Digitization and Classification of ECG Images: The George B. Moody PhysioNet Challenge 2024

Matthew A Reyna[1], Deepanshi[1], James Weigle[1], Zuzana Koscova[1], Kiersten Campbell[1], Salman Seyedi[1], Andoni Elola[2], Ali Bahrami Rad[1], Amit J Shah[3], Neal K Bhatia[4], Gari D Clifford[1,5], Reza Sameni[1,5]

[1]Department of Biomedical Informatics, Emory University, USA
[2]Department of Electronic Technology, University of the Basque Country UPV/EHU, Spain
[3]Department of Epidemiology, Emory University, USA
[4]Department of Medicine, Emory University, USA
[5]Department of Biomedical Engineering, Georgia Institute of Technology, USA

Abstract

The George B. Moody PhysioNet Challenge 2024 invited teams to develop algorithmic approaches for digitizing and classifying electrocardiograms (ECGs) from photographed or scanned images of paper ECGs.

Paper ECGs have existed for decades, capturing the variability and evolution of cardiovascular diseases (CVDs) across demographics, geography, and time. Physical and digital ECG images remain common in cardiac care. However, ECG-based interpretation algorithms typically require digital time-series representations of ECG data, so existing algorithms cannot interpret them, and new algorithms cannot learn from them. Therefore, digitizing ECG images is important for improving the accessibility and quality of cardiac care.

To support this goal, the Challenge introduced ECG-Image-Kit, a synthetic ECG image generator with various realistic distortions, such as wrinkles, creases, shadows, rotations, and handwriting, to allow teams to create arbitrary large and diverse datasets for training generalizable models. The Challenge also introduced ECG-Image-Database, a dataset of 35,595 real ECG papers from 1,977 distinct ECG records, to assess and support the generalizabilty of the Challenge approaches. A total of 62 teams participated in the Challenge, representing diverse approaches from both academia and industry worldwide.

1. Introduction

The electrocardiogram (ECG) is an accessible, non-invasive pre-screening tool for cardiovascular diseases (CVDs). Invented in 1895, the ECG has evolved significantly, progressing to portable devices in 1927 and paper-printing ECGs by 1948 [1]. By 1988, algorithms interpreted over half of the 100 million ECGs recorded annually in the U.S. [2]. Modern advances include digital ECG devices and more sophisticated algorithmic interpretation algorithms, both which have increased accessibility to CVD-based diagnosis.

Despite the rise of digital ECGs, paper ECGs remain prevalent, especially in the Global South [3]. These ECGs reflect the diversity and evolution of CVDs across demographics, geography, and time. However, ECG interpretation algorithms generally expect ECG time-series instead of images, limiting the utility of the paper ECGs. Moreover, photographs and scans of paper ECGs often have distortions and other artifacts, such as creases, tears, fading ink, and stains on the paper as well as shadows, skewing, and blurriness from image acquisition.

Therefore, digitizing ECGs to extract the plotted ECG time-series data is vital for aiding ECG-based diagnosis and improving global cardiac care access. The 2024 Challenge invited teams to digitize and classify photographed and scanned images of paper ECGs.

2. Methods

Algorithms for digitizing and classifying ECG images typically apply classical image processing and, more recently, deep learning techniques. Some approaches attempt to digitize the images and use the extracted time-series to classify the image, and other approaches attempt to classify the images directly without using the underlying time-series.

Classical image processing techniques include grayscale thresholding for grid removal, pixel scanning for ECG digitization, and template-based optical character recognition (OCR) for patient data extraction [4]; heuristics derived from pixel intensities for region-of-interest identification

[5]; the Hough transform for skew correction, color-based segmentation for grid removal, and median filtering for noise removal [6]. Deep learning techniques include a dense neural network for grid removal [7, 8] and U-Net architecture for segmentation [9].

Deep learning methods have the potential to be more robust than classical imaging processing approaches to paper distortions and image noise and artifacts. However, these methods are generally limited by a lack of diverse noise artifacts and a scarcity of ground truth ECG time-series data in available datasets.

2.1. Challenge Data

The Challenge data included data from multiple sources, including public and private databases of ECG waveforms, ECG images, and ECG-based diagnoses or labels [10].

We generated ECG images from real ECG time-series using ECG-Image-Kit [8, 11]. This package allows the generation of ECG images with and without synthetic artifacts that resemble the real-world image artifacts, such as as wrinkles, creases, shadows, rotations, and handwriting. Teams could include this code or other code in their entries to augment the provided training set to improve the performance of their model.

