

Machine Learning-Based Risk Stratification Model to Guide Cardiopulmonary Exercise Training in Cardiovascular Patients

Ha Eun Na¹, Gi Pyo Lee², Young Jae Kim³, Kwang Gi Kim⁴, and Ju Kang Lee⁵

¹Department of Physical Medicine and Rehabilitation, Gachon University Gil Medical Center, Incheon;

²Department of Biomedical Engineering, Gachon University, Seongnam;

³Gachon Biomedical & Convergence Institute, Gachon University Gil Medical Center, Incheon;

⁴Department of Biomedical Engineering, Gachon University Gil Medical Center, Gachon University College of Medicine, Incheon;

⁵Department of Physical Medicine and Rehabilitation, Gachon University Gil Medical Center, Gachon University College of Medicine, Incheon, Korea.

Purpose: In cardiovascular patients, precise risk stratification is crucial for minimizing exercise-related risks and optimizing treatment efficacy during cardiac rehabilitation. Existing classification criteria vary among organizations, necessitating a comprehensive assessment of the underlying conditions and medical tests. We developed machine-learning models to assist in standardizing risk classification for cardiopulmonary exercise training.

Materials and Methods: Using retrospective data from 1163 patients across three institutions, we developed an AI model to reproduce clinician-assigned risk stratification for exercise-related cardiovascular events. Data pre-processing involved excluding variables with numerous missing values and conducting multiple imputation using chained equations, resulting in 46 variables for analysis. Logistic regression, support vector machines, extreme gradient boosting (XGBoost), and random forests were employed, optimizing hyperparameters via a grid search. In feature selection, the least absolute shrinkage and selection operator (LASSO), with a range of regularization strengths, was applied to the L1 term to determine the optimal penalty value.

Results: Several machine-learning models were evaluated for their ability to replicate clinician-assigned cardiovascular risk-type classifications. XGBoost achieved the best performance with 76.72% accuracy and an area under the curve (AUC) of 88.76%. After applying LASSO-based feature selection, performance improved, with the best result ($\alpha=0.001$) showing 79.31% accuracy and an AUC of 89.40%. Feature importance analysis identified ventilatory efficiency, maximal systolic blood pressure, and maximal oxygen consumption as key features for risk stratification based on aerobic exercise load test data, along with weight.

Conclusion: A machine-learning model was established to categorize exercise-related risk types during cardiopulmonary rehabilitation, serving as a decision-support tool for the safe and effective training of cardiovascular patients.

Key Words: Cardiac rehabilitation, artificial intelligence, machine learning

Received: September 22, 2025 **Revised:** March 6, 2026 **Accepted:** March 31, 2026 **Published online:** June 9, 2026

Co-corresponding authors: Kwang Gi Kim, PhD, Department of Biomedical Engineering, Gachon University Gil Medical Center, Gachon University College of Medicine, 21 Namdong-daero 774beon-gil, Namdong-gu, Incheon 21565, Korea.

E-mail: kimkg@gachon.ac.kr and

Ju Kang Lee, MD, PhD, Department of Physical Medicine and Rehabilitation, Gachon University Gil Medical Center, Gachon University College of Medicine, 21 Namdong-daero 774beon-gil, Namdong-gu, Incheon 21565, Korea.

E-mail: pmrdoc@gilhospital.com

•The authors have no potential conflicts of interest to disclose.

© Copyright: Yonsei University College of Medicine 2026

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

INTRODUCTION

Cardiovascular diseases (CVD) are a leading cause of mortality in South Korea, with the death rate increasing from 52.5 to 65.8 per 100000 individuals between 2012 and 2022.¹ Since 2014, CVD has been consistently ranked as the second leading cause of death, with ischemic heart diseases, such as myocardial infarction, angina, and heart failure, accounting for a significant proportion.¹ Despite advances in acute treatment, the risks of sudden cardiac death and heart failure remain significant, adversely affecting long-term prognosis and quality of life.^{2,3}

Cardiac rehabilitation (CR) is an evidence-based intervention that improves patient survival and reduces disease recurrence. Major guidelines from the American College of Cardiology and American Heart Association (AHA) recommend CR with supervised exercise in patients with myocardial infarction, coronary revascularization, heart failure, or heart transplantation.⁴⁻⁶ Participation in CR has reduced 5-year mortality by 45%–47%,⁷ leading to its adoption and insurance coverage in several countries. In South Korea, CR programs are available at multiple institutions, including government-supported cardiovascular centers established in 2008.⁸

Recent developments in artificial intelligence (AI) have enabled real-time monitoring, exercise correction, and the integration of therapeutic devices for CR. AI-driven tools using biometric data and mobile applications are being applied to support individualized rehabilitation.⁹⁻¹¹ However, accurate risk stratification is essential for tailoring exercise prescriptions. Calculated parameters, such as target heart rate (HR) and target metabolic equivalent of task (MET), based on the Karvonen formula, are used to estimate individual risk levels and guide clinicians in determining appropriate exercise intensity, duration, and monitoring strategies.

Risk classification criteria vary among the guidelines published by the AHA, the American Association of Cardiovascular and Pulmonary Rehabilitation (AACVPR), the American College of Sports Medicine, and other national societies.^{12,13} These differences include the presence or absence of stress testing, inclusion of comorbidities, and thresholds for left ventricular ejection fraction (EF).⁴ For example, the AACVPR defines high risk as EF <40%, whereas the European Society of Cardiology defines it as <35%. The 2019 Korean guidelines for CR for acute coronary syndrome also differ, contributing to clinical inconsistencies.

To address these challenges, this study aimed to develop a machine learning-based model to standardize the risk stratification process by learning from expert clinical interpretations. Using a dataset in which risk levels were assigned by clinical specialists in accordance with the AACVPR guidelines, we trained an AI model to automate and reproduce expert-driven risk classification (Fig. 1). The primary objective was to establish a reliable clinical decision-support tool that ensures consistent and objective risk stratification, thereby supporting safe and individualized exercise prescription in cardiopulmonary rehabilitation settings. The performance of the proposed model was subsequently evaluated by comparing its concordance with clinician-assigned risk classifications in a real-world clinical setting.

