Original Article

Healthc Inform Res. 2023 April;29(2):132-144.
https://doi.org/10.4258/hir.2023.29.2.132
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Standardized Database of 12-Lead Electrocardiograms with a Common Standard for the Promotion of Cardiovascular Research: KURIAS-ECG

Hakje Yoo[1,*], Yunjin Yum[2,*], Soo Wan Park[1], Jeong Moon Lee[1], Moonjoung Jang[1], Yoojoong Kim[3], Jong-Ho Kim[4], Hyun-Joon Park[5], Kap Su Han[6], Jae Hyoung Park[4], Hyung Joon Joo[1,4,7]

[1]Korea University Research Institute for Medical Bigdata Science, Korea University College of Medicine, Seoul, Korea
[2]Department of Biostatistics, Korea University College of Medicine, Seoul, Korea
[3]School of Computer Science and Information Engineering, The Catholic University of Korea, Bucheon, Korea
[4]Department of Cardiology, Cardiovascular Center, Korea University College of Medicine, Seoul, Korea
[5]Korea University Research Institute for Healthcare Service Innovation, Korea University College of Medicine, Seoul, Korea
[6]Department of Emergency Medicine, Korea University Anam Hospital, Korea University College of Medicine, Seoul, Korea
[7]Department of Medical Informatics, Korea University College of Medicine, Seoul, Korea

Objectives: Electrocardiography (ECG)-based diagnosis by experts cannot maintain uniform quality because individual differences may occur. Previous public databases can be used for clinical studies, but there is no common standard that would allow databases to be combined. For this reason, it is difficult to conduct research that derives results by combining databases. Recent commercial ECG machines offer diagnoses similar to those of a physician. Therefore, the purpose of this study was to construct a standardized ECG database using computerized diagnoses. **Methods:** The constructed database was standardized using Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) and Observational Medical Outcomes Partnership–common data model (OMOP-CDM), and data were then categorized into 10 groups based on the Minnesota classification. In addition, to extract high-quality waveforms, poor-quality ECGs were removed, and database bias was minimized by extracting at least 2,000 cases for each group. To check database quality, the difference in baseline displacement according to whether poor ECGs were removed was analyzed, and the usefulness of the database was verified with seven classification models using waveforms. **Results:** The standardized KURIAS-ECG database consists of high-quality ECGs from 13,862 patients, with about 20,000 data points, making it possible to obtain more than 2,000 for each Minnesota classification. An artificial intelligence classification model using the data extracted through SNOMED-CT showed an average accuracy of 88.03%. **Conclusions:** The KURIAS-ECG database contains standardized ECG data extracted from various machines. The proposed protocol should promote cardiovascular disease research using big data and artificial intelligence.

Keywords: Electrocardiograms, Database, Biological Ontologies, Artificial Intelligence, Cardiovascular Diseases

Submitted: September 28, 2022, **Revised:** February 22, 2023, **Accepted:** March 10, 2023

Corresponding Author
Hyung Joon Joo
Department of Cardiology, Cardiovascular Center, Korea University Anam Hospital, Korea University College of Medicine, Seoul 02841, Korea.
Tel: +82-2-919-8565, E-mail: drjoohj@gmail.com (https://orcid.org/0000-0003-1846-8464)

*These authors contributed equally to this work.

I. Introduction

Electrocardiograms (ECGs) are the most basic test used to diagnose or screen cardiac diseases [1-3]. Many studies have recently been conducted to advance the pre-processing and diagnostic algorithms of ECG signals using artificial intelligence (AI) and deep learning technologies. For these studies, accurate and consistent annotations of ECG diagnosis and classification, as well as sufficient and high-quality ECG data of various ECG diagnoses and classifications, are very important [4,5]. Several ECG databases (DBs) have been introduced, and recently published datasets contain many more ECG data than earlier datasets [6-17] (Table 1). Despite the existing massive ECG DBs, ECG diagnosis and classification are not standardized, and their distributions are skewed. For this reason, the development of algorithms combining various ECG DBs is limited. Moreover, because the data capacity of a 12-lead ECG is very considerable, it is important to construct an efficient dataset that can be effectively studied by researchers who may have limited infrastructure.