The public data contained 21,799 12-lead ECG waveforms, labels, and images from the PTB-XL dataset [12, 13]. The PTB-XL dataset was available prior to the Challenge, and ECG-Image-Kit generated ECG images from the PTB-XL data for training; teams could generate additional images from these data.

The hidden data contained 1,977 12-lead ECG waveforms and labels and 35,595 ECG images from the PTB-XL dataset and an Emory University Hospital dataset. For these data, we printed, photographed, and/or scanned ECGs with various real-world artifacts, including (1) color scans, black-and-white scans, and mobile phone photographs of clean ECG papers; (2) mobile phone photos of stained papers; color scans, black-and-white scans, and mobile phone photos of deteriorated ECG papers; and mobile phone photos of a computer monitor [8, 10, 11].

In both the public and hidden data, the ECG waveforms were standard 12-lead ECGs that were 10 seconds long with sampling frequencies of either 250 Hz or 500 Hz. We encoded the ECG waveforms in a WFDB-compatible format using 16 bits of signal quantitization.

Each ECG record contained labels from the following classes: (1) acute myocardial infarction, (2) atrial fibrillation or atrial flutter, (3) bradycardia, (4) conduction disturbances, (5) hypertrophy, (6) normal, (7) old myocardial infarction, (8) premature atrial complex, (9) premature ventricular complex, (10) ST/T changes, and (11) tachycardia.

The labels for the PTB-XL dataset were taken directly from the data, and 12SL statement codes from the PTB-XL+ dataset defined separate acute MI and old MI classes [12, 13]. The labels for records from the Emory University Hospital dataset were derived from 12SL statement codes for the data and matched to the above classes.

All of the hidden data was sequestered during the Challenge to prevent overfitting on the data, but we plan to release it so that the community can use it to develop more robust and generalizable approaches. The leaderboard subset of the hidden data are the color and black-and-white scans from the PTB-XL dataset, and the extended hidden data are the other variants from the PTB-XL dataset.

2.2. Challenge Objective

For the 2024 Challenge, we asked the teams to design and implement open-source algorithms that could digitize the ECG, i.e., turn images of an ECG into ECG time-series data representing the same ECG and/or classify paper ECGs from the extracted time-series data or from the image itself. Teams could complete either or both tasks. The winners of the Challenge achieved the highest performance on the hidden leaderboard data.

2.2.1. Challenge Timeline

This year's Challenge was the 25th George B. Moody PhysioNet Challenge [14]. As in previous years, the Challenge had an unofficial phase and an official phase. The unofficial phase (25 January 2024 to 10 April 2024) introduced the teams to the Challenge. We publicly shared the Challenge objective, training data, example algorithms, and evaluation metric and invited the teams to submit their code for evaluation, scoring at most five entries from each team on the hidden data. Between the unofficial and official phases, we took a hiatus (11 April 2024 to 23 May 2024) to improve the Challenge. The official phase (24 May 2024 to 19 August 2024) continued the Challenge. We updated the Challenge data, example algorithms, and evaluation metric and again invited teams to submit their code for evaluation, scoring at most ten entries from each team on the hidden data.

We announced the results at the end of the Computing in Cardiology (CinC) 2024 conference, where the teams presented, defended, and published their work. Only teams that presented and published their work at the conference were eligible for rankings and prizes. We will publicly release the algorithms after the end of the Challenge and the publication of these papers.

The Challenge Organizers also held hackathons at the MidSouth Computational Biology and Bioinformatics Society 2024 conference in Atlanta, GA, USA on 24 March 2024; Data Science Africa 2024 summer school and workshop in Nyeri, Kenya from 3 June 2024 to 6 June 2024; and at CinC 2024 on 8 September 2024.

2.2.2. Challenge Evaluation

The evaluation metric for the ECG digitization task was the signal-to-noise ratio (SNR) of the digitized signal. Let $x = (x_i)_{i=1}^n$ be a signal in an ECG image, and let $y = (y_i)_{i=1}^n$ be a signal digitized from an ECG image. Since small horizontal and vertical translations are common but typically do not affect the interpretation of an ECG, we shifted y horizontally and vertically to maximize its cross-correlation with x for shifts smaller than ± 0.5 s and/or ± 1 mV. We then computed

$$\text{SNR} = 10 \log_{10} \frac{\sum_{i=1}^n (y_i - x_i)^2}{\sum_{i=1}^n x_i^2} \quad (1)$$

for each channel in each record. We did not score missing values in the digitized signal, and we scaled the SNR linearly by the fraction of samples in x that were not digitized in y. Higher SNR values are better, indicating that the model outputs better captured the ECG time-series with less noise. We computed the mean of the SNR values across all records and all channels in each record. The team with the highest mean SNR on the leaderboard data won the digitization task.

