scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A multimodal dataset for coronary microvascular disease biomarker discovery

Dantong Li[1,2,3,10], Xiaoting Peng[1,2,3,10], Lianting Hu[1,2,3,10], Jintai Chen[4], Xinyang Long[1,2,3], Xueli Zhang[2], Siting Ye[5], Xiaohe Bai[6], Chao Wu[7], Huan Yang[1,2,3], Shuai Huang[1,2,3], Lingcong Kong[1,2,3], Entao Liu[8], Shuxia Wang[8] ✉, Huan Ma[2] ✉, Qingshan Geng[9] ✉ & Huiying Liang[1,2,3] ✉

Coronary microvascular disease (CMD), particularly prevalent among women, is associated with increased morbidity and mortality, making clinical screening vital for effective management. However, limited publicly available screening-level data hinders disease-specific biomarker discovery. To address this gap, 80 female angina patients without obstructive coronary artery disease and 40 age-matched female controls were prospectively enrolled to curate a new dataset. All participants underwent adenosine stress with electrocardiogram (ECG) monitoring across Rest, Stress, and Recovery stages. CMD diagnosis was confirmed with the standard clinical criterion, i.e., coronary flow reserve (CFR) < 2.0 via PET/CT. Using ECG variables from different stages, we developed machine learning models to classify CMD, thus validating dataset's effectiveness in CMD identification. We also validated the potential of ECG for differential diagnosis through joint analysis with the published mental stress-induced myocardial ischemia (MSIMI) dataset, which is based on the same cohort under different stress conditions. Disease-specific ECG variable sets were identified. Our findings highlight the value of multi-stage ECG in CMD screening. We expect this dataset to significantly advance CMD research.

Background & Summary

Nearly half of all patients presenting with angina in the catheterization laboratory are found to have nonobstructive coronary artery disease (ANOCA), a condition particularly prevalent in women, accounting for approximately 75% of cases[1–3]. After excluding noncardiac causes, there is compelling evidence that the chest pain in the majority is linked to abnormalities in the coronary circulation, with coronary microvascular disease (CMD) being one of the primary contributors[4,5].

Clinical screening for CMD is essential for stratifying the management of angina patients, given its association with increased morbidity and mortality[6]. The CORonary MICrovascular Angina (CorMicA) study showed that accurate CMD diagnosis and treatment can reduce angina severity by 27% at 12 months post-diagnosis[7]. Moreover, accurate CMD diagnosis can reduce healthcare resource utilization, resulting in cost savings ranging from $2,100 to $7,300, benefiting both hospitals and patients.

[1]Medical Big Data Center, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, 510080, China. [2]Guangdong Provincial Cardiovascular Institute, Guangzhou, Guangdong Province, 510080, China. [3]Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, Guangdong Province, 510080, China. [4]Information Hub, the Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China. [5]Department of Ultrasound, The Second Affiliated Hospital of Guangzhou University, Guangzhou, China. [6]School of Physical Sciences, University of California San Diego, La Jolla, San Diego, CA, 92093, USA. [7]Institute for Heart and Brain Health, University of Michigan Medical Center, Ann Arbor, Michigan, USA. [8]Department of Nuclear Medicine, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China. [9]Department of Cardiology, Shenzhen People's Hospital (The Second Clinical Medical College, Jinan University, Guangzhou, China. [10]These authors contributed equally: Dantong Li, Xiaoting Peng, Lianting Hu. ✉e-mail: wang_shuxia2002@aliyun.com; mahuandoctor@163.com; gengqsh@163.net; lianghuiying@hotmail.com

