Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

CARDIO-ML-ED Dataset: A Multi-modal Clinical ECG Database for Machine Learning Applications and Data Analysis

Mohammad Hossein Moattar

mhmoattar@iau.ac.ir

Islamic Azad University

Shaghayegh Rahmani
MMS.C, Islamic Azad University

Franaz Soleimani
Mashhad Medical Sciences, Islamic Azad University

Moien Sohani Darban
Islamic Azad University

Saeedeh Baghyari
MMS.C. Islamic Azad University

Mohammad Mansouripour
MMS.C. Islamic Azad University

Seyed MohammadHosein Beheshti Shirazi
Islamic Azad University

Masoumeh Nejati Moharrami
Islamic Azad University

Parsa Khazraei
Islamic Azad University

Mohammad Reza Alipour
Islamic Azad University of Mashhad

KianaSadat Kimiaei
Islamic Azad University

Shirin Kazemi Esfahani
Islamic Azad University

Azin Homayounpour
Islamic Azad University

Abolfazl Pourtaheri
Mashhad Medical Sciences, Islamic Azad University

Hooman Movaghar Khoshbakht
Islamic Azad University

Amir Noori
Islamic Azad University of Mashhad

Ghazaleh Afkhami Teimouri
Islamic Azad University of Mashhad
Mobina Eslami
Islamic Azad University
Mobina Chavoshi
Islamic Azad University of Mashhad
Zeinab Sabour Karizno
Mashhad Medical Sciences, Islamic Azad University
Fatemeh DehghanNayeri
Mashhad Medical Sciences, Islamic Azad University
Farham Khaksari
Mashhad Medical Sciences, Islamic Azad University

Data Note

Keywords: Electrocardiography, Clinical Dataset, Multi-modal Data, Multi-label Data, Arrhythmia Detection, Cardiovascular Medicine, Clinical Decision Support

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11322784/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Objectives:

The CARDIO-ML-ED dataset was collected to address the critical shortage of publicly available cardiovascular databases that integrate electrocardiogram (ECG) imagery with comprehensive clinical metadata. Existing ECG repositories lack detailed patient-level information such as laboratory results, vital signs, and medical history—factors essential for developing clinically meaningful machine learning models. We aimed to create a curated, well-labelled, multimodal resource that supports research in automated ECG interpretation, multimodal fusion, and exploratory clinical phenotyping, while promoting reproducibility and innovation in artificial intelligence (AI) driven cardiovascular care.

Data description:

The CARDIO-ML-ED dataset comprises two complementary components: (1) structured clinical metadata for 582 subjects, and (2) a curated subset of 408 high-resolution ECG images derived from paper-based records digitized through standardized scanning, normalization, cropping, quality filtering, and noise reduction. The dataset includes 82 clinical features per subject—covering demographics, vital signs, laboratory tests (complete blood count, cardiac enzymes, electrolytes, inflammatory markers, arterial blood gases), and medical history. Diagnostic labels were independently assigned by a cardiologist and an emergency medicine physician, achieving substantial interrater agreement (Cohen's kappa = 0.82). This resource offers unique value through its dense integration of ECG data with clinical covariates, enabling multimodal fusion research, unsupervised clustering, imbalanced learning, and predictive analytics.

Objective

The CARDIO-ML-ED dataset was curated to address a critical gap in publicly available cardiovascular data resources: the lack of comprehensive, multimodal databases that pair high-quality electrocardiogram (ECG) imagery with extensive clinical metadata. Existing public ECG databases [1–10] typically suffer from limitations including small sample sizes, limited demographic diversity, inconsistent diagnostic annotations, and—most critically—an absence of detailed patient-level clinical information such as laboratory results, vital signs, and medical history.

Our primary objective was to create a resource that facilitates multimodal learning research, where the relationships between ECG morphological patterns and patient-specific clinical factors can be systematically explored. By integrating 82 structured clinical features with digitized ECG images, we aim to bridge the gap between computational data science and clinical cardiology, enabling the development of more interpretable and clinically aware artificial intelligence models.

This dataset was collected specifically for this data article and has not been previously published as part of another research study. The data were gathered prospectively over 12 months from two major medical centers in Mashhad, Iran, and are made publicly available to promote reproducibility and innovation in artificial intelligence driven cardiovascular diagnostics. To the best of our knowledge, no related publications have been produced using this specific dataset prior to this release.

Data description

The CARDIO-ML-ED dataset comprises three primary components, all linked by a unique file number for each subject and deposited in the Mendeley Data repository under the identifier https://doi.org/10.17632/zfszfmpb82.2. An overview is provided in Table 1.

