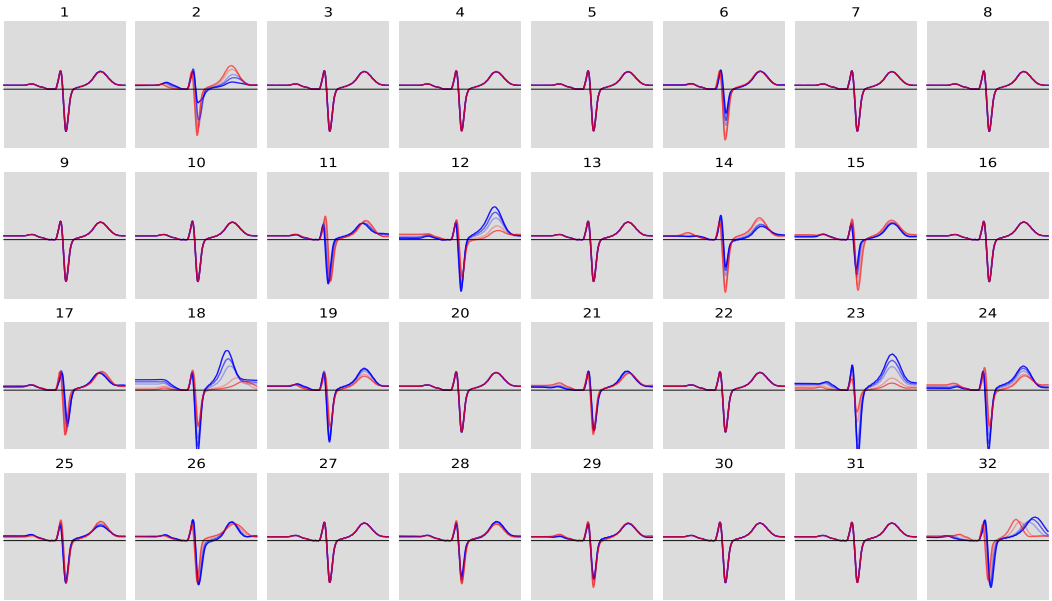
latent traversals of a subset of the 13 ECG factors (latent number 1, 2, 6, 7, 9, 11, 14, 15, 23, 25, 29, 30, 31) that hold significant information for correctly reconstructing electrocardiograms.

Abbreviations: ECG, electrocardiogram.

(A) Lead II



(B) Lead V3



(C) Lead V4

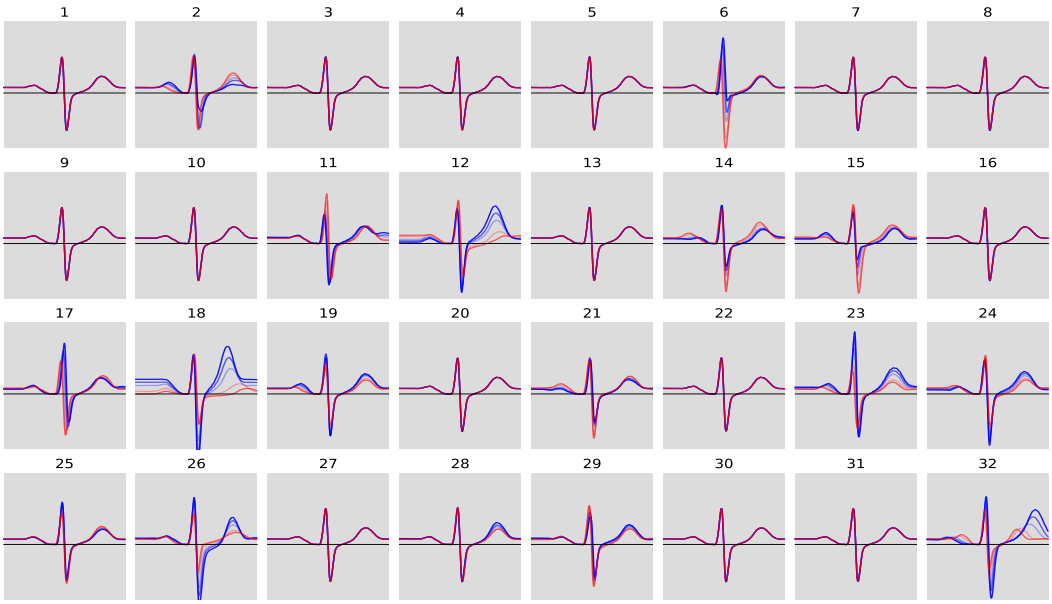


Figure 11. Latent traversals of all the ECG factors from Recovery phase pre-trained model (lead II, lead V3, and lead V4)

latent traversals of a subset of the 17 ECG factors (latent number 2, 6,11, 12, 14, 15, 17, 18, 19, 21, 23, 24, 25, 26, 28, 29, 32) that hold significant information for correctly reconstructing electrocardiograms.

Abbreviations: ECG, electrocardiogram.

3.3. Performance of coronary artery revascularization prediction model

Based on the latent variables derived from the VAE models, machine learning models were developed and tested to predict coronary revascularization. Accordingly, latent variables were extracted from Stage 1 (16 variables), Stage 2 (14 variables), Stage 3 (14 variables), Stage 4 (13 variables), and both early-recovery and mid-recovery phases (17 variables each). The model also included 12 variables indicating changes in systolic and diastolic blood pressure during each exercise phase. A total of 103 variables were used to train and evaluate the XGBoost model.

The AI-based prediction model achieved an AUROC of 0.84 (95% CI: 0.80-0.89) and an AUPRC of 0.25 (0.18-0.33) in the test dataset. The performance of the physician and the DTS was AUROC: 0.80 (95% CI: 0.76–0.85) and AUPRC: 0.11 (95% CI: 0.08–0.14), and AUROC: 0.78 (95% CI: 0.73–0.82) and AUPRC: 0.07 (95% CI: 0.05–0.09), respectively. (Figure 3 and Figure 4). According to DeLong's test, the AUROC of AI-based prediction model was significantly higher than that of DTS (DeLong test [unpaired, two-sided], AI vs. Physician P = 0.141, AI vs. DTS P = 0.002, and Physician vs. DTS P = 0.171). Table 3 shows the performance metrics of models. In AI-based prediction model, 0.068 was selected as a threshold of similar sensitivity as physicians.