Prior datasets used their own ECG diagnoses and classifications. In most small ECG datasets, there are five to 10 diagnostic labels. Zheng et al. [16] recently released a large ECG dataset consisting of 10,646 ECGs annotated with 11 cardiac rhythms and 56 cardiovascular conditions redefined through human labeling. The PTB-XL dataset contains 71 ECG statements in accordance with the SCP-ECG standard [17] and also provides cross-references for other ECG annotation systems, including an ECG REFID identifier, CDISC code, and DICOM code. These annotation systems focus more on data operations, such as data transmission and storage, than on suitability for use by computational processing. The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), which was adopted for the present dataset, has gained popularity as a global standard terminology system to improve interoperability by covering all areas of the medical field [18]. In the United States, the Health IT Standards Committee has recommended using SNOMED CT for health information exchange. Several medical clinical DBs have recently been built using SNOMED CT, which makes it easier to construct a DB that can be merged with other clinical data [19].

Modern ECG machines generate digitized waveform data, computerized ECG parameter measurements, and diagnostic statements. The computerized interpretation algorithms of each ECG machine vendor adopt standard ECG measurement and diagnosis classification systems, such as the Minnesota code manual [20]. Although verification by experts

Table 1. Related electrocardiography datasets

Dataset	Number of subjects	Number of ECG	Lead	Record	Original frequency (Hz)	Diagnosis category	Abnormality (%)
AHA [6]	N/A	154	2	3 hr	250	8	100
European ST-T [7]	79	90	2	120 min	250	2	100
Long-term ST [8]	80	86	1	21, 24 hr	250	1	100
MIT-BIH Arrhythmia [9]	47	48	2	30 min	360	1	100
MIT-BIH Noise Stress Test [10]	15	15	1	12 half-hour, 3 half-hour	360	1	100
STAFF-III [11]	104	108	12	Various conditions	1,000	1	100
PTB Diagnostic ECG [12]	290	549	15	2 min	1,000	9	81
St Petersburg [13]	32	75	12	30 min	257	1	100
T-wave Alternans Challenge [14]	N/A	100	12	2 min	500	1	100
LUDB [15]	N/A	200	12	10 s	500	6	19
Zheng et al. [16]	10,646	10,646	12	10 s	500	68	37
PTB-XL [17]	18,885	21,837	12	10 s	400	71	57
Proposed	13,862	20,000	12	10 s	500	10	76

Clinical summary, including age; gender; diagnosis; and (where applicable) data on medical history, medication and interventions, coronary artery pathology, ventriculography, echocardiography, and hemodynamics.

ECG: electrocardiography, N/A: not applicable.

is the gold-standard method to confirm ECG diagnoses, human validation cannot avoid intra- and inter-observer variability, and maintaining data consistency has been noted as a challenge [4]. Automated ECG interpretation is often used in studies that construct big data sets, such as large population-based cohort studies. Moreover, previous studies have shown that the performance of recent computerized ECG interpretations is comparable to that of expert physicians, with correct classification percentages of 91.3% for the computer program and 96.0% for cardiologists, respectively [21,22]. ECG statements in existing ECG datasets, which are mostly relatively small, were labeled by physicians. The PTB-XL dataset, which is the largest ECG dataset, contains a mixture of ECG statements labeled by physicians and ECG statements automatically interpreted by an ECG machine.

The purpose of this study was to construct a high-quality, well-defined, and evenly distributed ECG dataset of sufficient size for research compiled into a practically usable DB. In this study, a pair of strategies were used to construct a high-quality DB; the first sought to establish a standardized system using standard vocabularies and Concept_IDs of SNOMED CT and Observational Medical Outcomes Partnership–common data model (OMOP-CDM) to overcome differences in the diagnoses made by different devices, while the second pursued the removal of poor-quality ECGs to prevent unnecessary data from being included in the DB. Various types of noise were also removed through a denoising process. Through this approach, 147 detailed ECG diagnoses were classified into 10 categories, and a DB containing at least 2,000 ECG data points for each category was constructed.

II. Methods

1. Database Construction

ECG signals were measured using equipment from GE (Boston, MA, USA) and Philips (Eindhoven, the Netherlands) in hospital settings, and all signals were ultimately saved in the clinical information system (CIS; INFINITT Healthcare Co., Seoul, South Korea) in XML file format. A total of 434,938 standard 12-lead ECGs were obtained from the CIS of Korea University Anam Hospital between January 1, 2017, and December 31, 2020. The study protocol was approved by the Institutional Review Board of the Korea University Anam Hospital (No. 2021AN0261), and the need for written informed consent was waived because of the retrospective study design with minimal risk to participants. The study complied with all relevant ethical regulations and the principles of the Declaration of Helsinki.