MATERIALS AND METHODS

Data

This study was conducted using muscle exercise prescription data provided by the National Information Society Agency, South Korea, through the AI-Hub platform.¹⁴ The dataset used in this study was obtained from the AI-Hub platform, a national

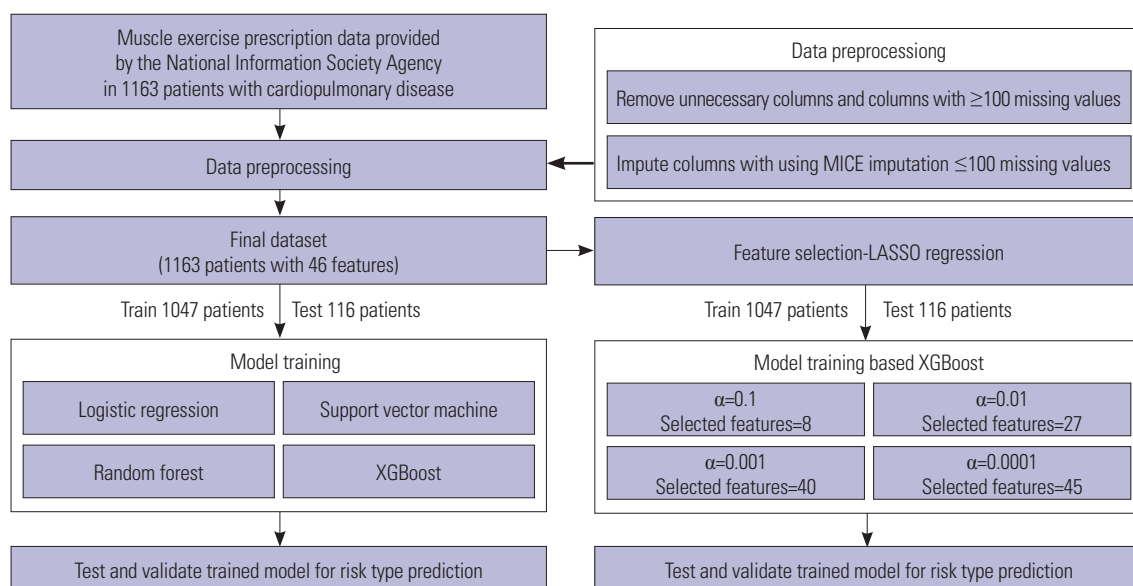


Fig. 1. Flowchart of study population. MICE, multiple imputation by chained equations; LASSO, least absolute shrinkage and selection operator.

data repository supported by the Ministry of Science and ICT of Korea. Detailed information regarding the data structure, variable definitions, and data collection protocols was obtained from the official dataset documentation provided by AI-Hub. The dataset was established in 2022 to prescribe resistance exercises to individuals with cardiovascular, respiratory, and musculoskeletal disorders, as well as healthy individuals. In this study, we retrospectively analyzed data collected from patients with CVD. The study protocol was reviewed and approved by the Institutional Review Board of the Gachon University Gil Medical Center (IRB No. GCIRB2024-326; approval date: November 15, 2024).

A dataset on CVDs was constructed retrospectively by selecting patients diagnosed with cardiovascular conditions at three medical institutions: Korea University Anam Hospital, Pusan National University Hospital, and Jeju National University Hospital. Specifically, data were collected from 303 patients at Korea University Anam Hospital (162 new patients and 141 returning patients), 476 patients at Pusan National University Hospital (199 new and 277 returning patients), and 384 patients at Jeju National University Hospital (230 new and 154 returning patients), totaling 1163 individuals.

A total of 147 variables were included in the initial feature set. Information regarding the dataset, including variable definitions and data structure, was verified using the official dataset documentation provided by AI-Hub,¹⁴ which was downloaded and reviewed prior to analysis. The analyzed variables were organized into clinically relevant domains and included demographic characteristics (age, sex), medical history (hypertension, diabetes mellitus, coronary artery disease, heart failure, prior revascularization), anthropometric measurements (height, weight, body mass index), questionnaire-derived variables related to physical activity and symptoms, cardiopulmonary exercise testing (CPET) parameters [peak oxygen uptake, oxygen uptake at anaerobic threshold, ventilatory efficiency (VE/VCO₂ slope), peak heart rate (PHR), blood pressure response]. A detailed list of all analyzed variables and their clinical definitions is provided in Supplementary Table 1 (only online).

“Risk type” was categorized into three levels: low, moderate, and high, which are commonly used for risk stratification of exercise-related cardiovascular events. Risk classification was performed in accordance with the guidelines of the AACVPR, based on a comprehensive assessment integrating CPET results and relevant clinical information. According to the AACVPR criteria, individuals were classified as high risk or moderate risk if they met at least one criterion corresponding to the respective risk category. In contrast, individuals were classified as low risk only if all low-risk criteria were satisfied and no moderate- or high-risk criteria were present. Using this guideline-based clinician assessment, 214 individuals were categorized as low risk, 650 as moderate risk, and 299 as high risk.

Expected outcomes defined for the risk stratification model

The target HR and target MET for each patient were calculated using the Karvonen and MET formulas, respectively.^{15,16} The Karvonen method determines the target HR during exercise based on the concept of heart rate reserve (HRR), which is the difference between the PHR, measured during a test exercise, and the resting heart rate (HR_{rest}). Similarly, the target MET is calculated by multiplying the proportionality constant by the difference between 1 and the peak MET values obtained from the exercise test.

Both formulae require a proportional constant, kappa (κ), which reflects the exercise intensity level based on the individual's risk stratification. Patients with higher risk levels are assigned lower exercise intensities, whereas those with lower risk levels are prescribed relatively higher intensities. The κ value typically ranges from 0.40 to 0.85 and is individualized for each patient. In this study, κ was determined for each patient using a machine learning-based risk stratification model that categorizes risk levels.

The target HR was calculated as follows:

$$\text{Target HR} = [(\text{HR}_{\text{max}} - \text{HR}_{\text{rest}}) \times \kappa] + \text{HR}_{\text{rest}} \quad (1)$$

The target MET was calculated as follows:

$$\text{Target MET} = 1 \text{ MET} + (\text{peak MET} - 1) \times \kappa \quad (2)$$

Where HR refers to HR, MET refers to metabolic equivalent, and κ is the exercise intensity coefficient determined by cardiovascular risk level. The κ values corresponding to risk categories are as follows: 1) low risk 0.70–0.85; 2) moderate risk 0.55–0.70; and 3) high risk 0.40–0.55.

Data preprocessing

The original dataset comprised 147 variables. After excluding the columns related to dataset information, patient identity numbers, and labels unsuitable for model training, 141 variables remained. Additionally, four variables representing the maximum and minimum recommended METs and target HRs for aerobic exercise, derived based on the κ values of risk classification, were excluded, resulting in 137 variables for analysis.

Due to the retrospective nature of data collection, some variables had missing values, particularly those derived from patient questionnaires. As machine-learning algorithms cannot process missing data directly, a series of preprocessing steps was undertaken to address this issue. Variables with more than 100 missing values (88 variables) were removed, along with two variables with identical values across all observations. Consequently, 46 variables were retained for model development, with risk type designated as the target label.

For the 12 variables with fewer than 100 missing values, missing entries were imputed using the multiple imputation by chained equations (MICE) method,¹⁷ specifically designed for

continuous data. MICE is a robust statistical technique that addresses missing data by iteratively modeling each variable to determine missing values as a function of other variables in the dataset. This approach generates plausible estimates by creating multiple imputed datasets that reflect the uncertainty inherent in the missing data. By leveraging the relationships among variables, MICE provides reliable and realistic imputations, thereby enhancing the integrity of the dataset for subsequent analysis. Through these preprocessing steps, a final training dataset was constructed with the imputed values to ensure its suitability for machine learning model development.

The constructed dataset was divided into training and test sets. The independent test set was strictly held out and used exclusively for final performance evaluation and was not involved in any stage of feature selection or model training. The dataset was split at an approximate ratio of 9:1. Specifically, data from 1047 individuals (approximately 90% of the dataset) were used for model development, including feature selection and model training, while data from 116 individuals (approximately 10%) were allocated to the independent test set. The test set was constructed by randomly sampling approximately 10% of individuals from each risk category to preserve the original class distribution, resulting in 26 low-risk, 65 moderate-risk, and 25 high-risk individuals.