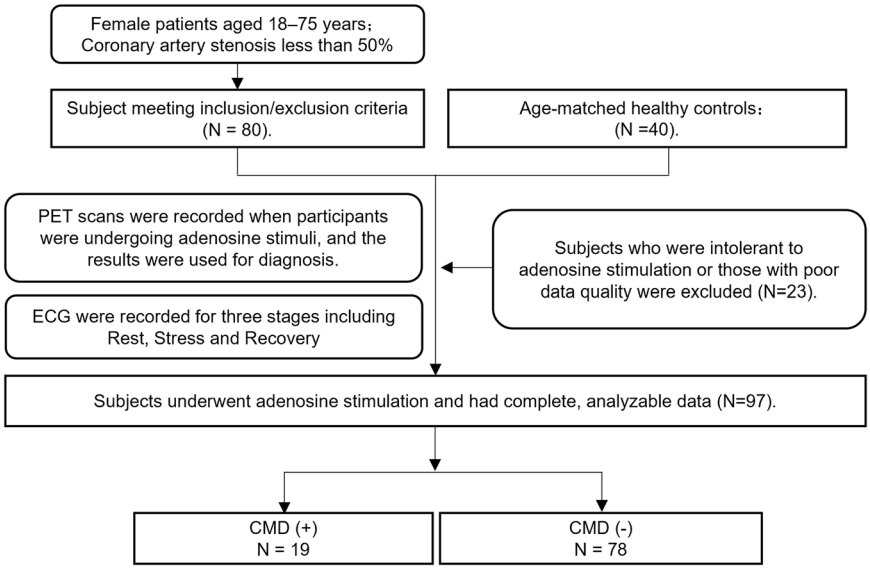

Fig. 1 Participant flow chart of the study.

Diagnosis of CMD can be made by coronary flow reserve (CFR)obtained by a combined pressure/thermodilution wire or noninvasively by positron emission tomography (PET). However, neither intracoronary testing nor PET are ideal for repeated investigations owing to limited access, invasive nature, cost, and radiation exposure[6]. Thus, the idea of simple, non-invasive biomarkers or screening tools for assessing CMD is particularly attractive.

Electrocardiogram (ECG) is a cost-effective tool, which utilization spans from primary healthcare institutions to residential applications, establishing it as the preferred method for the initial assessment of CMD in various medical and personal health monitoring environments[8–10]. However, there are two issues remain: (i) the conventional criteria (i.e., ST-segment depressions ≥ 1 mm) showed limited sensitivity for detecting CMD[11]; (ii) patients with similar clinical presentations may have different underlying conditions, such as CMD or mental stress-induced myocardial ischemia (MSIMI).Given that the intervention strategies vary depending on the etiology, it is crucial to obtain a preliminary understanding during the screening process.

To address the aforementioned issues, we prospectively enrolled 80 female patients with ANOCA and 40 age-matched healthy female controls (Fig. 1). Subjects were asked to undergo adenosine infusion at a constant rate of 140 μg/kg/min for 6 minutes using a syringe pump. ECG recordings were obtained for three distinct phases: rest (6 min), stress (6 min), and recovery (6 min) (Fig. 2). Myocardial perfusion was assessed during the stress phase using PET/CT imaging with a 13N-ammonia tracer, and CMD was diagnosed with a CFR < 2.0.23 subjects were excluded due to adenosine intolerance or poor data quality, leaving 97 subjects for analysis. Based on ECG data from three distinct stages, we calculated 88 interpretable ECG variables for each stage. These variables were then used to explore their predictive value for CMD by applying five classical machine learning models.

Furthermore, this same cohort of subjects also underwent mental stress stimulation, with ECG recordings captured for modified stages: rest (6 min), stress (12 min), and recovery (6 min) for each stimulus. MSIMI was diagnosed using a summed difference score (SDS) ≥ 3 obtained from PET/CT. Following rigorous quality control, complete data for both adenosine and mental stress stimulations were available for 93 subjects. We then assessed the correlation between ECG variables and the diagnostic criteria (CFR for CMD and SDS for MSIMI), resulting in the identification of two disease-specific sets of ECG variables[12,13]. We believe the joint analysis of these CMD and MSIMI datasets could further help researchers identify similarities and differences between the two conditions, aiding in differential diagnosis[12,13].

Notably, our data were derived from the same individuals all stimulations conducted under standardized laboratory conditions between 7:00–10:00 AM while participants were fasting, thereby minimizing potential confounding factors. We anticipate that this dataset will bridge research gaps related to CMD screening and serve as a foundational resource for future related investigations.