The digitized waveform data from the 12-lead ECG machine were automatically trimmed to 10 seconds with 500 Hz. The patient's basic information was input through the ECG machine by a nurse. The ECG data were stored in XML format on the CIS server and included metadata with each patient's basic personal information. The XML format contained basic examination information, technical data, eight ECG parameters, diagnosis statements, and waveform data. The basic examination information included the patient registration number, examination date and time, and examination equipment, as well as technical data such as the sampling rate, amplitude, and filtering frequency. A standard Python module (ElementTree XML API) was used to parse the data in the XML file of each ECG, and all associated programming source code was written in Python version 3.6.0.

2. ECG Diagnosis Standardization and Classification

The ECG machines automatically generated ECG diagnoses and ancillary descriptions through the approved computerized algorithm of each vendor (GE Medical and Philips Medical Systems). The ECG findings, including the ECG diagnosis and ancillary descriptions, were present in free text format in the "statement" section of the original XML files. These free texts were converted to the terminology of SNOMED CT and its cross-referenced terminology of OMOP-CDM [19,23]. OMOP-CDM is a standard data schema with a vocabulary [23]. The OMOP-CDM vocabulary adopts existing vocabularies rather than using de novo constructions; for example, the OMOP-CDM concept name "ECG normal" (Concept_ID: 4065279) originated from the SNOMED CT name "electrocardiogram normal (finding)" (SNOMED code 164854000). Both OMOP-CDM Concept_ID 4065279 and SNOMED code 164854000 define a "normal ECG." Standard terminology mapping for the ECG diagnosis was performed using web-based software, which incorporated an integrated algorithm using cosine similarity and a rule-based hierarchy (available at cdal.korea.ac.kr/ECG2CDM). The conversion accuracy was 99.9%. Using this software, free text in statements and comments in ECG XML files was converted into OMOP-CDM codes and terms, which could also be easily converted to SNOMED CT codes and terms using the concept table found at http://athena.ohdsi.org.

ECG diagnoses were further classified based on the Minnesota code manual, which has been used in many epidemiological studies and clinical trials. This system has also been reported to be predictive of future cardiovascular events and mortality [20]. The Minnesota classification includes nine categories, as follows: QRS axis deviation, high amplitude R wave, arrhythmia, atrioventricular (AV) conduction defect, ventricular conduction defect, Q and QS pattern, ST junction and segment depression, T wave item, and miscellaneous. Some ECG diagnoses were unclassified. Thus, the present study used 10 classification categories of the Minnesota code manual, including "unclassified." The Minnesota code manual also provided a list of minor and major code abnormalities that could be used for a subgroup analysis. Two professional cardiologists labeled the Minnesota code classification categories and abnormalities in terms of 147 detailed ECG diagnoses for the present dataset.

Figure 1 shows the distribution of the overall ECG diagnosis/classification; notably, the ECG diagnosis scheme within each Minnesota category was skewed. As shown in Table 2, the most common ECG diagnosis was "normal ECG" (41.51%) followed by "ECG: sinus rhythm" (5.42%) and "ECG: sinus bradycardia" (5.34%). The difference between the most common ECG diagnosis and the second most common (sinus rhythm) was more than 30 percentage points.

In this study, DB quality was improved by removing data containing severe noise based on machine-interpreted diagnoses. ECG signals contain common types of noise such as AC interference and baseline wander. However, data with patient movement, electrode attachment problems, and sudden increases in amplitude are unsuitable for research. Therefore, to improve the data quality, the present dataset excluded ECG data with either Concept_name (Concept_ID) "poor ECG quality" (OMOP-CDM code: 4088345) or "suspect arm ECG leads reversed" (OMOP-CDM code: 4088344). The original statements mapped onto "poor ECG quality," or "suspect arm ECG leads reversed" are shown in Table 3.

ECG cases were selected from the entire dataset through a total of three steps to improve the DB quality. In the first step, ECGs matching the ECG statement "poor ECG quality" or "suspect arm ECG lead reversed" were removed from the source data (n = 32,164). Second, ECG cases containing missing data for any ECG leads or sampling rates less than 500 Hz were also excluded (n = 7,644). In the third step, ECG data from patients' first visit to a hospital were selected from instances where there were multiple ECGs from the same patients, and 237,536 cases were excluded. ECGs from the first hospital visit were preferred to help reduce the confounding effect of subsequent data susceptible to other external factors, such as medications, and to reduce confounding and bias in the data due to treatment.

Among the remaining 157,594 ECG cases, all cases with fewer than 100 data of each ECG diagnosis between January 1, 2017, and December 31, 2020 were included. Then, 2,000 ECG data points (1 ECG per subject) corresponding to each of the 10 Minnesota classifications were consecutively selected from January 1, 2017 onwards. Finally, ECGs from 13,862 patients were included in the dataset.