Machine learning algorithms

To develop a classification model for risk-type categorization, four widely used machine-learning algorithms were employed: logistic regression (LR), support vector machine (SVM), extreme gradient boosting (XGBoost), and random forest (RF). LR¹⁸ is a linear model commonly applied to binary or multiclass classification tasks. It estimates the probability of class membership by applying a logistic function to a linear combination of input features. SVM¹⁹ is a supervised learning algorithm that identifies the optimal hyperplane that maximizes the margin between classes and employs kernel functions to address non-linear classification problems. XGBoost^{20,21} is an efficient and scalable implementation of the gradient boosting framework that sequentially constructs decision trees to minimize classification errors. It incorporates regularization techniques to reduce overfitting and enhance performance. RF²² is an ensemble learning method that generates multiple decision trees using random subsets of features and samples, with the final risk stratification determined by majority voting across all the trees.

Grid search for optimal hyperparameter selection

In this study, the hyperparameters for each machine learning model were optimized using the grid search^{23,24} method. Hyperparameters are predefined parameters that significantly influence model performance and must be set prior to the training process. A grid search is a systematic approach for identifying the optimal combination of hyperparameter values.²⁵ This involves specifying a discrete set of candidate values for each

hyperparameter and evaluating all possible combinations through model training and validation.

Cross-validation was performed for each combination to assess average model performance.²⁶ The combination that yielded the highest average accuracy was selected and applied to the final model. Using grid search, we defined the search space for the hyperparameters of each model and conducted cross-validation of the training data to identify the configuration that achieved the best classification performance. Table 1 summarizes the final hyperparameter configurations selected through grid search for each model.

Feature selection using LASSO

Feature selection was performed using the least absolute shrinkage and selection operator (LASSO)²⁷ method to enhance the stratification performance and interpretability of the training models. LASSO is a regularization technique for linear regression that improves both model interpretability and generalization by applying L1 regularization to penalize the absolute sum of regression coefficients. This penalty effectively shrinks some coefficients to zero, thereby performing variable selection and reducing multicollinearity among correlated clinical variables. As a result, LASSO yields a sparse and interpretable model, which is particularly advantageous for clinical decision-support applications. Given the relatively large number of candidate variables ($n=147$) and the moderate sample size of the dataset, inclusion of irrelevant or redundant variables could increase the risk of overfitting and reduce model generalizability. Feature selection was therefore applied to retain only variables that contributed meaningfully to model performance.

In this study, regression coefficients were initially estimated for all candidate variables, and variables with coefficients shrunk to zero were excluded from subsequent analyses. To identify the optimal level of regularization, a range of penalty

Table 1. Parameters for Each Machine Learning Model Selected Through Grid Search

Model	Hyperparameter	Selected value
Logistic regression	Penalty	L2
	Solver	LBFGS
	C (regularization)	1.0
Support vector machine	Kernel	RBF
	C	10
	Gamma	0.01
Random forest	Number of trees (n_estimators)	100
	Max depth	10
	Min samples split	2
Extreme gradient boosting	Learning rate (eta)	0.1
	Max depth	6
	Number of estimators	100
	Subsample	0.8

values (i.e., L1 regularization strengths) was evaluated. The number and types of features selected at each penalty level are summarized in Supplementary Table 2 (only online), which provides a complete list of the variables used for model development.

Model performance comparison

All four machine learning algorithms were trained using the full set of features from the training dataset to classify risk types according to clinician-assigned labels. To evaluate model performance, a confusion matrix was constructed by comparing the classified labels with the ground truth. Standard classification metrics, including true positives, true negatives, false positives, and false negatives, were derived from the confusion matrix.

Based on these values, the accuracy, precision, recall, and F1-score were determined to assess the classification performance of each model. In addition, receiver operating characteristic (ROC) curve analysis was performed, and the area under the curve (AUC) was computed to evaluate the discriminatory ability of the models. To address the black-box nature of machine learning models, we employed SHapley Additive exPlanations (SHAP) to clearly elucidate the model's logic in determining risk stratification.^{21,28} SHAP quantifies the contribution of each feature to risk stratification by applying the Shapley value principle from cooperative game theory, thereby enabling the visualization of feature importance and corresponding directional effects. Hence, we analyzed the SHAP values to assess how each variable contributed to the model's stratification decisions and evaluated whether the relationships among the clinically essential features aligned with those established in conventional medical analyses.

All data preprocessing, machine-learning model development, and performance evaluations were conducted using Python software (version 3.8.19; Python Software Foundation, Wilmington, DE, USA) along with Pandas (version 1.5.3) and Scikit-learn (version 1.2.2), setting significance at $p < 0.05$.

RESULTS

Participant characteristics

The demographic and clinical characteristics of the dataset used in this study, stratified by risk categories (low, moderate, and high), are presented in Tables 2 and 3. The p -values for continuous variables were derived from one-way analysis of variance, whereas those for categorical variables were obtained using the chi-square test.

We analyzed CR prescription data from 1163 patients, of whom 214 (18.39%) were classified as low risk, 650 (55.89%) as moderate risk, and 299 (25.71%) as high risk. The mean age was 59.14 ± 12.56 years, and 77.73% were male. Hypertension (64.66%), hyperlipidemia (44.63%), and diabetes mellitus (27.86%) were the most common comorbidities. Among comorbid conditions,

hypertension and hyperlipidemia were most frequently observed in the high-risk group and least frequent in the low-risk group. A history of cancer was also more common in the high-risk group. With respect to surgical history, prior cardiac surgery was considerably more frequent in the high-risk group, whereas non-cardiac surgeries were more frequently reported in the low-risk group.

Regarding cardiovascular diagnoses, ischemic heart disease was the most common overall, comprising non-ST-segment elevation myocardial infarction (24.94%), ST-segment elevation myocardial infarction (22.79%), and angina (17.28%). Congestive heart failure was present in 11.35% of the total cohort and occurred most frequently in the high-risk group. Similarly, atrial fibrillation was observed predominantly in the high-risk group. CPET results showed significant differences among the risk groups in terms of maximal oxygen consumption ($\text{VO}_2 \text{ max}$), maximal and HR_{rest}, predicted maximal % HR, MET, O_2 pulse, systolic blood pressure (SBP) (resting and maximal), exercise duration, rate pressure product, and perceived exertion. $\text{VO}_2 \text{ max}$ was lowest in the high-risk group (20.24 ± 7.09), followed by the moderate-risk group (25.46 ± 12.04) and the low-risk group (26.54 ± 5.20) ($p < 0.0001$).

Performance of the risk stratification model using the entire dataset

In this study, the classification performance of the risk stratification model was evaluated using a test dataset of 116 individuals. The machine learning models were trained using the full set of features, and their performance was validated by assessing their concordance with clinician-assigned risk types in an independent test dataset.