Methods

Ethical approval and consent. The study protocol and procedures, including data sharing, were approved by the Ethics Committee at Guangdong Provincial People's Hospital (Approval No. GDREC2019298H(R3)). Prior to enrolment, potential participants receive a verbal explanation of the study's objectives, methodology, and their rights as research subjects. They are provided ample time to deliberate on participation. Written informed consent is obtained from all participants before data collection, with explicit assurances that:

- Personal identifiers will be anonymized in shared datasets;
- Confidentiality of sensitive information is strictly maintained;
- Participation is voluntary and may be withdrawn at any stage.

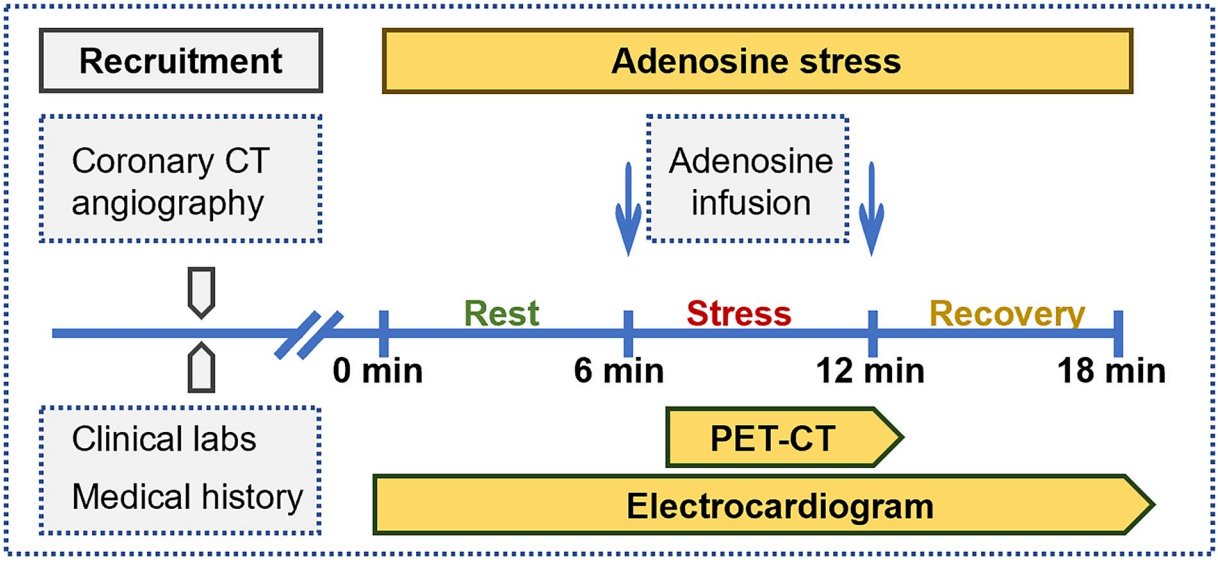

Fig. 2 The overview of data acquisition.

Participants. Women aged 18 to 75 years with chest pain or angina-like symptoms and confirmed coronary artery stenosis of less than 50% via coronary CT angiography were included under the ANOCA criteria. Exclusion criteria ruled out those with chest pain stemming from non-cardiac conditions or other severe diseases. Age-matched female control subjects were recruited through public notices and online platforms and underwent CT coronary angiography to exclude obstructive CAD. Further details can be found in our previously published dataset[12].

Procedures. All laboratory tests were performed between 7:00 and 10:00 AM with participants fasting, in order to minimize the impact of circadian rhythms and ensure accurate comparison of ECG signals. After a 6-minute supine rest, participants underwent an adenosine stress test within the diagnostic PET/CT unit. Two intravenous lines were used to administer the radiopharmaceutical while maintaining the adenosine infusion. Adenosine was infused at a constant rate of 140 μg/kg/min for 6 minutes using a syringe pump. During the first 2–3 minutes, a 'bolus-like' injection of 700–900 MBq of 13NH3 (5 mL) was administered. For patients at risk of complications, such as borderline hypotension, a reduced dose of 100–120 μg/kg/min was used. ECG and blood pressure were continuously monitored throughout the procedure.