Data file 1 – Clinical Metadata Repository

This Excel (.xlsx) file contains complete structured clinical data for all 582 subjects, organized across ten interconnected sheets. The master sheet includes core patient demographics (age, gender), admission metrics (vital signs including systolic/diastolic blood pressure, pulse rate, respiratory rate, temperature, and Peripheral Oxygen Saturation (SpO_2), along with triage level), and comprehensive laboratory results. Additional sheets detail diagnosis and treatment summaries, complete blood count with differential, basic metabolic panel and cardiac enzymes (troponin-I), inflammatory markers (erythrocyte sedimentation rate and C-reactive protein(CRP)), coagulation profile (Partial Thromboplastin Time (PTT), Prothrombin Time (PT), International Normalized Ratio (INR)), urinalysis, arterial blood gas analysis, and comprehensive medical history.

Data set 1 – ECG Image Repository: This collection comprises 408 high-quality, de-identified PNG images derived from paper-based 12-lead ECG traces. The original paper records were digitized using a Canon CanoScan LiDE 400 flatbed scanner at 600 Dots Per Inch (DPI) and saved as lossless Portable Network Graphics (PNG) images. A processing pipeline was applied to standardize the data: images were normalized to uniform dimensions (1500 × 1000 pixels) with contrast adjustment via histogram equalization; manually cropped to display at least four consecutive heartbeats focusing on leads I, II, III, aVR, aVL, aVF, and V1–V6; and enhanced using a median filter for noise reduction, Gaussian blur for smoothing, and unsharp masking for edge enhancement. Quality filtering excluded images with severe baseline wander, excessive muscle artifact, paper distortion, or incomplete lead representation. All personally identifiable information was removed using image inpainting techniques. Due to size constraints, the complete dataset is hosted externally:

https://drive.google.com/drive/folders/1u1jzVb8r2Rekel5BSrZ2moNMh-56DKiL?usp=drive_link

Data file 2 – ECG Diagnostic Labels

This Comma-Separated Values (CSV) file contains the diagnostic labels assigned to each ECG image by two independent medical experts (a board-certified cardiologist and an emergency medicine physician) using standardized American College of Cardiology/American Heart Association criteria.

Data file 3 – README

A text file providing a short overview of the dataset structure and usage notes.

Table 1
Overview of data files/data sets.

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data file 1	Datasheet1 (Edited)	Microsoft Excel file (.xlsx)	Mendeley Data (https://doi.org/10.17632/zfszfmpb82.2)
Data set 1	ECG Image Repository	PNG image files (.png)	https://drive.google.com/drive/folders/1u1jzVb8r2Rekel5BSrZ2moNMh-56DKiL?usp=drive_link
Data file 2	ECG Labels	CSV file (.csv)	Mendeley Data (https://doi.org/10.17632/zfszfmpb82.2)
Data file 3	README	Text file (.txt)	Mendeley Data (https://doi.org/10.17632/zfszfmpb82.2)

Limitations

Several limitations should be carefully considered by users. First, the dataset exhibits severe class imbalance, with the non-cardiac class dominating (≈ 70%). Clinically important categories such as atrial fibrillation (count = 3), Atrioventricular (AV) block (count = 4), and Premature Ventricular Contraction (PVC) (count = 4) have very small sample sizes, limiting the dataset's utility for training robust supervised classifiers for rare conditions. Users are advised to implement appropriate imbalance handling techniques and use evaluation metrics suitable for imbalanced data. Second, this database contains digital images of ECGs instead of raw waveforms. This format limitation restricts the application of standard signal-processing methods (e.g., R-peak detection, QT interval measurement) and the reproducibility of bio-signal algorithms. The dataset is best suited for computer vision and multimodal fusion approaches. Third, the usable ECG subset (n = 408) is relatively small for deep learning applications, particularly for multi-class classification. Users should consider transfer learning, data augmentation, or simpler models. Fourth, while data were collected from two hospitals in Mashhad, Iran, providing some demographic diversity (age 6–98 years), the dataset may not fully represent global patient populations, requiring external validation for generalization. Lastly, researchers using the full 582-subject metadata should be aware that ECG images are only available for 408 subjects, limiting multimodal analyses to this subset.

Abbreviations

Abbreviation
Full Form
AI
Artificial Intelligence
AV
Atrioventricular
CRP
C-Reactive Protein
CSV

Comma-Separated Values
DPI
Dots Per Inch
ECG
Electrocardiogram / Electrocardiography
INR
International Normalized Ratio
ML
Machine Learning
PNG
Portable Network Graphics
PT
Prothrombin Time
PTT
Partial Thromboplastin Time
PVC
Premature Ventricular Contraction
SpO_2
Peripheral Oxygen Saturation

Declarations

Ethics approval and consent to participate

This study was conducted in full compliance with the ethical standards of the institutional and national research committee and with the 1964 Helsinki Declaration and its later amendments. The study protocol was approved by the Islamic Azad University Ethics Committee (Approval Number: IR.IAU.MSHD.REC.1403.118). The requirement for individual informed consent was waived by the ethics committee for this retrospective, de-identified study in accordance with national regulations. All data were fully anonymized prior to analysis to protect patient privacy and confidentiality.