Table 4 lists the performance metrics of the models, including accuracy, precision, recall, F1-score, and AUC. Among the LR, SVM, XGBoost, and RF algorithms, XGBoost demonstrated the highest overall performance, achieving an accuracy of 76.72%, a precision of 79.82%, a recall of 69.33%, an F1-score of 72.66%, and an AUC of 88.76%. In terms of the AUC, XGBoost outperformed the other models, followed by RF (88.32%), SVM (84.76%), and LR (80.98%). These results indicate that XGBoost yielded the most accurate and balanced performance for classifying risk types across the test dataset. The ROC curve for the XGBoost model presented in Fig. 2 demonstrates the high discriminatory capacity of the model for classifying risk types. The AUC value of 88.76% further supports the robustness of this model compared with those of the other tested algorithms.

To investigate the contribution of individual features to the classification performance of the XGBoost model, the feature importance was assessed using the F-score, which reflects the frequency at which a feature is used to split nodes in decision trees; a higher score indicates greater influence on the model's risk-type categorization. The top 20 most important features identified using the XGBoost model are presented in Fig. 3. The x-axis represents the relative importance of each feature and

Table 2. Demographic and Clinical Characteristics of Participants According to Risk Type

Characteristic	Total (n=1163)	Risk type			p
		Low (n=214)	Moderate (n=650)	High (n=299)	
Age (yr)	59.14±12.56	58.13±12.42	59.32±12.21	59.47±13.40	0.4240
Sex (Male : Female)	904 : 259	146 : 68	528 : 122	230 : 69	0.0144
Height (cm)	165.44±8.67	164.75±8.96	165.70±8.78	165.38±8.18	0.3759
Body weight (kg)	68.89±12.97	67.77±11.18	69.58±13.07	68.19±13.87	0.1177
Body mass index (kg/m ²)	25.07±3.61	24.93±3.18	25.23±3.52	24.83±4.05	0.2333
Comorbidity					
Hypertension	752 (64.66)	100 (46.73)	437 (67.23)	215 (71.91)	<0.0001*
Diabetes	324 (27.86)	46 (21.50)	193 (29.69)	85 (28.43)	0.4879
Hyperlipidemia	519 (44.63)	64 (29.91)	302 (46.46)	153 (51.17)	0.0004*
Liver	25 (2.15)	8 (3.74)	12 (1.85)	5 (1.67)	0.7864
Lung disease	25 (2.15)	2 (0.93)	14 (2.15)	9 (3.01)	0.8623
Cancer	45 (3.87)	11 (5.14)	12 (1.85)	22 (7.36)	0.0066
Surgery/procedure					
Heart surgery	204 (17.54)	25 (11.68)	94 (14.46)	85 (28.43)	<0.0001*
Heart procedure	315 (27.09)	48 (22.43)	176 (27.08)	91 (30.43)	0.6703
Other surgeries	391 (33.62)	99 (46.26)	204 (31.38)	88 (29.43)	0.0039
Cardiovascular disease					
Angina	201 (17.28)	49 (22.90)	101 (15.54)	51 (17.06)	0.4107
NSTEMI	290 (24.94)	47 (21.96)	176 (27.08)	67 (22.41)	0.7275
STEMI	265 (22.79)	30 (14.02)	176 (27.08)	59 (19.73)	0.0069
Congestive heart failure	132 (11.35)	26 (12.15)	36 (5.54)	70 (23.41)	<0.0001*
Mitral stenosis	19 (1.63)	4 (1.87)	7 (1.08)	8 (2.68)	0.7641
Mitral regurgitation	63 (5.42)	18 (8.41)	32 (4.92)	13 (4.35)	0.5800
Aortic stenosis	29 (2.49)	0 (0.00)	22 (3.38)	7 (2.34)	0.2670
Aortic regurgitation	18 (1.55)	6 (2.80)	3 (0.46)	9 (3.01)	0.0756
Pulmonary stenosis	1 (0.09)	1 (0.47)	0 (0.00)	0 (0.00)	0.6176
Atrial fibrillation	91 (7.82)	5 (2.34)	44 (6.77)	42 (14.05)	0.0002*
Paroxysmal supraventricular tachycardia	1 (0.09)	1 (0.47)	0 (0.00)	0 (0.00)	0.6176
Arteriovenous block	5 (0.43)	0 (0.00)	5 (0.77)	0 (0.00)	0.6817
Others	48 (4.13)	14 (6.54)	19 (2.92)	15 (5.02)	0.4084
Aortic dissection	15 (1.29)	0 (0.00)	10 (1.54)	5 (1.67)	0.7499
Aortic aneurysm	6 (0.52)	0 (0.00)	1 (0.15)	5 (1.67)	0.1030

NSTEMI, non-ST-segment elevation myocardial infarction; STEMI, ST-segment elevation myocardial infarction.

Data are presented as mean±standard deviation or n (%).

*p<0.001.

not an absolute scale. Among all the features, the minute ventilation carbon dioxide output ratio (VE/VCO₂) exhibited the highest importance, followed by SBP max and VO₂ max. Other influential factors include exercise duration, predicted VO₂, and age.

The SHAP beeswarm plot for the high-risk class (Fig. 4) illustrates how individual feature values influence the model's assignment of the high-risk label. Each dot represents one patient; the actual value of the feature is shown in color, with red indicating higher values and blue indicating lower values; the horizontal position corresponds to the SHAP value. Positive SHAP values push the model output toward high-risk classification, whereas negative values push it toward low-risk categorization. Overall, VE/VCO₂, VO₂ max, HR_{rest}, and age are the principal determinants that push the model toward high-risk classification.

Performance of the risk stratification model via feature selection

To improve model interpretability and reduce overfitting, a risk stratification model was developed using a subset of features selected using a LASSO-based feature selection method. The degree of regularization was controlled by varying the α parameter. Feature selection was applied before model development using the XGBoost algorithm, which demonstrated the highest performance when trained on a full set of features.

Table 5 summarizes the performance metrics of the XGBoost models trained using different α values. Among the tested values, an α of 0.001 yielded the best overall performance, with an accuracy of 79.31%, precision of 82.07%, recall of 72.46%, F1-score of 75.49%, and AUC of 89.40%. These results suggest that

Table 3. Aerobic Exercise Load Test Results of the Participants According to Risk Type