3. Waveform Data Denoising

The KURIAS-ECG DB was constructed using raw information obtained from the ECG equipment and systems. KURIAS is an abbreviation for the Korea University Research Institute for Medical Bigdata Science, and KURIAS-ECG refers to the 12-lead ECG DB constructed by this research institution. The ECG signal acquired from ECG equipment contains complex noise due to the device's data transmission/reception, the location of electrodes, the patient's movement, muscle activity, and human body differences. Therefore, a preprocessing step was applied to the ECG signals to reduce low-quality data caused by noise (Figure 2).

In general, a signal containing unnecessary noise is ac-

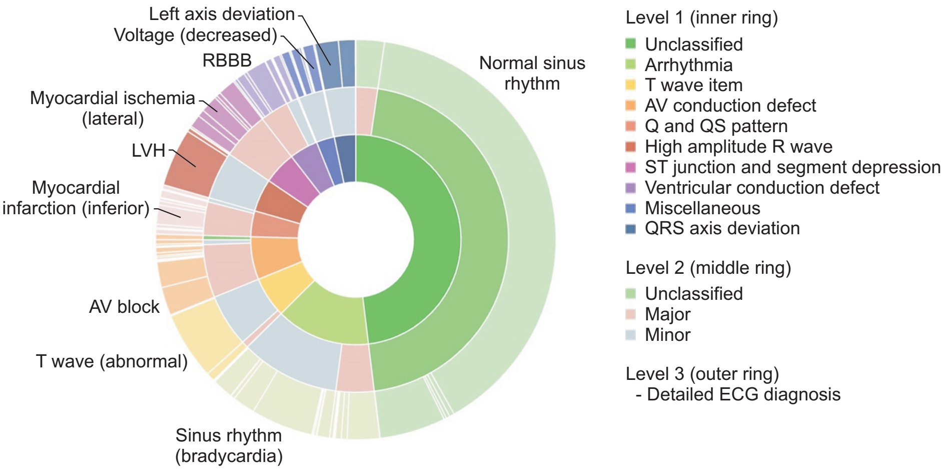

Figure 1. Graphical summary of the distribution of ECG diagnoses and classifications in the original source data of the ECG dataset (n = 434,938). Note that the distribution of ECG diagnoses is highly skewed. ECG: electrocardiography, LVH: left ventricular hypertrophy, RBBB: right bundle branch block, AV: atrioventricular.

Table 2. ECG diagnoses and classifications (top 10)

Concept name	Concept_ID	SNOMED_name	SNOMED_code	Minnesota classification	Abnormality	Proportion (%)
ECG normal, ECG: normal sinus rhythm	4065279, 4142265	Electrocardiogram normal (finding), Electrocardiogram: normal sinus rhythm (finding)	164854000, 426285000	Unclassified	Unclassified	41.51
ECG: sinus rhythm	4145513	Electrocardiogram: sinus rhythm (finding)	426783006	Unclassified	Unclassified	5.42
ECG: sinus bradycardia	4138456	Electrocardiogram: sinus bradycardia (finding)	426177001	Arrhythmia	Minor	5.34
EKG: T-wave abnormal	4065390	Electrocardiographic T-wave abnormal (finding)	164934002	T wave item	Minor	5.11
Atrioventricular block	316135	Atrioventricular block (disorder)	233917008	AV conduction defect	Major	2.37
EKG: left ventricle hypertrophy	4065282	Electrocardiographic left ventricle hypertrophy (finding)	164873001	High amplitude R wave	Minor	2.27
First-degree atrioventricular block	314379	First-degree atrioventricular block (disorder)	270492004	AV conduction defect	Major	2.12
Prolonged QT interval	4008859	Prolonged QT interval (finding)	111975006	Unclassified	Major	2.01
Left-axis deviation	4215406	Left-axis deviation (finding)	39732003	QRS axis deviation	Minor	1.83
ECG: atrial fibrillation	4064452	Electrocardiographic atrial fibrillation (finding)	164889003	Arrhythmia	Major	1.72

ECG: electrocardiography, SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms, EKG: Elektrokardiogramm, AV: atrioventricular.

quired due to various factors during an ECG examination. For diagnostic accuracy through monitoring, it is essential to remove the noise generated in the ECG signal. Butterworth filters are among the most commonly used signal-processing methods in the field of biomedical engineering [24]. In this study, the cut-off frequency was set from 0.05 to 150 Hz to minimize the distortion of the ST segment and to maintain the post-potential information of the QRS wave [25]. The baseline wander is the noise of low-frequency components caused by body movements, electrode movements, and breathing. In this study, the asymmetrically reweighted penalized least squares smoothing (arPLS) method was applied to overcome the baseline fluctuation problem due to the low-frequency components. Conventional polynomial methods have been proposed as effective techniques for removing baseline wander [26]. However, the arPLS method is effectively used to calculate a baseline for signals with various signal spectra, such as ECGs, by repeatedly changing the weights while estimating the baseline [27].