Table 4 summarizes the odds ratios of coronary revascularization for the three methods in high-risk groups. According to the AI model, the OR for the high-risk group was 19.06 (95% CI: 13.18–27.76). Positive physician-assessed results were associated with a similar OR of 19.85 (95% CI: 13.72–28.90), whereas the DTS high-risk group had a lower OR of 8.14 (95% CI: 5.07–12.65).

The most important global ECG latent variables for the prediction of coronary revascularization were high values for latent 26 and 18 of Mid-recovery, latent 2 of Stage 2 and latent 4 of Stage 25. (Figure 3)

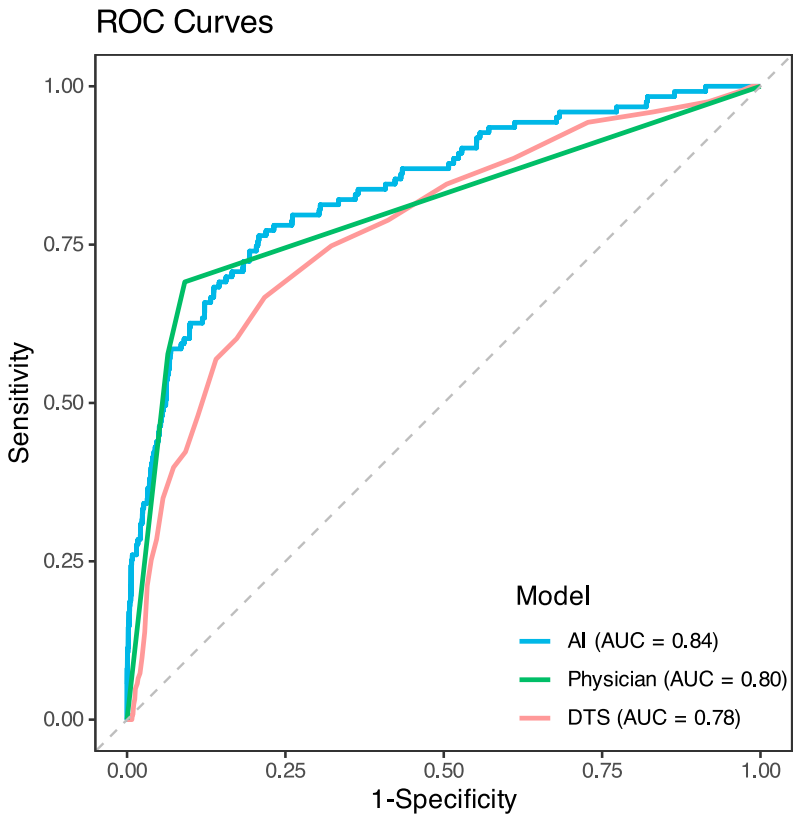


Figure 12. ROC curves of AI model, Physician, and DTS

Abbreviations: ROC, eceiver operating curve; AI, artificial intelligence; DTS, duke treadmill score.

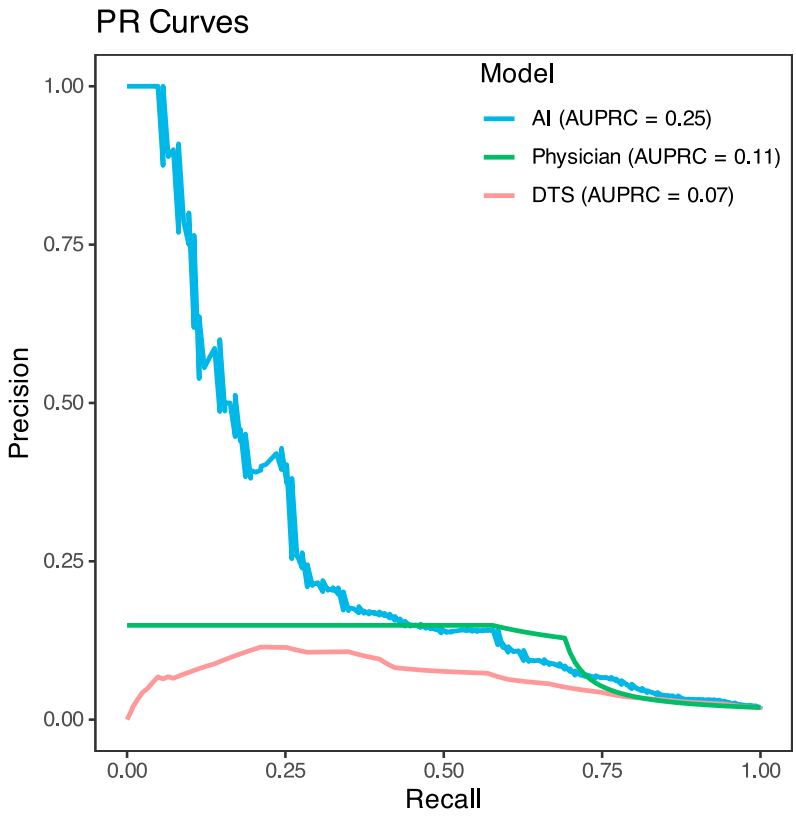


Figure 13. PR curves of AI model, Physician, and DTS

Abbreviations: PR, precision-recall; AI, artificial intelligence; DTS, duke treadmill score.

Table 3. Comparison of AI performance with existing criteria and validation

	AI *	Physician	DTS
Sensitivity	0.59	0.58	0.14
Specificity	0.93	0.94	0.97
F1-score	0.23	0.24	0.11
PPV	0.14	0.15	0.09
NPV	0.99	0.99	0.98
Balanced Accuracy	0.76	0.76	0.56

* The AI model's cut-off value is 0.068 which was selected as the cutoff value, which yielded similar sensitivity as physician.

Abbreviations: DTS, duke treadmill score; PPV, positive predictive value; NPV, negative predictive value.

Table 4. Odds ratio of coronary revascularization according to risk stratification by AI model and existing criteria

Analysis	Risk group	Outcome/ N (N=6,431)	Odds ratio (95% CI)
AI	Low risk (under cut-off*)	51/5,924	1.00 (reference)
	High risk (over cut-off*)	72/507	19.06 (13.18-27.76)
Physician	Negative & Equivocal	52/5,954	1.00 (reference)
	Positive	71/477	19.85 (13-72-28.90)
DTS	Low-Medium	97/6,204	1.00 (reference)
	High	26/227	8.14 (5.07-12.65)

* For internal validation, 0.068 was selected as the cutoff value, which yielded similar sensitivity as physician.