Characteristic	Total (n=1163)	Risk type			p
		Low (n=214)	Moderate (n=650)	High (n=299)	
VO ₂ max (mL/kg/min)	24.32±10.24	26.54±5.20	25.46±12.04	20.24±7.09	<0.0001*
Maximal HR (bpm)	138.39±23.86	144.88±20.79	138.14±23.04	134.28±26.61	<0.0001*
Resting HR (bpm)	75.46±13.23	72.45±11.98	74.42±11.95	79.86±15.51	<0.0001*
% of predicted maximal HR	86.00±13.92	89.65±11.88	85.82±13.92	83.78±14.78	<0.0001*
Maximal MET	6.97±3.27	7.61±1.51	7.28±3.96	5.83±2.03	<0.0001*
Maximal ventilatory threshold (mL/kg/min)	21.37±66.20	23.91±69.01	22.89±78.87	16.25±10.18	0.2941
Minute ventilation-carbon dioxide output ratio	35.29±90.27	28.80±5.16	36.73±120.54	36.79±8.35	0.5084
Maximal oxygen pulse (mL/beat)	12.44±5.19	12.98±2.98	13.04±6.09	10.76±3.75	<0.0001*
Predicted VO ₂ max (%)	92.19±29.82	94.78±20.90	93.10±27.92	88.36±37.90	0.0280
Maximal SBP (mm Hg)	169.88±56.18	182.18±30.15	170.70±69.54	159.31±30.08	<0.0001*
Maximal DBP (mm Hg)	74.37±12.84	75.20±13.48	74.68±12.71	73.09±12.59	0.1185
Resting SBP (mm Hg)	122.10±16.92	126.19±16.24	121.67±16.91	120.12±16.97	0.0002*
Resting DBP (mm Hg)	75.00±11.84	76.84±12.11	74.76±11.78	74.21±11.66	0.0335
Exercise duration (sec)	769.96±353.84	861.72±148.06	798.46±421.78	642.34±246.88	<0.0001*
Maximal respiratory exchange ratio	1.19±2.98	1.12±0.10	1.27±3.98	1.06±0.20	0.5601
Maximal rate pressure product (mm Hg · beats/min)	23453.48±14955.83	26329.10±16456.37	23907.34±16927.42	20408.71±6236.69	<0.0001*
Maximal rate perceived exertion	17.32±1.88	17.19±1.98	17.66±1.70	16.68±2.02	<0.0001*

VO₂ max, maximal oxygen consumption; HR, heart rate; MET, metabolic equivalents of task; SBP, systolic blood pressure; DBP, diastolic blood pressure.

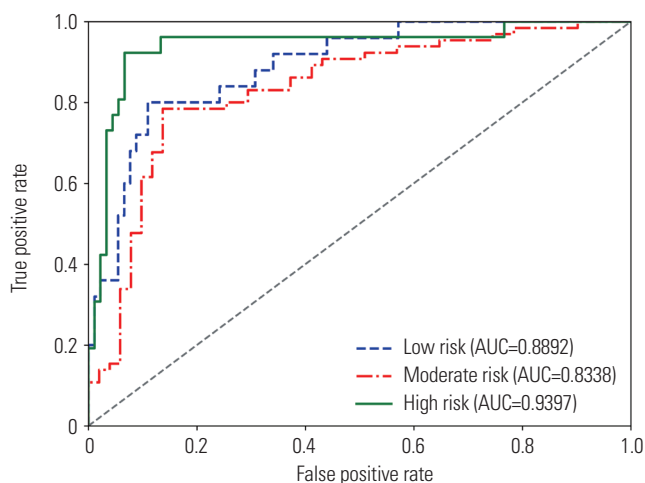
Data are presented as mean±standard deviation.

*p<0.001.

Table 4. Performance Metrics of the Risk Stratification Model Using the Entire Dataset

	Accuracy	Precision	Recall	F1-score	AUC
Logistic regression	66.38 (57.76–74.14)	65.42 (53.36–76.56)	58.46 (49.21–68.13)	60.66 (49.94–70.01)	80.98
Support vector machine	71.55 (63.79–80.17)	70.78 (59.30–81.37)	63.13 (54.02–71.96)	65.55 (55.74–74.95)	84.76
Extreme gradient boosting	76.72 (68.97–84.48)	79.82 (69.82–88.54)	69.33 (59.54–78.58)	72.66 (62.89–81.36)	88.76
Random forest	71.55 (62.93–79.31)	76.72 (64.68–87.29)	60.62 (51.80–70.11)	63.57 (53.67–72.64)	88.32

AUC, area under the curve.

**Fig. 2.** Receiver operating characteristic curve analysis of the extreme gradient boosting-based risk type prediction model. AUC, area under the curve.

moderate regularization balances the trade-off between model complexity and generalization, effectively reducing overfitting while retaining essential features. Conversely, a higher α value

(e.g., 0.1) imposed stronger regularization, substantially reducing model performance across all evaluation metrics. This indicates that excessive regularization may eliminate informative features and degrade the model's classification accuracy in identifying the expected outcomes.

The top 20 features most critical to the performance of the XGBoost model following LASSO-based selection are ranked by their F-scores in Fig. 5. The F-score quantifies the usage frequency of each feature in node splitting within the decision trees, thereby reflecting the relative contribution to stratification accuracy. Among these features, VE/VCO₂ and weight achieved the highest importance, followed sequentially by predicted maximal VO₂, maximal SBP, and HR_{rest}.

The corresponding SHAP beeswarm plot for the high-risk class (Fig. 6) was derived from the XGBoost model trained on the LASSO-selected features ($\alpha=0.001$). This analysis reveals that impaired ventilatory efficiency, reduced aerobic capacity, elevated HR_{rest}, advanced age, and specific comorbidities (hyperlipidemia and heart failure) exert the greatest positive influence on high-risk category assignment.

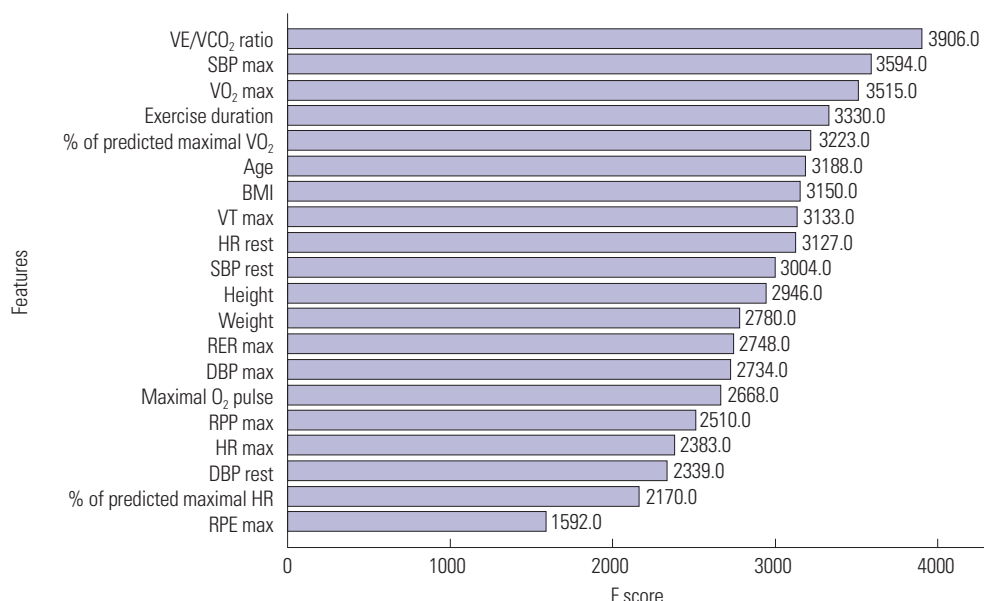


Fig. 3. Top 20 feature importance plot of the extreme gradient boosting model using all features. The VE/CO₂ ratio contributed most to risk prediction (F-score=3906), followed by maximal systolic blood pressure (3594), maximal oxygen consumption (3515), exercise duration (3330), and percentage of predicted maximal VO₂ (3223).