4. Validation of the Waveform Database

To demonstrate that the KURIAS-ECG DB was objectively of high quality, a pair of validation methods were employed. The first aimed to quantitatively verify whether noisy data were reduced due to the removal of poor-quality ECGs by analyzing the waveform difference between the DB from which poor-quality ECGs were removed and the DB without the removal of such ECGs. Second, to verify the effect of the diagnosis standardized by SNOMED-CT and Concept_ID, the possibility of standardized diagnoses was verified by extracting data for seven diagnoses according to Concept_ID and developing a classification AI model for each group.

Poor-quality ECGs were removed using the corresponding Concept_ID. To verify the quality of the waveforms constituting the DB, signal-processing analysis was performed on the waveform of lead II, which is most often used for rhythm [28]. All ECG waveforms have baseline wander. In this study, the baseline wander present in the ECG waveform was acquired through a signal-processing method, and both data sets were quantitatively compared using the difference between the highest and smallest baseline wander values.

Table 3. Exclusion criteria

Concept name	Vendor	Statement
Poor ECG quality	GE	Poor data quality, interpretation may be adversely affected Acquisition hardware fault prevents reliable analysis, carefully check ECG record before interpreting Baseline wander Current undetermined rhythm precludes rhythm comparison, needs review Electrode noise Muscle tremor Poor data quality Poor data quality in current ECG precludes serial comparison
	Philips	All 12 leads are missing Artifact in lead(s) Artifact in lead(s) and baseline wander in lead(s) Baseline wander in lead(s) Incomplete analysis due to missing data in precordial lead(s) Missing lead(s) Missing lead(s) and partial lead(s) Poor-quality data - please repeat ECG!
Suspect arm ECG leads reversed	GE	Suspect arm lead reversal, interpretation assumes no reversal Arm lead reversal
	Philips	Left arm and left leg electrode reversal Probable extremity electrode reversal Right and left arm electrode reversal Right arm and left leg electrode reversal

In addition, to confirm the usefulness of diagnostic standardization, a single classification model for seven diagnoses was developed based on the ECG waveforms. In this study, the residual blocks-based network (ResNet) was used as a model to develop seven diagnostic classification models. In this study, ResNet was used to develop seven diagnostic classification models. ResNet has been effectively used to classify cardiovascular diseases using ECG waveforms in previous studies [29]. As input variables of the model, the waveforms of lead I, lead II, and V2 and min-max normalization were applied. To check the data quality of various diagnoses, seven diagnoses (normal sinus rhythm, sinus bradycardia, left-axis deviation, atrial fibrillation, first-degree atrioventricular block, Wolff-Parkinson-White syndrome, and prolonged QT interval) were defined as target variables. All diagnoses were classified based on the Concept_ID used for standardization.

The dataset was divided into subsets for training, validation, and testing at a 6:2:2 ratio, and the accuracy of the model was evaluated through 10-fold cross-validation. An Adam optimizer was adopted, and the learning rate was set to 0.0001. The accuracy of the model was evaluated by calculating the average of the accuracy, recall, and F1-score.

III. Results

The established DB consisted of 13,862 patients, including 7,840 men and 6,022 women. Detailed characteristics of the patients and ECGs are shown in Table 4. Most of the ECG signals were measured using GE (88.43%). The average number of Minnesota code categories per ECG test was 2 ± 1.05, and the average number of ECG diagnoses per ECG test was 3 ± 1.49. The three most-common ECG diagnoses within each Minnesota code category are presented in Table 4. The ECG diagnosis within each category was less skewed than that in the original dataset. In addition, ECG diagnoses with a low diagnostic rate, such as indeterminate axis, were also included in a relatively high proportion.