Abbreviations: AI, artificial intelligence; CI, confidence interval; DTS, duke treadmill score

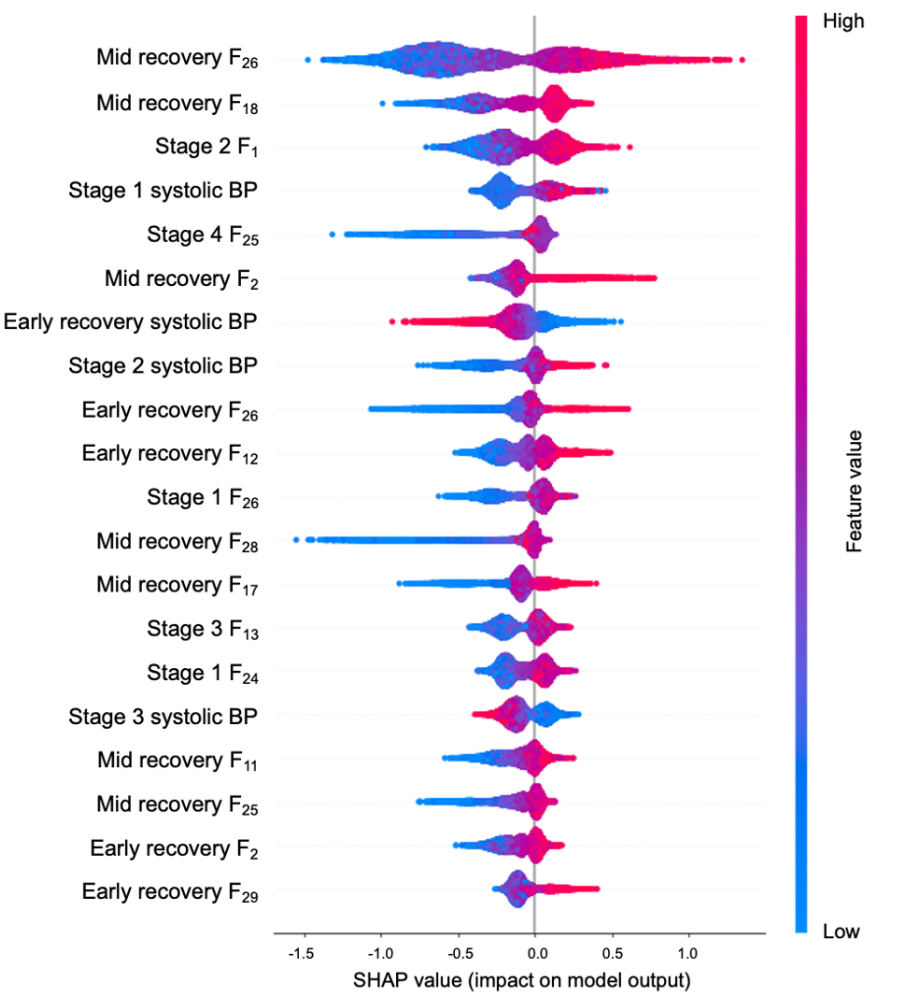


Figure 14. Explanations for the coronary artery revascularization using Shapley Additive exPlanations values

In the figure, F followed by a number represents the sequential position of the latent feature. Early recovery is defined as the first two minutes of recovery and the next two minutes as the mid-phase.

Abbreviations: BP, blood pressure; SHAP value, Shapley Additive exPlanations values.

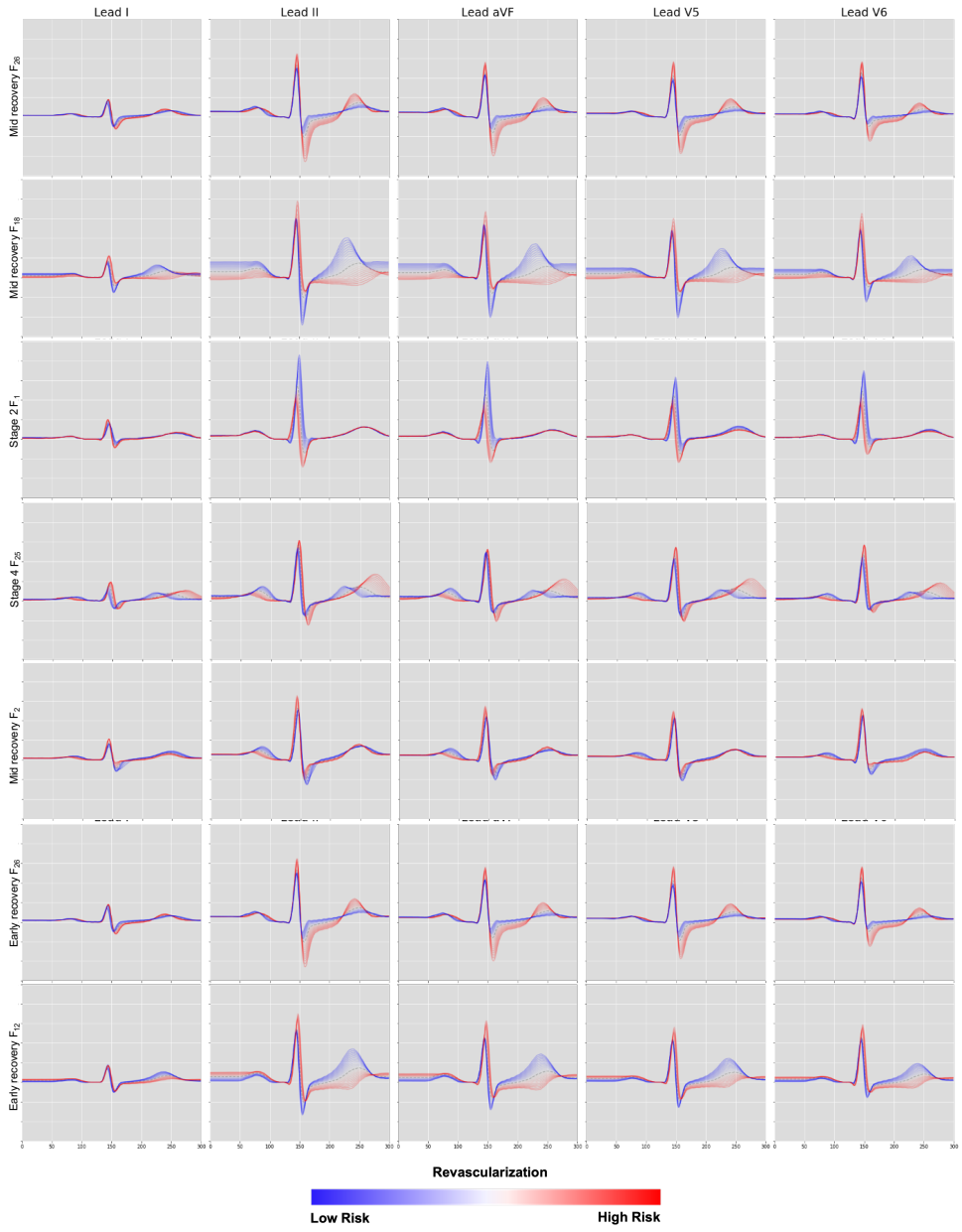


Figure 15. Latent traversals of Top 10 latent from coronary revascularization prediction model