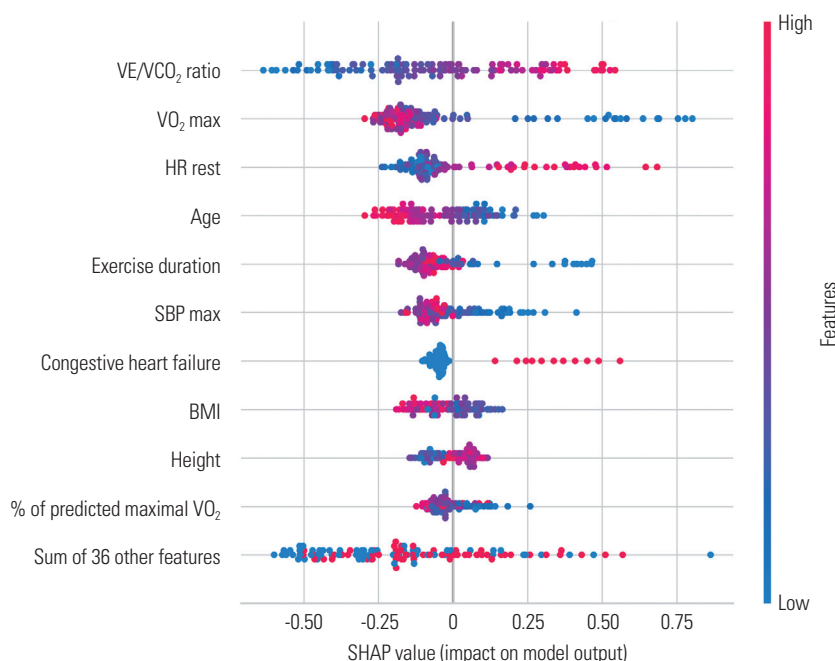


Fig. 4. SHAP beeswarm plot for the high-risk class, generated from an extreme gradient boosting model trained on all 46 features. The VE/CO₂ ratio, VO₂ max, resting heart rate, age, and exercise duration are the strongest predictors of high cardiovascular risk. Higher cardiopulmonary fitness (as indicated by VO₂ max) is protective, whereas elevated VE/CO₂ and resting heart rate reflect greater cardiovascular strain and increase risk.

Table 5. Performance Metrics of the Risk Stratification Model Using LASSO-Based Feature Selection

Alpha (α)	Accuracy	Precision	Recall	F1-score	AUC (mean)
0.1	66.38 (57.76–75.00)	69.87 (56.53–80.89)	54.46 (45.32–63.93)	57.02 (46.57–66.35)	78.66
0.01	74.14 (65.52–81.90)	76.25 (65.94–85.66)	66.15 (56.79–75.33)	69.11 (59.24–78.59)	87.40
0.001	79.31 (71.55–86.21)	82.07 (72.99–90.25)	72.46 (63.47–81.51)	75.49 (65.90–83.37)	89.40
0.0001	73.28 (65.52–81.03)	76.20 (65.52–86.42)	64.87 (56.53–73.84)	67.96 (57.96–76.36)	88.14

LASSO, least absolute shrinkage and selection operator; AUC, area under the curve.

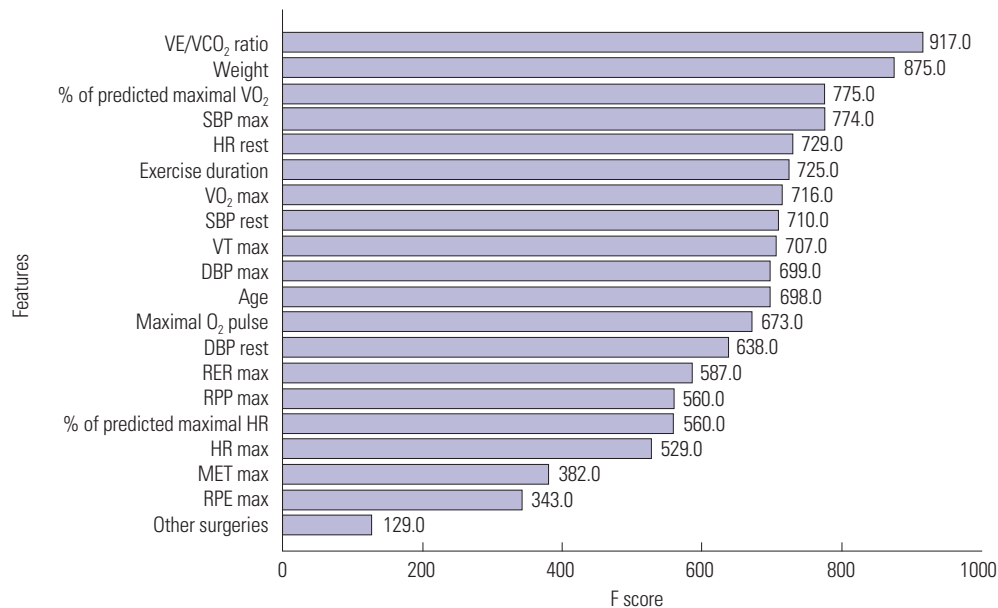


Fig. 5. Feature importance of the top 20 predictors selected by LASSO feature selection ($\alpha=0.001$) and evaluated using the extreme gradient boosting F-score. The VE/VO₂ ratio exhibits the highest importance (F-score=917), followed by weight (875), percentage of predicted maximal VO₂ (775), maximal systolic blood pressure (774), and resting heart rate (729).



Fig. 6. SHAP beeswarm plot for the high-risk class, generated from an XGBoost model trained on the features selected by least absolute shrinkage and selection operator ($\alpha=0.001$). The VE/VO₂ ratio, VO₂ max, resting heart rate, age, and exercise duration were identified as the strongest predictors of high cardiovascular risk. Higher VE/VO₂, elevated resting heart rate, and the presence of hyperlipidemia or congestive heart failure drive the model toward high-risk prediction. Conversely, higher VO₂ max and the percentage of predicted VO₂ levels, along with younger age, reduce the predicted risk.

DISCUSSION

CR is beneficial in reducing secondary cardiovascular events and cardiovascular death rates.⁴ In this study, we aimed to develop a machine learning-based model to classify risk levels for exercise-related cardiovascular events in patients with CVD, us-

ing clinician-assigned risk levels as the reference standard.

We developed an XGBoost-based machine learning model incorporating 46 variables to stratify the risk levels associated with exercise-related cardiovascular events in patients with CVD. The model was trained using risk categories assigned by clinicians, who determined the stratification by integrating

CPET results and clinical data in accordance with the AACVPR guidelines. The model demonstrated high accuracy in risk stratification, achieving an accuracy of 76.72% and an AUC of 88.76%, reflecting XGBoost's ability to capture complex non-linear relationships and interactions among clinical features. Among the most influential factors for risk stratification identified by the XGBoost model, ventilatory efficiency, aerobic capacity, HR_{rest}, and age were the most important. SHAP analysis further confirmed that elevated VE/VCO₂ and reduced VO₂ max were strongly associated with high-risk classification, underscoring the critical role of ventilatory and aerobic efficiencies in risk stratification.