Data acquisition. *ECG data.* ECG monitoring during the adenosine test was carried out continuously using a standard 12-lead system (Tim Software, Beijing Co., Beijing, China) with 16-bit precision and a 500 Hz sampling frequency.

PET/CT data. Following a CT scan for PET attenuation correction, PET data acquisition was conducted using 13NH3 as the tracer. All PET/CT examinations were performed on a single clinical scanner (Biograph HI-REZ 16, Siemens Medical Solution) in accordance with a standardized acquisition protocol and international PET/CT guidelines.

Data labelling and processing. *ECG data.* The ECG recordings were categorized into Rest (6 min), Stress (6 min), and Recovery (6 min) phases, corresponding to the start and end times of the adenosine infusion. These recordings were manually segmented and could be analysed individually or as a whole.

We also provided labels for pharmacologically induced myocardial ischemia (PSIMI) based on ECG data. The diagnosis of myocardial ischemia was determined through a consensus among three senior cardiologists, each with over a decade of clinical experience, using the standard criterion of ST depression greater than 0.1 mV. The final diagnosis was collectively agreed upon by the cardiologists. The MedEx MECG-200 ECG analysis system was employed to filter the signals and analyse the heart's electrical activity. Patient IDs, ages, and data acquisition dates were obtained from the hospital's health record database.

To eliminate noise from power line interference, baseline wander, and muscle contraction, we applied two median filters (200 ms and 600 ms) in combination with the Daubechies wavelet at a level 6 decomposition tree. After noise was removed, wave peaks were detected, and both morphological features and HRV indices were extracted using NeuroKit2, an open-source Python package suitable for both novice and advanced users. Comprehensive installation instructions for Python and NeuroKit2 can be found at https://github.com/neuropsychology/NeuroKit. For details on MSIMI data collection and labeling, please refer to our previously published work[12].

PET/CT data. PET data were collected in list mode and analysed by two experienced readers in accordance with the American Society of Nuclear Cardiology guidelines. Prior to analysis, the PET images were thoroughly

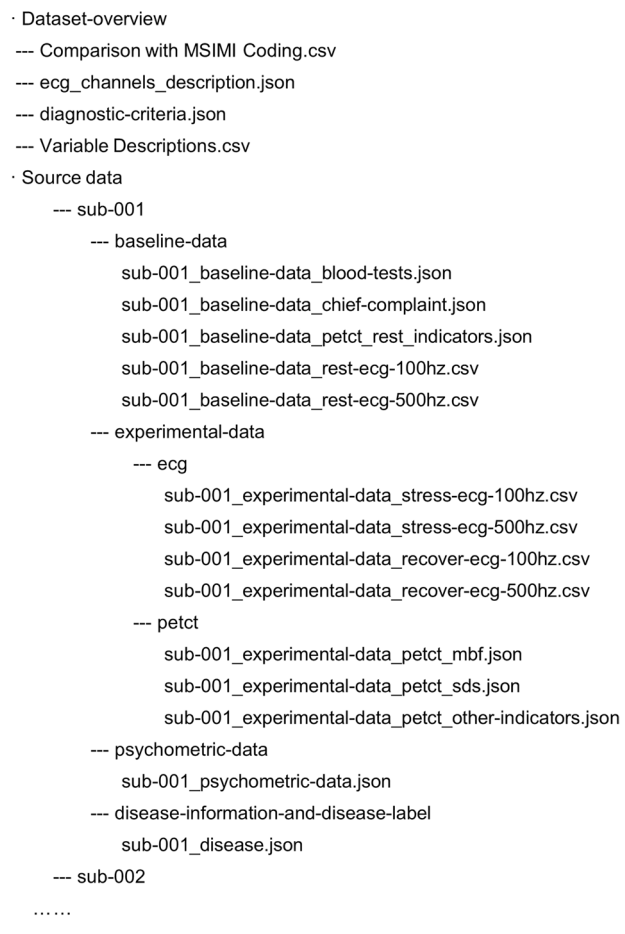

Fig. 3 Directory tree for the repository with previews.

inspected for any patient movement, attenuation issues, reconstruction artifacts, or low count density. Perfusion was assessed using QGS + QPS Automatic Quantification software, Version 2013.1, from Cedars-Sinai Medical Center, Los Angeles, USA.