3.4 Performance of coronary artery revascularization prediction model in the external validations

The external validations showed AUROCs of 0.74 and 0.62 and AUPRCs of 0.06 and 0.04 in the AI model and DTS model, respectively. Table 5 shows the performance metrics of models. The same threshold of 0.068, derived from the internal tests, was used for external validation. The odds ratios of AI model and DTS in high-risk groups are shown in Table 6. AI model showed an OR of 5.87 (95% CI: 1.94–15.52) for the high-risk group, whereas DTS showed a higher OR of 7.24 (95% CI: 3.25–14.93).

Table 5. Comparison of AI performance in internal and external validation

	Internal		External	
	AI model	DTS	AI model	DTS
AUROC (95%CI)	0.84 (0.80-0.89)	0.78 (0.73-0.82)	0.74 (0.66-0.81)	0.62 (0.52-0.73)
AUPRC (95%CI)	0.25 (0.17-0.33)	0.07 (0.05-0.09)	0.06 (0.03-0.11)	0.05 (0.03-0.09)
Sensitivity	0.59	0.14	0.13	0.26
Specificity	0.93	0.97	0.98	0.95
F1-score	0.23	0.11	0.11	0.15
PPV	0.14	0.09	0.10	0.10
NPV	0.99	0.98	0.98	0.988
Balanced Accuracy	0.76	0.56	0.55	0.61

* The same threshold of 0.068, derived from the internal tests, was used for external validation.

Abbreviations: AI, artificial intelligence; DTS, duke treadmill score; AUROC, the area under the receiver operating curve; AUPRC, the area under the precision recall curve; CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value.

Table 6. Odds ratio of coronary revascularization according to risk stratification by AI model and duke treadmill score in external validation

Analysis	Risk group	Outcome/N (N=1,867)	Odds ratio (95% CI)
AI	Low risk (under cut-off*)	46/1,829	1.00 (reference)
	High risk (over cut-off*)	5/38	5.87 (1.94-14.52)
DTS	Low-Medium	86/1,829	1.00 (reference)
	High	10/38	7.24 (3.25-14.93)

The same threshold of 0.068, derived from the internal tests, was used for external validation.

Abbreviations: AI, artificial intelligence; DTS, duke treadmill score; CI, confidence interval.

3.5. Subgroup analysis

Figure 9 presents subgroup analyses according to gender, age, achieved exercise Stage, and history of coronary revascularization for the prediction of coronary revascularization. For each model (AI, physician, and DTS), interaction P-values were calculated to determine whether their predictive performance differed significantly by subgroup.

There was a relatively better performance from all three models in predicting coronary revascularization in males than in female. The AUROCs of 0.83, 0.80, and 0.75 and AUPRCs of 0.28, 0.13, and 0.08 in AI model, physician, and DTS, respectively. The AUROC of DTS was significantly lower than that of AI model and physician (DeLong test [unpaired, two-sided], AI vs. Physician P = 0.211, AI vs. DTS P < 0.001, and Physician vs. DTS P = 0.042).

For young group, AI model outperformed than two models. For young people, all three models performed better than for older people in predicting coronary revascularization. In the young group, the AUROCs for AI model, physician, and DTS were 0.92, 0.83, and 0.80; and the AUPRCs were 0.33, 0.14, and 0.07. AI model had a significantly higher AUROC than physician and DTS (DeLong test [unpaired, two-sided], AI vs. Physician P = 0.002, AI vs. DTS P < 0.001, and Physician vs. DTS P = 0.311).

According to the achieved exercise Stage, the Stage 1 & 2 group showed the best performance in both AUROC and AUPRC. The AUROCs were 0.88, 0.81, and 0.85, respectively, and the AUPRCs were 0.73, 0.44, and 0.49. There were no statistical differences between the models (DeLong test [unpaired, two-sided]: AI vs. Physician, P = 0.174; AI vs. DTS, P = 0.513; Physician vs. DTS, P = 0.530). In Stage 4 group, the AUROCs for AI model, physician, and DTS were 0.81, 0.80, and 0.72; and the AUPRCs were 0.10, 0.08, and 0.03. The AUROC of DTS was significantly lower than that of AI model and physician (DeLong test [unpaired, two-sided], AI vs. Physician P = 0.732, AI vs. DTS P = 0.014, and Physician vs. DTS P = 0.014).

Patients without a history of coronary revascularization had AUROCs of 0.84, 0.80, and 0.78, respectively, while AUPRCs were 0.26, 0.12, and 0.07, respectively. The AUROC of DTS was significantly lower than that of AI model and physician (DeLong test [unpaired, two-sided], AI vs. Physician P = 0.158, AI vs. DTS P = 0.003, and Physician vs. DTS P = 0.301). With at least one prior revascularization, the AUROC values for the AI model, physician, and DTS were 0.84, 0.80, and 0.74, respectively; and the AUPRC values were 0.23, 0.08, and 0.08. There were no statistical differences

between the models (DeLong test [unpaired, two-sided]: AI vs. Physician, P = 0.567; AI vs. DTS, P = 0.194; Physician vs. DTS, P = 0.273).

The interaction P-values were greater than 0.05 in most subgroups, indicating that performance differences were generally consistent regardless of subgroup characteristics. However, for the age subgroup in AI, the interaction P-value was less than 0.01, demonstrating statistically significant differences in model performance by age.