To enhance interpretability and the accuracy of risk stratification, we applied LASSO-based feature selection, which improved performance (accuracy=79.31%, AUC=89.40%) by reducing redundant features and streamlining the model. Using a LASSO-based feature selection approach ($\alpha=0.001$), the XGBoost model identified VE/VCO₂, body weight, predicted maximal VO₂, maximal SBP, and HR_{rest} as the most critical factors driving the risk stratification, based on F-scores. SHAP beeswarm plot analysis for the high-risk class derived from this LASSO-refined model indicated that impaired ventilatory efficiency, reduced aerobic capacity, elevated HR_{rest}, older age, and comorbid conditions such as hyperlipidemia and heart failure were the strongest contributors to high-risk classification.

The ventilatory equivalent of carbon dioxide (VE/VCO₂ slope), which represents the amount of ventilation required to eliminate carbon dioxide during exercise, is a well-established parameter of ventilatory efficiency measured using CPET. It reflects not only pulmonary function but also cardiovascular pathophysiology such as impaired cardiac output, increased pulmonary vascular resistance, and elevated left ventricular filling pressures. Previous studies have consistently linked elevated VE/VCO₂ values to poorer cardiovascular outcomes and greater diastolic dysfunction severity.^{29,30} Similarly, VO₂ max is a robust and objective indicator of cardiac function, with modest improvements linked to clinically meaningful reductions in cardiovascular-related hospital readmissions and all-cause mortality in patients with coronary artery disease and heart failure.^{31,32} Our findings are consistent with this evidence and reinforce the clinical significance of CPET-derived parameters in determining risk categories.

A notable finding of this study is the model's ability to achieve a high level of concordance with clinician-assigned risk stratification despite being trained on a comparatively limited dataset. Whereas clinician-based assessment incorporates multidimensional clinical information—such as echocardiographic parameters (e.g., left ventricular EF), continuous electrocardiographic monitoring for exercise-induced arrhythmias and hemodynamic responses (blood pressure and HR), and post-treatment clinical stability (including recent coronary artery disease, percutaneous coronary intervention, or coronary artery bypass grafting)—in accordance with the AACVPR guidelines, the pro-

posed AI model relied primarily on baseline medical history and CPET-derived variables (Supplementary Table 1, only online).

The model's ability to closely approximate expert risk classification in the absence of direct echocardiographic data or real-time electrocardiographic waveforms suggests that physiological responses elicited during exercise testing may act as robust surrogates for more complex clinical and imaging-based parameters. These findings underscore the potential utility of the proposed AI-based approach as a streamlined and automated screening tool, particularly in clinical environments where immediate access to advanced cardiac imaging or specialized electrocardiographic interpretation is limited.

The strengths of this study include a large, multicenter, retrospective dataset from the Korean population, which provides a tailored risk stratification tool to guide CR prescriptions. The model maintained high accuracy even with limited non-exercise test data, thereby enhancing its clinical utility.

However, several limitations of this study should be acknowledged. First, a limitation of this study relates to the definition of the study population. As this work constitutes a secondary analysis of an open-access database, detailed information regarding the original inclusion and exclusion criteria and screening procedures of the primary cohort was not fully available. Although the dataset was generated as part of a standardized national project using a unified protocol to promote consistency across participating institutions, the lack of granular recruitment details limits precise characterization of the study population. This limitation arises from the methodological characteristics of secondary analyses using national or administrative datasets. Second, the retrospective design and the use of a dataset derived from a single cohort across three hospitals may limit the generalizability of our findings. While internal validation was conducted to assess model performance, the absence of external validation using independent datasets from diverse institutions or ethnic populations restricts the applicability of the model to the broader population of CR patients. Furthermore, as this study was conducted exclusively on Korean patients, our results may not fully reflect the clinical characteristics of other populations. Therefore, international validation involving different healthcare systems and diverse disease profiles is essential to confirm the global generalizability of our findings. Third, exercise intensity prescriptions in the dataset were based solely on HR. This approach may be less reliable for patients taking beta-blockers, which can blunt HR responses. Future studies should incorporate alternative measures, such as the Borg Rating of Perceived Exertion (RPE), to ensure more robust exercise prescriptions. Fourth, this study did not include outcome-based validation. The model was trained and evaluated against clinician-assigned risk classifications rather than longitudinal clinical outcomes, such as major adverse cardiovascular events, mortality, or exercise-related complications. Consequently, the findings primarily reflect the model's ability to reproduce expert clinical judgment rather than its direct

prognostic performance. Prospective validation against objective clinical outcomes is an essential next step to establish the clinical utility of this approach. Finally, the reference labels used for training were based on expert clinician judgment, which is inherently subject to inter-rater variability and subjectivity. Although risk stratification followed established guidelines, individual clinician interpretation may have influenced the labeling process. Future research employing more objective labeling strategies or consensus-based multi-rater assessments could help mitigate this limitation and further improve model reliability.

To the best of our knowledge, this is among the first studies to develop a machine-learning model specifically aimed at classifying CR risk levels, which may assist in standardizing risk stratification for benchmarking existing clinical classifications of exercise-related cardiovascular events and potentially help optimize exercise interventions pending clinical validation. While the criteria for this stratification can vary across clinical guidelines, our model successfully integrates complex factors—such as age, comorbidities, EF, and history of intervention—to replicate the expert clinical judgment required to identify these specific risk categories.

To ensure the generalizability of these findings, a concrete external validation strategy is warranted. Future studies should prioritize multicenter collaborations involving diverse clinical settings, from primary clinics to tertiary academic centers, and encompass patient populations with varied ethnic backgrounds and healthcare systems. Supporting generalizability would require demonstrating consistent discriminative performance, specifically maintaining high sensitivity and specificity, along with robust AUC-ROC and F1-scores, across different disease profiles and institutional datasets. Looking ahead, data accumulation, including prescriptions and actual patient outcomes, will enable precise analysis of exercise-related events and side effects, thereby improving the accuracy and safety of CR prescriptions using evidence-based AI models.

This approach has the potential to assist clinical decision-making by offering objective and standardized risk assessments. However, while the model reproduces clinician-assigned risk classifications, its ability to actually reduce adverse events or improve clinical outcomes has not been validated. Prospective studies incorporating clinical outcome measures are essential before routine clinical implementation.

ACKNOWLEDGEMENTS

This work was supported by the GRRC program of Gyeonggi province [GRRC-Gachon2023 (B01), Development of AI-based medical imaging technology], and by the Gachon University (GCU-202403860001).

We would like to thank Editage (www.editage.co.kr) for English language editing. This research used datasets from “The Open AI Dataset Project (AI-Hub, S. Korea).” All data information can be accessed through AI-Hub (www.aihub.or.kr).

AUTHOR CONTRIBUTIONS

Conceptualization: all authors. **Data curation:** Ha Eun Na, Gi Pyo Lee, Young Jae Kim, and Ju Kang Lee. **Formal analysis:** Gi Pyo Lee. **Funding acquisition:** Kwang Gi Kim and Ju Kang Lee. **Investigation:** Ha Eun Na and Gi Pyo Lee. **Methodology:** Ha Eun Na, Gi Pyo Lee, and Young Jae Kim. **Project administration:** Young Jae Kim, Kwang Gi Kim, and Ju Kang Lee. **Resources:** Kwang Gi Kim. **Software:** Gi Pyo Lee. **Supervision:** Young Jae Kim and Ju Kang Lee. **Validation:** Ha Eun Na and Gi Pyo Lee. **Visualization:** Ha Eun Na and Gi Pyo Lee. **Writing—original draft:** Ha Eun Na and Gi Pyo Lee. **Writing—review & editing:** all authors. **Approval of final manuscript:** : all authors.