Consent for publication

Not applicable. This manuscript does not contain any individual person's data in any form (including individual details, images, or videos). The dataset has been fully de-identified, and no identifiable patient information is presented.

Competing interests

The authors declare that they have no competing interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

The authors declare that no funding was received for this research.

Author Contribution

M.H.M. (Corresponding Author): Supervision, Methodology, Project Administration, Data Curation, Writing – Original Draft, Writing – Review & Editing.S.R.: Resources, Supervision, Data Curation, Funding Acquisition, Writing – Review & Editing.F.S.: Supervision, Investigation, Data Curation, Writing – Review & Editing.M.S.D.: Software, Data Curation, Formal Analysis, Validation.S.B. and M.M.: Data Curation, ValidationS.M.H.B.S.: Software, Data Curation, Formal Analysis, Validation.M.N.M.: Software, Data Curation, Validation.P.K.: Software, Data Curation, Validation.M.R.A.: Resources, Investigation.K.K. and S.K.E. and A.H. and A.P. and H.M.K. and A.N. and G.A.T. and M.E. and M.C. and Z.S.K. and F.D. and F.K.: Data Curation, Validation.

Acknowledgement

The authors extend their sincere gratitude to the medical and administrative staff of 22 Bahman Hospital and Ghaem Hospital in Mashhad, Iran, for their invaluable cooperation and support in providing access to the clinical records necessary for this study. We are profoundly thankful to the board-certified cardiologists and emergency medicine physicians who dedicated their time and expertise to the meticulous annotation and labeling of the ECG signals. Their clinical insight was fundamental to ensuring the accuracy and reliability of CARDIO-ML-ED dataset. This project would not have been possible without the tireless efforts of our data curation team.

Data Availability

The data described in this Data note can be freely and openly accessed on Mendeley Data under the identifier https://doi.org/10.17632/zfszfmpb82.2. Please see Table 1 for details and links to the specific data files and data sets. The dataset is openly available for reuse under the terms specified by the repository.

References

1. Moody, G. B., & Mark, R. G. (2001). The impact of the MIT-BIH Arrhythmia Database. *IEEE Eng in Med and Biol*, *20*(3), 45–50. PMID: 11446209.
2. Bousseljot, R., Kreiseler, D., & Schnabel, A. (1995). Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet, Biomedizinische Technik, Band 40, Ergänzungsband 1 S 317.
3. Wagner, P., Strodthoff, N., Bousseljot, R., Samek, W., & Schaeffter, T. (2022). PTB-XL, a large publicly available electrocardiography dataset (version 1.0.3). PhysioNet. RRID:SCR_007345. https://doi.org/10.13026/kfzx-aw45
4. Clifford, G. D., Liu, C., Moody, B. (2017). AF classification from a short single lead ECG recording: The PhysioNet/computing in cardiology challenge 2017, in: 2017 Computing in Cardiology (CinC), IEEE, pp. 1–4. https://doi.org/10.22489/CinC.2017.065-469
5. Perez Alday, E. A., Gu, A., Shah, A. J., et al. (2020). Classification of 12-lead ECGs: the Physionet/computing in Cardiology challenge 2020. *Physiological Measurement*, *41*(12), 124003. 10.1088/1361-6579/abc960

6. O'Sullivan, J., Odinaka, I., Lai, P. H. (2010). ECG Biometrics: A Robust Short-time Frequency Analysis, in: 2010 IEEE International Workshop on Information Forensics and Security.
7. Perez Alday, E. A., Gu, A., Shah, A. J. (2020). Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020, Physiol Meas. Nov 11. http://doi.org/10.1088/1361-6579/abc960
8. Perez Alday, E. A., Gu, A., Shah, A. (2022). Classification of 12-lead ECGs: The PhysioNet/Computing in Cardiology Challenge 2020 (version 1.0.2). PhysioNet. RRID:SCR_007345. https://doi.org/10.13026/dvyd-kd57
9. Goldberger, A., Amaral, L., Glass, L., et al. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation [Online]*, *101*(23), e215–e220. RRID:SCR_007345.
10. Wagner, P., Strodthoff, N., Bousseljot, R. D. (2020). PTB-XL: A Large Publicly Available ECG Dataset, Scientific Data (https://doi.org/10.1038/s41597-020-0495-6