The overall distribution of the constructed DB is shown in Figure 3. Since the same number of ECG data points were selected for each Minnesota code classification, the data-composition ratio of ECG diagnoses with low frequency, such as QRS axis deviation, increased. Thus, the distribution

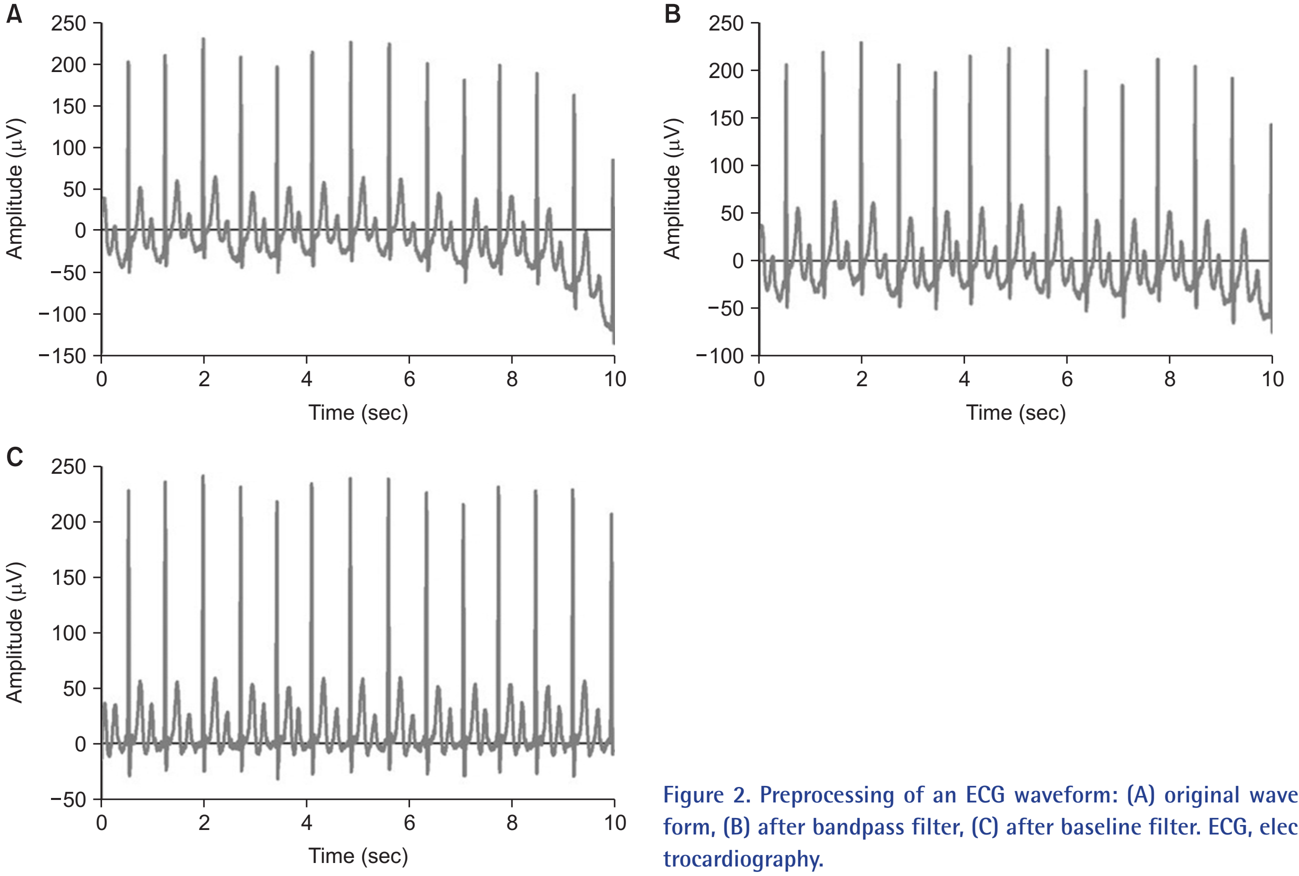

Figure 2. Preprocessing of an ECG waveform: (A) original waveform, (B) after bandpass filter, (C) after baseline filter. ECG, electrocardiography.

of ECG diagnoses for the 10 Minnesota categories was less skewed than the original data for abnormalities classified as normal, minor, and major.

1. Database Content

The data attributes of the constructed ECG DB are described in Table 5. The ECG DB consisted of four sections—namely, general metadata, analyzed parameters, diagnosis statements, and standard and waveform data.