Blood flow was measured in the left anterior descending artery (LAD), right coronary artery (RCA), left circumflex artery (LCX), as well as the total blood flow across all coronary arteries (FLOW-TOT) during both rest and stress conditions using PET/CT. Coronary flow reserve (CFR) was subsequently calculated as follows:

$$CFR = \frac{FLOW - TOT_{\text{Stress}}}{FLOW - TOT_{Rest}}$$

A CFR of less than 2.0 was established as the diagnostic threshold for CMD. Corresponding to the ECG data, we also provided PSIMI labels based on PET/CT findings. An SDS of 4 or higher is typically used as the criterion for diagnosing PSIMI. For details on MSIMI data collection and labeling, please refer to our previously published work[12].

Data Records

The dataset is available for download from the Science Data Bank, as referenced in citation number[14].

Dataset overview. The dataset's foundational information, such as disease definition, ECG channel description, diagnostic criteria, disease labels for CMD as well as MSIMI (see in our previously published dataset[12]), were organized into individual JSON-formatted files[15]. Figure 3 shows the directory tree for our repository and previews of the meta-data while Fig. 4 illustrate the shared and unique data within MSIMI and PSIMI dataset.

Dataset description. The ECG data for each participant during each stage was saved in a 'csv'-formatted file. The data for each subject (number: 001, 002, …) was stored as a first-level directory and was identical with the ones in previously published dataset for MSIMI[12]. The naming convention adopted for these files adhered to a specific pattern:

Sub-xxx_baseline-data(or experimental-data or null)datatype(petct_rest_indicators. *et al.*)where 'xxx' stands for the subject number (001, 002, …, 020).