Analysis	AUROC (95% CI)	AUPRC (95% CI)	OR (95% CI)	P value for Interaction
Gender				
AI (Male)	0.83 (0.79-0.87)	0.28 (0.19-0.35)	15.81 (10.77-23.42)	0.68
AI (Female)	0.71 (0.50-0.92)	0.10 (0.00-0.38)	11.01 (1.56-51.96)	
Physician (Male)	0.80 (0.75-0.84)	0.13 (0.10-0.17)	15.48 (10.55-22.87)	0.55
Physician (Female)	0.76 (0.56-0.96)	0.03 (0.00-0.08)	24.86 (4.82-114.95)	
DTS (Male)	0.75 (0.71-0.80)	0.08 (0.06-0.10)	4.23 (2.39-7.08)	0.97
DTS (Female)	0.73 (0.50-0.97)	-	-	
Age				
AI (< 60)	0.92 (0.87-0.96)	0.33 (0.20-0.46)	59.23 (31.42-114.26)	<0.01
AI (≥ 60)	0.78 (0.72-0.84)	0.25 (0.16-0.34)	11.62 (7.34-18.56)	
Physician (< 60)	0.83 (0.76-0.90)	0.14 (0.08-0.20)	29.29 (15.85-54.81)	0.12
Physician (≥ 60)	0.79 (0.73-0.84)	0.11 (0.07-0.14)	15.78 (9.94-25.37)	
DTS (< 60)	0.80 (0.72-0.87)	0.07 (0.04-0.10)	5.09 (2.04-11.01)	0.66
DTS (≥ 60)	0.76 (0.70-0.82)	0.07 (0.05-0.10)	6.49 (3.04-12.55)	
Achieved exercise stage				
AI (Stage 1&2)	0.88 (0.79-0.97)	0.73 (0.52-0.89)	13.22 (3.49-86.85)	0.84
AI (Stage 3)	0.79 (0.72-0.86)	0.21 (0.11-0.33)	12.77 (6.98-24.48)	
AI (Stage 4)	0.81 (0.75-0.87)	0.10 (0.05-0.16)	16.47 (9.13-29.00)	
Physician (Stage 1&2)	0.81 (0.73-0.90)	0.44 (0.28-0.59)	16.72 (4.99-77.16)	0.31
Physician (Stage 3)	0.76 (0.69-0.83)	0.12 (0.07-0.17)	11.00 (6.11-19.87)	
Physician (Stage 4)	0.72 (0.64-0.79)	0.03(0.02-0.04)	20.58 (11.88-35.90)	
DTS (Stage 1&2)	0.85 (0.77-0.93)	0.49 (0.29-0.73)	6.25 (1.49-27.97)	0.93
DTS (Stage 3)	0.77 (0.70-0.84)	0.11(0.07-0.15)	4.49 (1.49-10.97)	
DTS (Stage 4)	0.80 (0.73-0.86)	0.08 (0.05-0.11)	5.05 (2.05-10.69)	
History of coronary artery revascularization				
AI (None)	0.84 (0.80-0.88)	0.26 (0.18-0.35)	21.79 (14.59-32.80)	0.27
AI (At least one)	0.84 (0.76-0.93)	0.23 (0.05-0.44)	12.09 (4.67-33.41)	
Physician (None)	0.80 (0.76-0.85)	0.12 (0.09-0.16)	21.31 (14.27-32.04)	0.55
Physician (At least one)	0.80 (0.70-0.91)	0.08 (0.04-0.14)	15.50 (6.00-43.01)	
DTS (None)	0.78 (0.74-0.83)	0.07 (0.05-0.09)	4.40(2.25-7.90)	0.06
DTS (At least one)	0.74 (0.60-0.89)	0.08 (0.03-0.15)	15.04 (4.56-43.26)	

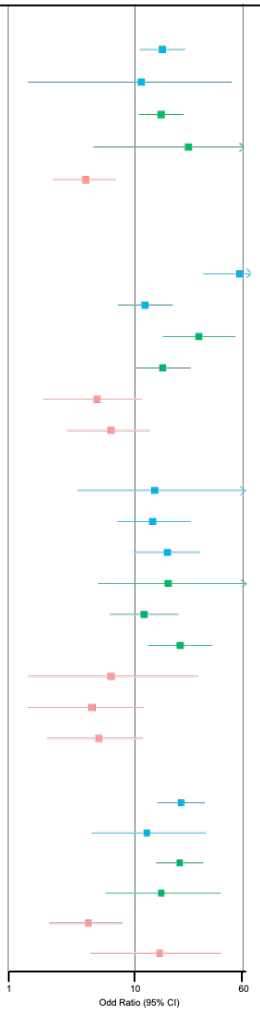


Figure 16. Subgroup analysis

The numbers of subjects in each subgroup are as follows: Male (N=4,170), Female (N=2,261), age < 60 years (N=2,829), age ≥ 60 years (N=3,552), Stage 1&2 (N=99), Stage 3 (N=1,668), Stage 4 (N=4,664), no comorbidities (N=5,370), and at least one comorbidity (N=1,061).

Abbreviations: AI, artificial intelligence; DTS, duke treadmill score; AUROC, the area under the receiver operating curve; AUPRC, the area under the precision recall curve; CI, confidence interval.

3.6. Evaluation of clinical validation

Without AI assisted, physicians achieved a microaveraged sensitivity of 0.75 (95% CI, 0.59-0.91), specificity of 0.38 (95% CI, 0.34-0.43), and an accuracy of 0.41 (95% CI, 0.36-0.46). With AI assisted, the physicians achieved a microaveraged sensitivity of 0.43 (95% CI, 0.25-0.61), specificity of 0.66 (95% CI, 0.61-0.71), and an accuracy of 0.64 (95% CI, 0.60-0.69). The underlying model had a sensitivity of 0.57 (95% CI, 0.21-0.94), specificity of 87 (95% CI, 0.80-0.94), and accuracy of 0.85 (95% CI, 0.78-0.92). Performance improvements across clinicians are detailed in the Table 3.

Table 7. Clinical Performance Metrics with and without AI assisted

Metric	Physicians	AI assisted physicians	Mean Increase (95% CI)
Sensitivity	0.750	0.429	-0.323 (-0.963-0.318)
Specificity	0.384	0.659	0.278 (-0.161-0.716)
Accuracy	0.410	0.643	0.233 (-0.182-0.647)
PPV	0.094	0.116	0.022 (-0.123-0.167
NPV	0.959	0.946	-0.013 (-0.123- 0.098)
F1-score	0.162	0.139	-0.024 (-0.177-0.129)

Abbreviations: AI, artificial intelligent; CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value.