The general metadata section included person ID, sex, age, acquisition data, acquisition time, and information from the device manufacturer. For this study, person ID was a randomly assigned number for pseudonymization different from the hospital's patient ID. The analysis parameters included heart rate, PR interval, ARS duration, QT interval, QT corrected, P axis, R axis, and T axis, which were automatically analyzed during a post-processing period by machine. "NULL" indicated parameters not calculated automatically in the waveform. The diagnosis statements and standards section consisted of the diagnosis statements automatically analyzed by the post-processing system and the results of mapping the OMOP-CDM vocabulary, SNOMED-CT codes, and Minnesota classifications corresponding to the diagnosis statements. The Concept_ID was obtained by analyzing the similarity between the diagnosis statement and the OMOP-CDM vocabulary and automatically mapping the result as a SNOMED-CT code. The Concept_name was the clinical vocabulary for each Concept_ID. Because a single ECG signal can have multiple diagnostic statements, multiple Concept _ID and Concept _name results can exist for that ECG. In addition, the diagnosis statement was also mapped to SNOMED-CT because the SNOMED_code and SNOMED_name were mapped to Concept_ID and Concept_name, respectively. The Minnesota code classified as abnormality was obtained by standardizing the diagnosis by categorizing the Concept_ID according to the Minnesota classification. The waveform data consisted of 12 ECG signals, and noise was removed through signal processing.

2. Validation of Waveforms Using the Baseline Gap

The difference between the datasets with and without poor-quality ECGs was analyzed by calculating the magnitude of the baseline variability of the waveforms acquired under both conditions. The dataset of waveforms without poor-quality ECGs had a narrow baseline displacement, as shown in Figure 4A, with an average of 44.54 μV (max, 244.11 μV; min, 7.70 μV). In contrast, the dataset including poor-quality ECGs had a wide baseline displacement, as shown in Figure

Table 4. Characteristics of the ECGs

	Total (n = 13,862)	Men (n = 7,840)	Women (n = 6,022)
Age (yr)	58.81 ± 20.12	56.94 ± 19.88	61.34 ± 20.17
Vendor			
GE	12,258 (88.43)	6,899 (88.00)	5,359 (88.99)
Philips	1,604 (11.57)	941 (12.00)	663 (11.01)
Number of Minnesota codes per ECG	2 ± 1.05	2 ± 1.06	2 ± 1.03
Number of ECG diagnoses per ECG	3 ± 1.49	3 ± 1.50	2 ± 1.47
Unclassified			
Sinus rhythm	329 (16.45)	178 (8.90)	151 (7.55)
Sinus arrhythmia	329 (16.45)	153 (7.65)	176 (8.80)
QT interval (prolonged)	329 (16.45)	155 (7.75)	174 (8.70)
QRS axis deviation			
Left axis deviation	944 (47.20)	647 (32.35)	297 (14.85)
Right axis deviation	943 (47.15)	472 (23.60)	471 (23.55)
Indeterminate axis	110 (5.50)	66 (3.30)	44 (2.20)
High-amplitude R-wave			
LVH	859 (42.95)	564 (28.20)	295 (14.75)
RVH	860 (43.00)	502 (25.10)	358 (17.90)
Ventricular hypertrophy	281 (14.05)	203 (10.15)	78 (3.90)
Arrhythmia			
Sinus rhythm (bradycardia)	119 (5.95)	64 (3.20)	55 (2.75)
Atrial fibrillation	119 (5.95)	82 (4.10)	37 (1.85)
Sinus rhythm (tachycardia)	119 (5.95)	61 (3.05)	58 (2.90)
AV conduction defect			
AV block	164 (8.20)	107 (5.35)	57 (2.85)
AV block (1st degree)	164 (8.20)	108 (5.40)	56 (2.80)
PR interval (short)	164 (8.20)	79 (3.95)	85 (4.25)
Ventricular conduction defect			
RBBB	205 (10.25)	129 (6.45)	76 (3.80)
RBBB (incomplete)	205 (10.25)	144 (7.20)	61 (3.05)
rSr pattern in V1 and V2	205 (10.25)	110 (5.50)	95 (4.75)
Q and QS pattern			
Myocardial infarction (inferior)	205 (10.25)	117 (5.85)	88 (4.40)
Myocardial infarction (septal)	205 (10.25)	145 (7.25)	60 (3.00)
Myocardial infarction (anterior)	205 (10.25)	120 (6.00)	85 (4.25)
ST junction and segment depression			
Myocardial ischemia (lateral)	297 (14.85)	152 (7.60)	145 (7.25)
ST–T abnormality (non-specific)	297 (14.85)	120 (6.00)	177 (8.85)
Myocardial ischemia (anterior)	297 (14.85)	97 (4.85)	200 (10.00)
T wave item			
T wave (abnormal)	884 (44.20)	365 (18.25)	519 (25.95)
T wave (inverted)	883 (44.15)	382 (19.10)	501 (25.05)
T wave (flattened)	233 (11.65)	135 (6.75)	98 (4.90)

Continued on the next page.