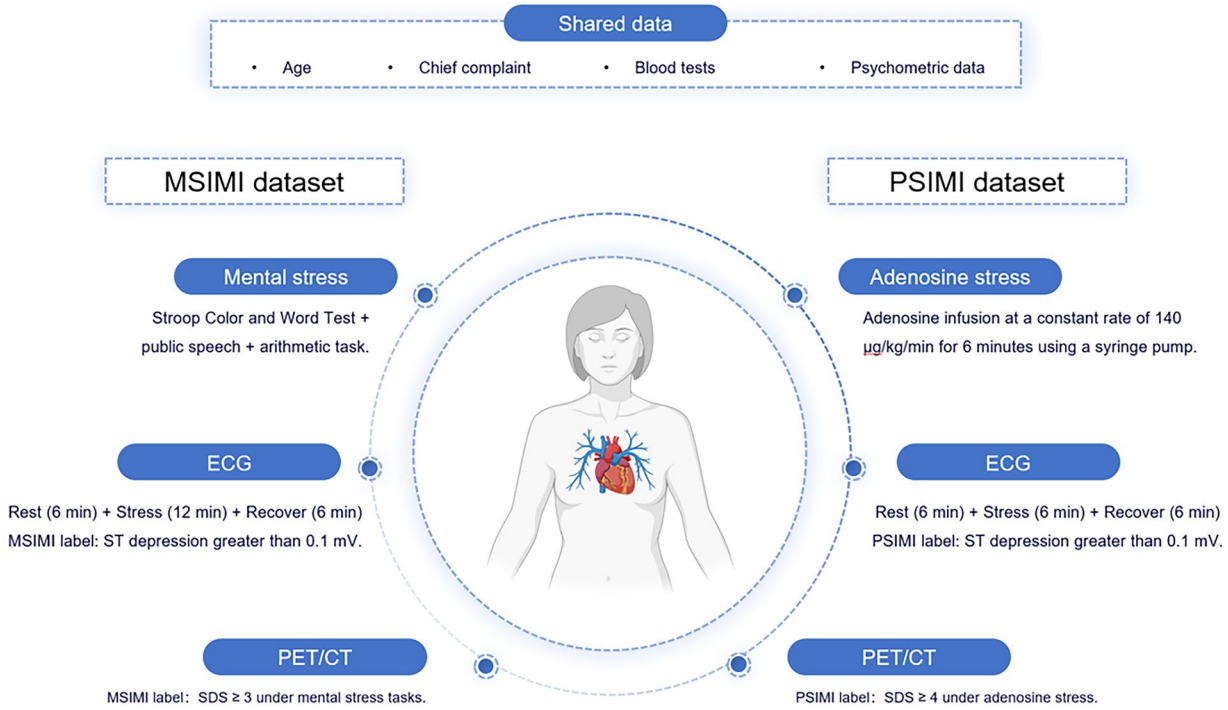

Fig. 4 Overview of MSIMI and PSIMI Datasets.

The labels, including CMD (-), and CMD(+), were provided in sub-xxx disease.json. For specific diagnostic criteria, please refer to diagnostic-criteria.json (as seen in the PET/CT data section).

Source data. We categorized the source data based on the recorded stages (before, during or after the adenosine stress) of the subject. One can locate the baseline data for all subjects in the current dataset as well as the MSIMI dataset. We pre-processed the raw ECG data, saving it as '.dat' files named after the subject number and stages. The PET/CT data consisted of structural information extracted using the software (QGS + QPS Automatic Quantification, Version 2013.1, Cedars-Sinai Medical Center, Los Angeles, USA).

Technical Validation

Disease classification. We conducted a detailed literature review to identify 88 ECG-based variables, including ECG morphology and heart rate variability (HRV) variables, which are listed in Table S1. The ECG recordings were extracted and segmented into three specific stages according to the timeline of the adenosine stress: Rest, Stress, and Recovery. For each stage, 88 ECG variables were calculated independently, resulting in a total of 264 variables. Additionally, 88 ECG variables were derived from the complete ECG data records. These variables from different stages, were used to develop five classic machine learning models—K-Neighbors, Logistic Regression, Random Forest, SVM, and XGBoost—to differentiate CMD (+) (Table 1). CFR < 2.0 obtained from PET/CT was used as diagnosis criteria. The average accuracy of the models was calculated across different sets using leave-one-out cross-validation. The models' accuracy varied based on the ECG variables from different stages, suggesting that ECG signals differ across stages and that ECG variables from recovery stage might potentially contribute to CMD diagnosis.

ECG variables comparison between CMD and MSIMI. A total of 264 ECG variables across three stages were calculated for both CMD and MSIMI. Statistical analyses were performed using t-tests to compare CMD (+) vs. CMD (−) and MSIMI (+) vs. MSIMI (−) (Table S2, with full details provided in the attachment). We further evaluated the correlations between ECG variables and diagnostic criteria separately (CFR for CMD and SDS for MSIMI). The results indicated that most ECG variables were correlated with both CFR and SDS (Table S3). However, HRV variables in the time domain (e.g., MeanNN, SDANN2, CVSD, SDSD, MedianNN) and entropy-based variables (e.g., SampEn, MSEn, CMSEn, RCMSEn) were specifically correlated with CFR, suggesting their potential as biomarkers for differential diagnosis during initial screening of these two conditions.

Usage Notes

This dataset has multiple potential uses for mental stress evaluation and daily ischemia detection. The hereby presented dataset and processing tools are provided for public use and may be used with proper citation to the current paper.