The evaluation metric for the ECG classification task was the macro F-measure. For each class in 2.1, we computed the per-class F-measure by comparing the ground truth and classifier labels in a one-vs.-rest manner for all records in a database. Higher F-measure values are better, indicating that the model better classified the ECG image. We computed the macro F-measure as the mean of the per-class F-measures across all classes. The team with the highest macro F-measure on the leaderboard data won the classification task.

3. Challenge Results

A total of 62 teams submitted 568 algorithms during the Challenge, including 53 teams with 74 successful entries and 133 unsuccessful entries during the unofficial phase and 43 teams with 98 successful entries and 263 unsuccessful entries during the official phase. After the end of the official phase, we attempted to score one entry from each team on the extended hidden data in Section 2.1. For the digitization task, we were able to score 22 teams on the leaderboard data and 15 teams on the entire hidden data; a total of 11 teams met all of the requirements to be ranked. For the digitization task, we were able to score 23 teams on the leaderboard data and 17 teams on the entire hidden data; a total of 12 teams met all of the requirements to be ranked. Tables 1 and 2 summarize the highest-ranked teams for the digitization and classification tasks, respectively. Team summaries, additional scores, and the full Challenge criteria for rankings are available in [15].

Rank	Team name	Leaderboard score	Mean extended score
1	SignalSavants	12.15	0.57
2	BAPORlab	5.49	1.47
3	wavie_ABI	5.47	N/A

Table 1: The three teams with the highest signal-to-noise ratio (SNR) score on the hidden leaderboard data and the extended hidden data; only ranked teams are shown.

Rank	Team name	Leaderboard score	Mean extended score
1	AIMED	0.82	0.73
2	Intentec AIC	0.74	0.70
3	BAPORlab	0.73	0.66

Table 2: The three teams with the highest macro F-measure score on the hidden leaderboard data and the extended hidden data; only ranked teams are shown.

4. Discussion

Teams with higher scores in the digitization task generally had higher scores in the classification task, but several teams with negative SNR scores, i.e, more noise than signal, could achieve F-measure classification scores that were close to, but lower than, the highest-performing teams.

This observation may suggest that intermediate time-series representations may not be necessary for ECG image interpretation, but it is more likely an indictment of our signal fidelity measure. The SNR is a standard metric, but it is sensitive to perturbations and errors in high-amplitude regions of the signal that are generally not critical for ECG interpretation; we modified the standard SNR calculation so that it was less sensitive to these regions. However, despite these changes, the SNR still does not directly capture or reflect clinical measurements that are likely to influence the downstream interpretation of an ECG; such a metric would typically require these annotations be sensitive to the choice of algorithm for providing them from the data.

There are several existing approaches for the digitization of ECG images. Unfortunately, we were unable to evaluate any of them, including ones for which code was available. Some solved a different problem, i.e., they did not produce ECG time-series, or they would not share their code or allow the independent evaluation of their algorithm. PowerfulMedical's PMcardio was an exception [16, 17]. PowerfulMedical provided us with API access to PMcardio for independent evaluation. We found that PMcardio outperformed the Challenge algorithms but, like the Challenge teams, they demonstrated variable performance across the different variants of the hidden data. A deeper analysis of this will be provided in a follow-up journal publication.

5. Conclusions

This year's Challenge explored the potential for the algorithmic digitization and classification of paper ECGs. We asked the Challenge participants to design working, open-source algorithms for extracting ECG time-series representations from ECG images and for classifying the ECGs from the extracted time-series and/or the image itself. Such algorithms have the potential to improve the interpretation of a ECGs and improve access to cardiovascular care. In addition, approaches such as those described in this Challenge may facilitate historical epidemiological studies of archival data and enable the preservation of analog patient histories.