4. Discussion

In this study, ExECG pre-trained VAE model was developed by using large-scale ExECG data for the first time compared to a prior study. The explainable AI model that utilizes ECG morphological features across a variety of exercise Stages have demonstrated strong prediction performance for coronary revascularization. In the test dataset, the model achieved AUROCs of 0.84 and AUPRCs of 0.25, outperforming both physician interpretation (AUROCs: 0.80, AUPRCs: 0.11), and DTS (AUROCs: 0.78, AUPRCs: 0.07). Incorporating SHAP and latent traversal into the AI model supports model interpretability, fostering clinical trust and adoption. Importantly, the AI identified important ECG markers, including subtle ST-segment deviations and heart rate recovery patterns, which were strongly associated with significant coronary disease. The findings in this study demonstrate that AI assisted ExECG can be used to improve risk stratification in coronary revascularization.

4.1. Pre-trained ExECG Model based on VAE

Several studies have demonstrated that VAE are effective for ECG compression, augmentation, clustering, and feature extraction, with several factors sufficient to encode a single or median beat ECG [37-41]. In addition, Van de Leur et al. (2022) demonstrated ECG morphology by a limited number of underlying factors and the median beat ECG can be encoded effectively using 21 continuous latent factors[26]. They improved the clinical utility and interpretability of VAE-derived features by relating them to established ECG measurements, integrating visualization tools, and validating them in predictive tasks.

The VAE approach was applied to ExECG in the present study to extract latent explanatory factors corresponding to each exercise Stage. Based on the modeling of dynamic morphological changes across phases of ExECG, each Stage-specific VAE model demonstrated high reconstruction accuracy, with Pearson correlation coefficients ranging from 0.933 to 0.940 ($P < 0.001$). This suggests that high-dimensional ExECG data can be effectively represented using only a small number of latent variables. Specifically, 16 latent variables were used in Stage 1, 14 in both Stage 2 and Stage 3, 13

in Stage 4, and 17 during recovery, indicating that ECG morphology can be well captured by a limited set of underlying factors (Figure 7-11).

4.2. Explainable AI for ExECG Interpretation

The present study extends ExECG AI research by using ECG morphological features and introducing explainable AI, which can provide deeper insights into signal interpretation and improve transparency in clinical prediction. Unlike conventional post hoc explainability methods commonly used to address the “black box” nature of deep learning in ECG analysis[31], the VAE-based approach in this study enables reliable and quantitative characterization of morphological changes in the ECG, rather than merely highlighting their temporal locations[17-19,26]. To improve transparency and clinical applicability, the SHAP framework was employed. SHAP provides quantitative attribution of individual feature contributions to model predictions and has demonstrated superior consistency and generalizability across diverse medical datasets compared to earlier interpretability methods[42,43]. This method is particularly useful in ExECG, where transient ECG changes occur dynamically across different exercise Stages.

The most significant predictor of coronary revascularization according to SHAP-based analysis was ST-segment depression during the mid-recovery phase. The recovery phase latent waveform features have been previously under-recognized as predictive markers for myocardial ischemia. The findings demonstrate the potential of explainable AI to provide both accurate prediction and physiologically relevant insights, thereby enhancing ExECG-based risk stratification.

4.3. AI-based ExECG Prediction of Coronary Revascularization

Using AI-based ExECG interpretation, this study provides direct evidence that prediction on coronary revascularization is comparable or even superior to interpretation of physician. Traditionally, physician rely on established criteria to assess ExECG signals, with an AUROC of 0.80, whereas my AI model achieved an AUROC of 0.84, thereby extracting greater predictive information from ExECG signals. Nevertheless, external validation indicated limited generalizability across institutional settings. This discrepancy may be attributed to heterogeneity in data acquisition protocols, population

characteristics, or underlying clinical workflows across institutions. Notably, even the DTS—a traditional risk stratification tool based on fixed, well-defined parameters and using the same revascularization outcome—also exhibited decreased performance in the external validation dataset. This indicates that the observed performance drop is not solely due to the complexity or overfitting of the AI model. Rather, it reflects the variability across institutions like patient selection, or clinical decision-making may significantly impact outcome labeling and model generalizability[44,45]. While the AI-based model showed reduced performance in the external validation, it consistently outperformed the traditional DTS-based approach. Accordingly, the model captures subtle morphological and dynamic features in ExECG that conventional scoring systems may ignore, making it an effective tool for detecting disease across a variety of clinical environments.

Diagnostic performance of AI was better in men compared with women. There were also gender differences in the conventional treadmill test algorithm[46]. There may be a reason for the gender differences observed in this study, since many of the features of the present model were associated with ST-segment depression. The diagnostic performance of ST-segment depression during TET is lower for women[46,47]. Previous studies have reported a relatively high prevalence of recovery-only ST-segment depression among asymptomatic, apparently healthy individuals[48], which may explain the superior model performance observed in younger patients (under 60 years). The interaction P-value below 0.01 indicates a significant performance difference by age, suggesting that the AI model may be especially effective in younger patients. By achieved exercise Stage, the prediction performance was highest in STAGE 1 and STAGE 2 subgroups compared to STAGE 3 and STAGE 4. Clinically, exercise tolerance during ExECG is an important indicator of underlying CAD[12,47]. Therefore, lower achieved Stages may reflect reduced functional capacity or the presence of CAD. The model appears to capture these clinical characteristics well and contributes to risk estimation.

4.4. AI-Based ExECG vs. Traditional Physician Interpretation

AI assistance in clinical decision-making substantially improved physician specificity (from 0.38 to 0.66) while reducing sensitivity (from 0.75 to 0.43). There is a possibility that the AI model prioritized specificity over sensitivity, resulting in fewer positive classifications. The threshold selection should reflect the intended clinical purpose of the

model. In screening, sensitivity should be maximized to detect as many true positives as possible, whereas in diagnostic settings, specificity should be prioritized to minimize false positives and avoid unnecessary interventions[15,49-51]. Since the primary purpose of this study was to evaluate an explainable AI framework and compare its predictive capability with physician interpretations, the threshold was chosen to reflect this objective. The thresholds of AI model may be revised according to their intended clinical use-diagnoses or screening decisions. Physicians also may have underutilized AI recommendations, particularly in borderline positive cases. The degree to which physicians accept AI recommendations is often determined by the transparency of the AI system, its interpretability, and the extent to which it is incorporated into the clinical workflow[52]. When physicians do not have enough information or do not trust the AI's performance, they may disregard AI recommendations and rely instead on their own clinical judgment[53]. In this study, physicians were not informed about the AI model's validated performance; as a result, they may have been reluctant to override their initial clinical judgment when AI predictions contradicted their expectations.