Stage	Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	AUC
Complete	Logistic Regression	80.52%	50.00%	6.67%	98.39%	11.76%	52.53%
Rest	Logistic Regression	81.82%	100.00%	6.67%	100.00%	12.50%	53.33%
Stress	Logistic Regression	84.42%	100.00%	20.00%	100.00%	33.33%	60.00%
Recovery	Logistic Regression	87.01%	85.71%	40.00%	98.39%	54.55%	69.19%
Rest + Stress + Recovery	Logistic Regression	87.01%	100.00%	33.33%	100.00%	50.00%	66.67%
Complete	Support Vector Machine	77.92%	43.75%	46.67%	85.48%	45.16%	66.08%
Rest	Support Vector Machine	76.62%	40.00%	40.00%	85.48%	40.00%	62.74%
Stress	Support Vector Machine	83.12%	57.14%	53.33%	90.32%	55.17%	71.83%
Recovery	Support Vector Machine	81.82%	53.33%	53.33%	88.71%	53.33%	71.02%
Rest + Stress + Recovery	Support Vector Machine	88.31%	80.00%	53.33%	96.77%	64.00%	75.05%
Complete	Random Forest	87.01%	85.71%	40.00%	98.39%	54.55%	69.19%
Rest	Random Forest	76.62%	36.36%	26.67%	88.71%	30.77%	57.69%
Stress	Random Forest	88.31%	80.00%	53.33%	96.77%	64.00%	75.05%
Recovery	Random Forest	84.42%	60.00%	60.00%	90.32%	60.00%	75.16%
Rest + Stress + Recovery	Random Forest	90.91%	100.00%	53.33%	100.00%	69.57%	76.67%
Complete	K-Nearest Neighbors	77.92%	41.67%	33.33%	88.71%	37.04%	61.02%
Rest	K-Nearest Neighbors	81.82%	66.67%	13.33%	98.39%	22.22%	55.86%
Stress	K-Nearest Neighbors	89.61%	100.00%	46.67%	100.00%	63.64%	73.33%
Recovery	K-Nearest Neighbors	94.81%	100.00%	73.33%	100.00%	84.62%	86.67%
Rest + Stress + Recovery	K-Nearest Neighbors	96.10%	100.00%	80.00%	100.00%	88.89%	90.00%
Complete	XGBoost	90.91%	100.00%	53.33%	100.00%	69.57%	76.67%
Rest	XGBoost	84.42%	66.67%	40.00%	95.16%	50.00%	67.58%
Stress	XGBoost	96.10%	100.00%	80.00%	100.00%	88.89%	90.00%
Recovery	XGBoost	92.21%	90.91%	66.67%	98.39%	76.92%	82.53%
Rest + Stress + Recovery	XGBoost	94.81%	100.00%	73.33%	100.00%	84.62%	86.67%

Table 1. Model performance based on ECG variables from different stages to classify CMD (+).

Code availability

For technical validation, we utilized publicly available code without any restrictions. Specifically, we employed the following functions/scripts:

- **nk.ecg_peaks.py** from the NeuroKit2 package to identify R-peaks in an ECG signal (https://github.com/neuropsychology/NeuroKit).
- **find_peaks.py** from the SciPy package to identify R-peaks in an ECG signal (https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.find_peaks.html).
- **nk.ecg_delineate.py** from the NeuroKit2 package to delineate the QRS complex for morphology features (https://neuropsychology.github.io/NeuroKit/_modules/neurokit2/ecg/ecg_delineate.html#ecg_delineate).
- **nk.hrv.py** from the NeuroKit2 package to compute Heart Rate Variability (https://neuropsychology.github.io/NeuroKit/_modules/neurokit2/hrv/hrv.html#hrv).
- **linear_model.LogisticRegression**, **svm.SVC**, **ensemble.RandomForestClassifier**, **neighbors.KNeighbors Classifier**, and **XGBClassifier** from the Scikit-learn package for classification (https://scikit-learn.org/stable/).

Received: 25 August 2024; Accepted: 16 April 2025;
Published online: 12 June 2025