ORCID iDs

Ha Eun Na	https://orcid.org/0009-0006-5957-7324
Gi Pyo Lee	https://orcid.org/0000-0002-9785-3776
Young Jae Kim	https://orcid.org/0000-0003-0443-0051
Kwang Gi Kim	https://orcid.org/0000-0001-9714-6038
Ju Kang Lee	https://orcid.org/0000-0002-8335-9785

REFERENCES

1. Ministry of Data and Statistics. [2022 cause of death statistics]. Daejeon: Ministry of Data and Statistics; 2023. Korean
2. Dagues N, Hindricks G. Sudden cardiac death in acute coronary syndromes. *Card Electrophysiol Clin* 2017;9:725-30.
3. Jenča D, Melenovský V, Stehlik J, Staněk V, Kettner J, Kautzner J, et al. Heart failure after myocardial infarction: incidence and predictors. *ESC Heart Fail* 2021;8:222-37.
4. Thomas RJ. Cardiac rehabilitation—challenges, advances, and the road ahead. *N Engl J Med* 2024;390:830-41.
5. Kim C, Sung J, Lee JH, Kim WS, Lee GJ, Jee S, et al. Clinical practice guideline for cardiac rehabilitation in Korea: recommendations for cardiac rehabilitation and secondary prevention after acute coronary syndrome. *Korean Circ J* 2019;49:1066-111.
6. Visseren FLJ, Mach F, Smulders YM, Carballo D, Koskinas KC, Bäck M, et al. 2021 ESC guidelines on cardiovascular disease prevention in clinical practice: developed by the task force for cardiovascular disease prevention in clinical practice with representatives of the European Society of Cardiology and 12 medical societies with the special contribution of the European Association of Preventive Cardiology (EAPC). *Eur Heart J* 2021;42:3227-337.
7. Goel K, Lennon RJ, Tilbury RT, Squires RW, Thomas RJ. Impact of cardiac rehabilitation on mortality and cardiovascular events after percutaneous coronary intervention in the community. *Circulation* 2011;123:2344-52.
8. Kim C, Lee SH, Kim WH. [Current status of cardiac rehabilitation in regional cardiocerebrovascular center]. *Public Health Wkly Rep* 2022;15:1363-76. Korean
9. Raja MA, Loughran R, Mc Caffery F. A review of applications of artificial intelligence in cardiorespiratory rehabilitation. *Inform Med Unlocked* 2023;41:101327.
10. De Cannière H, Corradi F, Smeets CJP, Schoutteten M, Varon C, Van Hoof C, et al. Wearable monitoring and interpretable machine learning can objectively track progression in patients during cardiac rehabilitation. *Sensors (Basel)* 2020;20:3601.
11. Rawstorn JC, Ball K, Oldenburg B, Chow CK, McNaughton SA, Lamb KE, et al. Smartphone cardiac rehabilitation, assisted self-management versus usual care: protocol for a multicenter randomized controlled trial to compare effects and costs among people

- with coronary heart disease. *JMIR Res Protoc* 2020;9:e15022.
12. Milani JGPO, Milani M, Verboven K, Cipriano G Jr, Hansen D. Exercise intensity prescription in cardiovascular rehabilitation: bridging the gap between best evidence and clinical practice. *Front Cardiovasc Med* 2024;11:1380639.
13. Silva AK, Barbosa MP, Bernardo AF, Vanderlei FM, Pacagnelli FL, Vanderlei LC. Cardiac risk stratification in cardiac rehabilitation programs: a review of protocols. *Rev Bras Cir Cardiovasc* 2014;29:255-65.
14. National Information Society Agency. Muscle exercise prescription data [Internet] [accessed on 2025 May 23]. Available at: <https://ai-hub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=&dataSetSn=71331>.
15. Frontera WR, DeLisa JA, Gans BM, Robinson LR, Bockenek W, Chae J. DeLisa's physical medicine and rehabilitation: principles and practice. 6th ed. Philadelphia: Wolters Kluwer; 2020.
16. American Association of Cardiovascular and Pulmonary Rehabilitation. AACVPR stratification algorithm for risk of event. Chicago: American Association of Cardiovascular and Pulmonary Rehabilitation; 2012.
17. van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. *J Stat Softw* 2011;45:1-67.
18. Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. 3rd ed. Hoboken: John Wiley & Sons; 2013.
19. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;20:273-97.
20. Chen T, Guestrin C. XGBoost: a scalable tree boosting system [accessed on 2025 May 23]. Available at: <https://doi.org/10.1145/2939672.2939785>.
21. Ryu SY, Lee SM, Kim YJ, Kim KG. Machine learning-based prediction of in-hospital mortality in patients with chronic respiratory disease exacerbations. *Digit Health* 2025;11:20552076251326703.
22. Breiman L. Random forests. *Mach Learn* 2001;45:5-32.
23. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York: Springer; 2009.
24. Bhat PC, Prosper HB, Sekmen S, Stewart C. Optimizing event selection with the random grid search. *Comput Phys Commun* 2018;228:245-57.
25. Jeon Y, Kim YJ, Jeon J, Nam KH, Hwang TS, Kim KG, et al. Machine learning based prediction of recurrence after curative resection for rectal cancer. *PLoS One* 2023;18:e0290141.
26. Browne MW. Cross-validation methods. *J Math Psychol* 2000;44:108-32.
27. Tibshirani R. Regression shrinkage and selection via the LASSO. *J R Stat Soc Ser B (Methodol)* 1996;58:267-88.
28. Lundberg SM, Lee SI. A unified approach to interpreting model predictions [accessed on 2025 May 23]. Available at: <https://dl.acm.org/doi/10.5555/3295222.3295230#purchase-access>.
29. Naylor M, Xanthakis V, Tanguay M, Blodgett JB, Shah RV, Schoenike M, et al. Clinical and hemodynamic associations and prognostic implications of ventilatory efficiency in patients with preserved left ventricular systolic function. *Circ Heart Fail* 2020;13:e006729.
30. Ojima S, Kubozono T, Kawasoe S, Kawabata T, Salim AA, Ikeda Y, et al. VE/VCO₂ slope in cardiopulmonary exercise testing was associated with left ventricular diastolic dysfunction in patients with reduced ejection fraction. *Eur Heart J* 2023;44(Supplement 2):ehad655.2585.
31. Mikkelsen N, Cadarso-Suárez C, Lado-Baleato O, Díaz-Louzao C, Gil CP, Reeh J, et al. Improvement in VO₂peak predicts readmissions for cardiovascular disease and mortality in patients undergoing cardiac rehabilitation. *Eur J Prev Cardiol* 2020;27:811-9.
32. Taylor JL, Bonikowske AR, Olson TP. Optimizing outcomes in cardiac rehabilitation: the importance of exercise intensity. *Front Cardiovasc Med* 2021;8:734278.