4.5. Limitations

Despite promising findings, this study has several limitations. Even though external validation was conducted, further studies, including multicenter validation and prospective trials, are required to ensure that the model is applicable and generalizable across diverse ethnic groups and populations. Due to its clinical relevance and retrospective design, coronary revascularization was selected as the primary endpoint, even though not all patients underwent invasive angiography. It could have missed cases of true ischemia in patients who were not revascularized, thereby underestimating its diagnostic accuracy. Furthermore, future research should explore the potential clinical utility of ExECG AI framework by exploring its application in other relevant contexts, including early detection of autonomic dysfunction, longitudinal monitoring, and risk stratification for major cardiovascular events. To ensure safe and effective implementation of AI assisted ExECG analysis into routine clinical practice, a comprehensive physician education program, clear regulatory pathways, and efficient workflow integration strategies are needed to ensure that it is safe and effective.

5. Conclusion

Based on a large-scale dataset of ExECG signals, this study presents the first pre-trained model designed to extract and represent ExECG signal characteristics, enabling task-specific, explainable AI frameworks. The framework improved prediction performance for coronary revascularization and reduced interobserver variability by providing visualization. Specifically, the dynamic morphological and temporal patterns of ExECG have been captured in the model, and they are showing strong alignment with clinically relevant features. The integration of explainable AI into clinical workflows could represent a significant advance in cardiovascular diagnostics, improving patient outcomes through more precise and personalized risk assessment. There will be a need for further validation and integration efforts to ensure successful clinical deployment and widespread adoption.

References

1 Roth, G. A. *et al.* Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. *Journal of the American college of cardiology* **76**, 2982-3021 (2020).

2 Lee, H.-H. *et al.* Korea heart disease fact sheet 2020: analysis of nationwide data. *Korean circulation journal* **51**, 495-503 (2021).

3 Herman, A. Coronary artery disease: The plaque plague. *Nursing made Incredibly Easy* **11**, 34-43 (2013).

4 Members, W. C. *et al.* 2021 ACC/AHA/SCAI guideline for coronary artery revascularization: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Journal of the American College of Cardiology* **79**, e21-e129 (2022).

5 Cassar, A., Holmes Jr, D. R., Rihal, C. S. & Gersh, B. J. in *Mayo Clinic Proceedings.* 1130-1146 (Elsevier).

6 Cagle, S. D. & Cooperstein, N. Coronary artery disease: diagnosis and management. *Primary Care: Clinics in Office Practice* **45**, 45-61 (2018).

7 Members, C. *et al.* ACC/AHA 2002 guideline update for exercise testing: summary article: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee to Update the 1997 Exercise Testing Guidelines). *Journal of the American College of Cardiology* **40**, 1531-1540 (2002).

8 Knuuti, J. *et al.* 2019 ESC Guidelines for the diagnosis and management of chronic coronary syndromes: The Task Force for the diagnosis and management of chronic coronary syndromes of the European Society of Cardiology (ESC). *European heart journal* **41**, 407-477 (2020).

9 Members, W. C. *et al.* 2021 AHA/ACC/ASE/CHEST/SAEM/SCCT/SCMR guideline for the evaluation and diagnosis of chest pain: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Journal of the American College of Cardiology* **78**, e187-e285 (2021).

10 Banerjee, A., Newman, D. R., Van den Bruel, A. & Heneghan, C. Diagnostic accuracy of exercise stress testing for coronary artery disease: a systematic review and meta-analysis of prospective studies. *International journal of clinical practice* **66**, 477-492 (2012).

11 Knuuti, J. *et al.* The performance of non-invasive tests to rule-in and rule-out significant coronary artery stenosis in patients with stable angina: a meta-analysis focused on post-test disease probability. *European heart journal* **39**, 3322-3330 (2018).

12 Mark DB, H. M., Harrell FE Jr, Lee KL, Califf RM, Pryor DB. Exercise Treadmill Score for Predicting Prognosis in Coronary Artery Disease. *Annals of Internal Medicine* **106**, 793-800 (1987). https://doi.org/10.7326/0003-4819-106-6-793 %m 3579066

13 Mark, D. B. *et al.* Prognostic value of a treadmill exercise score in outpatients with suspected coronary artery disease. *New England Journal of Medicine* **325**, 849-853 (1991).

14 Ribeiro, A. H. *et al.* Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature communications* **11**, 1760 (2020).

15 Attia, Z. I. *et al.* An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *The Lancet* **394**, 861-867 (2019).

16 Raghunath, S. *et al.* Deep neural networks can predict mortality from 12-lead

electrocardiogram voltage data. *arXiv preprint arXiv:1904.07032* (2019).
17 Raghunath, S. *et al.* Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nature medicine* **26**, 886-891 (2020).
18 Hempel, P. *et al.* Explainable AI associates ECG aging effects with increased cardiovascular risk in a longitudinal population study. *npj Digital Medicine* **8**, 25 (2025).
19 Cho, S. *et al.* Artificial intelligence–derived electrocardiographic aging and risk of atrial fibrillation: a multi-national study. *European heart journal* **46**, 839-852 (2025).
20 Hannun, A. Y. *et al.* Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine* **25**, 65-69 (2019).
21 Bailly, A. *et al.* Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models. *Computer Methods and Programs in Biomedicine* **213**, 106504 (2022).
22 Fan, J. & Li, R. in *Proceedings of the international Congress of Mathematicians.* 595-622 (European Mathematical Society Zurich).
23 Kim, H. E. *et al.* Transfer learning for medical image classification: a literature review. *BMC medical imaging* **22**, 69 (2022).
24 Arrieta, A. B. *et al.* Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* **58**, 82-115 (2020).
25 Yang, G., Ye, Q. & Xia, J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information Fusion* **77**, 29-52 (2022).
26 van de Leur, R. R. *et al.* Improving explainability of deep neural network-based electrocardiogram interpretation using variational auto-encoders. *European Heart Journal-Digital Health* **3**, 390-404 (2022).
27 Wouters, P. C. *et al.* Electrocardiogram-based deep learning improves outcome prediction following cardiac resynchronization therapy. *European heart journal* **44**, 680-692 (2023).
28 Lee, Y.-H. *et al.* Machine learning of treadmill exercise test to improve selection for testing for coronary artery disease. *Atherosclerosis* **340**, 23-27 (2022).
29 Yilmaz, A. *et al.* Machine learning approach on high risk treadmill exercise test to predict obstructive coronary artery disease by using P, QRS, and T waves' features. *Current Problems in Cardiology* **48**, 101482 (2023).
30 Bock, C. *et al.* Enhancing the diagnosis of functionally relevant coronary artery disease with machine learning. *Nature Communications* **15**, 5034 (2024).
31 Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* **1**, 206-215 (2019).
32 Healthcare, G. Marquette 12SL ECG Analysis Program: Physician's Guide. *GE Healthcare: Chicago, IL, USA* (2008).
33 Kingma, D. P. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
34 Snoek, J., Larochelle, H. & Adams, R. P. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems* **25** (2012).
35 Chen, T. & Guestrin, C. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.* 785-794.
36 Lundberg, S. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874* (2017).
37 Yildirim, O., San Tan, R. & Acharya, U. R. An efficient compression of ECG signals using deep convolutional autoencoders. *Cognitive Systems Research* **52**, 198-211 (2018).