Acknowledgements

This research is supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB: R01EB030362); the National Center for Advancing Translational Sciences of the National Institutes of Health (NCATS: UL1TR002378); as well as AliveCor, Amazon Web Services, IEEE SPS, and MathWorks under unrestricted gifts. AE received support by the MCIN/AEI/10.13039/501100011033/, by FEDER Una manera de hacer Europa through grant PID2021-122727OB-I00, and by the Basque Government through grant IT1717-22. GDC has financial interests in AliveCor, Nextsense and Mindchild Medical and holds a board position with Mindchild Medical. None of these entities influenced the design of or provided data for this year's Challenge. The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the above entities.

References

[1] Yang XL, Liu GZ, Tong YH, Yan H, Xu Z, Chen Q, et al. The history, hotspots, and trends of electrocardiogram. Journal of Geriatric Cardiology 09 2015;12:448–56.

[2] Drazen E, Mann N, Borun R, Laks M, Bersen A. Survey of computer-assisted electrocardiography in the united states. Journal of Electrocardiology 1988;21:S98–S104.

[3] Tison GH, Zhang J, Delling FN, Deo RC. Automated and interpretable patient ECG profiles for disease detection, tracking, and discovery. Circulation Cardiovascular Quality and Outcomes 2019;12:e005289.

[4] Ravichandran L, Harless C, Shah AJ, Wick CA, Mcclellan JH, Tridandapani S. Novel Tool for Complete Digitization of Paper Electrocardiography Data. IEEE Journal of Translational Engineering in Health and Medicine 2013; 1:1800107 –1800107.

[5] Baydoun M, Safatly L, Abou Hassan OK, Ghaziri H, El Hajj A, Isma'eel H. High precision digitization of paper-based ECG records: a step toward machine learning. IEEE Journal of Translational Engineering in Health and Medicine 2019;7:1–8.

[6] Garg DK, Thakur D, Sharma S, Bhardwaj S. ECG paper records digitization through image processing techniques. International Journal of Computer Applications 2012;48(13):35–38.

[7] Mishra S, Khatwani G, Patil R, Sapariya D, Shah V, Parmar D, et al. ECG paper record digitization and diagnosis using deep learning. Journal of Medical and Biological Engineering 2021;41(4):422–432.

[8] Shivashankara KK, Deepanshi, Shervedani AM, Reyna MA, Clifford GD, Sameni R. ECG-Image-Kit: a synthetic image generation toolbox to facilitate deep learning-based electrocardiogram digitization. Physiological Measurement May 2024;45(5):055019.

[9] Li Y, Qu Q, Wang M, Yu L, Wang J, Shen L, et al. Deep learning for digitizing highly noisy paper-based ECG records. Computers in Biology and Medicine 2020; 127:104077.

[10] Reyna MA, Deepanshi, Weigle J, Koscova Z, Campbell K, Shivashankara KK, et al. ECG-Image-Database: A dataset of ECG images with real-world imaging and scanning artifacts; a foundation for computerized ECG image digitization and analysis, 2024. URL https://arxiv.org/abs/2409.16612.

[11] Deepanshi, Shivashankara KK, Clifford GD, Reyna MA, Sameni R. ECG-Image-Kit: A Toolkit for Synthesis, Analysis, and Digitization of Electrocardiogram Images, January 2024. Online at: https://github.com/alphanumericslab/ecg-image-kit.

[12] Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, et al. PTB-XL, a large publicly available electrocardiography dataset. Scientific Data 2020;7:154.

[13] Strodthoff N, Mehari T, Nagel C, Aston PJ, Sundar A, Graff C, et al. PTB-XL+, a comprehensive electrocardiographic feature dataset. Scientific Data 2023;10:279.

[14] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation 2000;101(23):e215–e220.

[15] George B. Moody PhysioNet Challenge 2024. https://physionetchallenges.org/2024/. Accessed: 2024-09-30.

[16] Herman R, Demolder A, Vavrik B, Martonak M, Boza V, Kresnakova V, et al. Validation of an automated artificial intelligence system for 12-lead ECG interpretation. Journal of Electrocardiology 2024;82:147–154.

[17] Demolder A, Kresnakova V, Hojcka M, Boza V, Iring A, Rafajdus A, et al. High precision ECG digitization using artificial intelligence. medRxiv 2024;URL https://www.medrxiv.org/content/early/2024/09/01/2024.08.31.24312876.

Address for correspondence:

Matthew A Reyna

DBMI, 101 Woodruff Circle, 4th Floor East, Atlanta, GA 30322

matthew.a.reyna@emory.edu