References

1. Pepine, C. J. *et al.* Emergence of Nonobstructive Coronary Artery Disease: A Woman's Problem and Need for Change in Definition on Angiography. *J Am Coll Cardiol* **66**, 1918–1933, https://doi.org/10.1016/j.jacc.2015.08.876 (2015).
2. Murthy, V. L. *et al.* Effects of sex on coronary microvascular dysfunction and cardiac outcomes. *Circulation* **129**, 2518–2527, https://doi.org/10.1161/circulationaha.113.008507 (2014).
3. Boerhout, C. K. M., Beijk, M. A. M., Damman, P., Piek, J. J. & van de Hoef, T. P. Practical Approach for Angina and Non-Obstructive Coronary Arteries: A State-of-the-Art Review. *Korean Circ J* **53**, 519–534, https://doi.org/10.4070/kcj.2023.0109 (2023).
4. Voruganti, D. & Mehta, J. Coronary microvascular disease: diagnostic evaluation. *Vessel Plus* **5**, https://doi.org/10.20517/2574-1209.2021.60 (2021).
5. Belmonte, M. *et al.* Gender-related differences in absolute coronary flow and microvascular resistance in patients with angina and non-obstructed coronary arteries (ANOCA). *European Heart Journal* **44**, https://doi.org/10.1093/eurheartj/ehad655.2138 (2023).
6. Sinha, A., Rahman, H. & Perera, D. Coronary microvascular disease: current concepts of pathophysiology, diagnosis and management. *Cardiovasc Endocrinol Metab* **10**, 22–30, https://doi.org/10.1097/xce.0000000000000223 (2021).
7. Ford, T. J. *et al.* 1-Year Outcomes of Angina Management Guided by Invasive Coronary Function Testing (CorMicA). *JACC Cardiovasc Interv* **13**, 33–45, https://doi.org/10.1016/j.jcin.2019.11.001 (2020).
8. Sinha, A. *et al.* Rethinking False Positive Exercise Electrocardiographic Stress Tests by Assessing Coronary Microvascular Function. *Journal of the American College of Cardiology* **83**, 291–299, https://doi.org/10.1016/j.jacc.2023.10.034 (2024).
9. Lopez, D. M. *et al.* Role of Exercise Treadmill Testing in the Assessment of Coronary Microvascular Disease. *JACC Cardiovasc Imaging* **15**, 312–321, https://doi.org/10.1016/j.jcmg.2021.07.013 (2022).

10. Vandeloo, B. *et al*. Diagnostic performance of exercise stress tests for detection of epicardial and microvascular coronary artery disease: the UZ Clear study. *EuroIntervention* **18**, e1090–e1098, https://doi.org/10.4244/eij-d-22-00270 (2023).
11. Knuuti, J. *et al*. 2019 ESC Guidelines for the diagnosis and management of chronic coronary syndromes. *Eur Heart J* **41**, 407–477, https://doi.org/10.1093/eurheartj/ehz425 (2020).
12. Peng, X. *et al*. A multimodal physiological and psychological dataset for human with mental stress induced myocardial ischemia. *Sci Data* **11**, 704, https://doi.org/10.1038/s41597-024-03462-2 (2024).
13. Ma, H. *et al*. Assessing mental stress on myocardial perfusion and myocardial blood flow in women without obstructive coronary disease: protocol for a mechanistic clinical trial. *BMJ Open* **10**, e038362, https://doi.org/10.1136/bmjopen-2020-038362 (2020).
14. Peng, X. *et al*. A multimodal dataset to explore the potential biomarker for coronary microvascular disease. *ScienceDB* https://doi.org/10.57760/sciencedb.12045.
15. Chetran, A. *et al*. ECG and Biomarker Profile in Patients with Acute Heart Failure: A Pilot Study. Diagn. *Basel Switz*. **12**, 3037 (2022).

Acknowledgements

This study was funded by the Guangdong Provincial Medical Science and Technology Research Fund Project (A2023027 to D.L.), the Guangzhou Municipal 2025 Basic and Applied Basic Research Program (2025A04J4723 to D.L.), High-Level Talent Program (KY012023371 to D.L.), the Young Scientists Fund of National Natural Science Foundation of China (82200558 to D.L.), and the Tian Yuan Mathematical Foundation (12326612 to H.L.).

Author contributions

Conceptualization, formal analysis, writing, Dantong Li and Xiaoting Peng; resources, Huan Ma, Jun Quan, Shuxia Wang, Qingshan Geng; data acquisition, Shuai Huang, Xinyang Long, Xiaohe Bai, Xueli Zhang, Siting Ye, Chao Wu, Huan Yang, Lingcong Kong, Entao Liu; data curation, Lianting Hu, Jintai Chen; supervision, Huiying Liang.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-05022-8.

Correspondence and requests for materials should be addressed to S.W., H.M., Q.G. or H.L.