38 Kuznetsov, V., Moskalenko, V., Gribanov, D. & Zolotykh, N. Y. Interpretable feature generation in ECG using a variational autoencoder. *Frontiers in genetics* **12**, 638191 (2021).
39 Jang, J.-H., Kim, T. Y., Lim, H.-S. & Yoon, D. Unsupervised feature learning for electrocardiogram data using the convolutional variational autoencoder. *PLoS One* **16**, e0260612 (2021).
40 Singh, P. & Sharma, A. Attention-based convolutional denoising autoencoder for two-lead ECG denoising and arrhythmia classification. *IEEE Transactions on Instrumentation and Measurement* **71**, 1-10 (2022).
41 Arslan, N. N., Ozdemir, D. & Temurtas, H. ECG heartbeats classification with dilated convolutional autoencoder. *Signal, Image and Video Processing* **18**, 417-426 (2024).
42 Ayano, Y. M., Schwenker, F., Dufera, B. D. & Debelee, T. G. Interpretable machine learning techniques in ECG-based heart disease classification: a systematic review. *Diagnostics* **13**, 111 (2022).
43 Anand, A., Kadian, T., Shetty, M. K. & Gupta, A. Explainable AI decision model for ECG data of cardiac disorders. *Biomedical Signal Processing and Control* **75**, 103584 (2022).
44 Ramspek, C. L., Jager, K. J., Dekker, F. W., Zoccali, C. & van Diepen, M. External validation of prognostic models: what, why, how, when and where? *Clinical kidney journal* **14**, 49-58 (2021).
45 Van Calster, B., Steyerberg, E. W., Wynants, L. & Van Smeden, M. There is no such thing as a validated prediction model. *BMC medicine* **21**, 70 (2023).
46 Kwok, Y., Kim, C., Grady, D., Segal, M. & Redberg, R. Meta-analysis of exercise testing to detect coronary artery disease in women. *The American journal of cardiology* **83**, 660-666 (1999).
47 Gibbons, R. J. *et al.* ACC/AHA guidelines for exercise testing. A report of the American College of Cardiology/American Heart Association task force on practice guidelines (Committee on Exercise Testing). *Journal of the American College of Cardiology* **30**, 260-311 (1997).
48 Lanza, G. *et al.* Diagnostic and prognostic value of ST segment depression limited to the recovery phase of exercise stress test. *Heart* **90**, 1417-1421 (2004).
49 Vijan, S., Hofer, T. P. & Hayward, R. A. Cost-utility analysis of screening intervals for diabetic retinopathy in patients with type 2 diabetes mellitus. *Jama* **283**, 889-896 (2000).
50 Areia, M. *et al.* Cost-effectiveness of artificial intelligence for screening colonoscopy: a modelling study. *The Lancet Digital Health* **4**, e436-e444 (2022).
51 Wang, Y. *et al.* Economic evaluation for medical artificial intelligence: accuracy vs. cost-effectiveness in a diabetic retinopathy screening case. *NPJ Digital Medicine* **7**, 43 (2024).
52 Rosenbacke, R., Melhus, Å., McKee, M. & Stuckler, D. How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Health Care: Systematic Review. *JMIR AI* **3**, e53207 (2024).
53 Zheng, R. *et al.* Investigating Clinicians' Intentions and Influencing Factors for Using an Intelligence-Enabled Diagnostic Clinical Decision Support System in Health Care Systems: Cross-Sectional Survey. *Journal of Medical Internet Research* **27**, e62732 (2025).

ABSTRACT IN KOREAN

관상동맥 혈관재형성술 예측 네트워크 개발 연구: 운동부하심전도 사전 학습 모델 구축

배경: 운동부하 심전도(ExECG)는 관상동맥질환 평가에 널리 사용되지만, 진단 정확도의 변동성이 커 해석에 어려움이 있다. 본 연구에서는 운동부하 심전도 기반으로 관상동맥 재혈관술이 필요한 환자를 예측할 수 있는 설명 가능한 인공지능(AI) 모델을 개발하고 검증하고자 한다.

방법: 브루스 프로토콜을 사용한 운동부하 심전도 검사를 받은 20,534명의 환자를 대상으로 연구를 수행하였다. 변분 오토인코더(variational autoencoder)를 활용해 운동부하 심전도 상 중요한 심전도 특징을 먼저 학습한 후, 운동부하 심전도 검사 이후 90일 이내에 시행된 경피적 관상동맥중재술(PCI) 또는 관상동맥우회술(CABG)을 '관상동맥 재혈관술'로 정의하여 이에 대한 예측 모델을 학습하였다. 모델의 성능은 임상의 판단과 Duke Treadmill Score와 비교 평가하였다.

결과: 본 모델은 수신자 조작 특성 곡선 아래 면적(AUROC)은 0.84 (신뢰구간 95%, 0.80–0.88)로 우수한 성능을 보였으며, 임상의 판단은 0.75 (신뢰구간 95% , 0.71–0.80), Duke Treadmill Score은 0.78 (신뢰구간 95% , 0.73–0.82) 보였다. AI 기반의 관상동맥 재혈관술 고위험군의 오즈비는 12.37(8.43-18.49)인 반면, Duke Treadmill Score와 의사 진단으로 기반으 오즈비 각각 5.65(3.02-9.40)와 19.65(13.56-28.65)이었다. 특히, 회복기 중간 단계에서의 ST분절 하강이 관상동맥 재혈관화 필요성의 가장 중요한 예측 인자임을 보였다.

결론: 운동부하 심전도를 활용하여 관상동맥 재혈관술을 예측하는 설명 가능한 인공지능 모델을 개발하고 검증하였습니다. 고도화된 AI 예측력과 해석 가능한 심전도 특징 제시함으로써, 임상 현장에서 운동부하 심전도의 진단적 유용성을 향상시킬 수 있을 것으로 기대된다.

핵심되는 말: 운동부하 심전도, 인공지능, 변분 오토인코더, 관상동맥 재혈관술