Table 4. Continued

	Total (n = 13,862)	Men (n = 7,840)	Women (n = 6,022)
Miscellaneous			
ST segment elevation	319 (15.95)	288 (14.40)	31 (1.55)
P wave (abnormal)	319 (15.95)	143 (7.15)	176 (8.80)
Voltage (decreased)	319 (15.95)	141 (7.05)	178 (8.90)

Values are presented as mean ± standard deviation or number (%).

The 3 most-common ECG diagnoses comprising each Minnesota classification were used, abbreviated, and slightly modified to reduce space.

ECG: electrocardiography, AV: atrioventricular, RBBB: right bundle branch block, LVH: left ventricular hypertrophy, RVH: right ventricular hypertrophy.

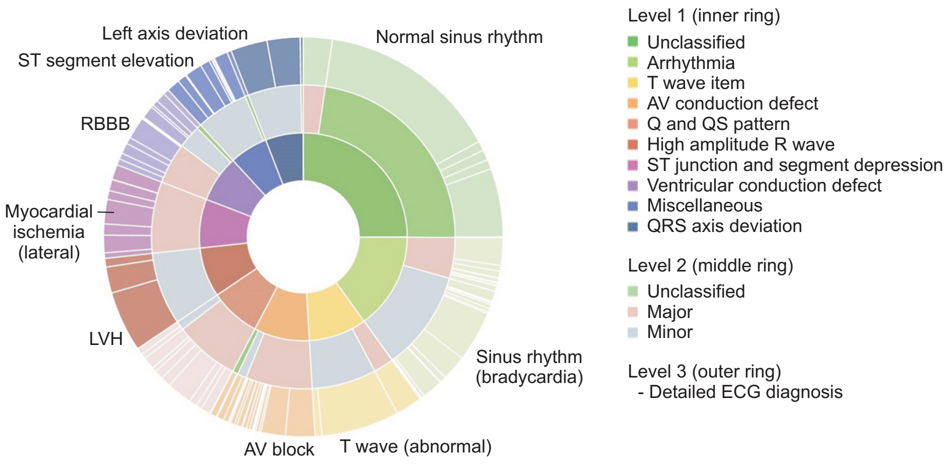

Figure 3. Graphical summary of the distribution of ECG diagnoses and classifications in the extracted source data of the original ECG dataset. Note that the distribution of ECG diagnoses is less skewed than in the original source data. ECG: electrocardiography, LVH: left ventricular hypertrophy, RBBB: right bundle branch block, AV: atrioventricular.

4B, with an average of 50.33 μV (min–max, 5.36–431.59 μV). As shown in Figure 4D, the baseline displacement of the dataset applying case sampling was statistically significantly lower than that of the dataset where case sampling was not applied ($p < 0.01$).

3. Validation of the Database Using Deep Learning

To verify the quality of the waveform data of the KURIAS-ECG DB, classification models were developed using waveform data for seven diagnostic categories extracted based on Concept_ID. The deep learning model developed to verify the waveform quality of the DB showed an average accuracy of 88.03% in the classification model for seven categories. Furthermore, the average F1-score was 0.88, and the average values of precision and recall were 0.87 in both results. The lowest accuracy was obtained for the classification model for prolonged QT interval (82.25%), and the highest accuracy was obtained for the classification model for atrial fibrillation (90.84%) (Table 6).

IV. Discussion

The KURIAS-ECG DB represents a new type of DB constructed by standardizing various types of 12-lead ECGs and applying a method to extract high-quality data. To construct the KURIAS-ECG DB, a total of 434,938 12-lead ECGs acquired over 4 years at a general hospital were used. The KURIAS-ECG DB overcomes the differences in diagnostic information among devices by establishing a common management standard using Concept_ID. In addition, this DB is balanced by subdividing diagnostic information from 147 ECG diagnoses into 10 categories using the Minnesota classification. As research on cardiac disease using AI is carried out, the importance of ECG waveforms is growing. In previous studies, AI models were developed to classify cardiovascular diseases, such as atrial fibrillation, arrhythmias, and heart failure based on ECG waveforms, and showed an accuracy of 85%–95% [29]. Moreover, Yoo et al. [30] applied ECG waveforms to an AI model to classify neurological diseases, such as Parkinson disease, that are accompanied by changes in cardiac movement. The Parkinson disease classification model using ECG waveforms achieved an 87% accuracy for

Table 5. Data attributes

Section	Variables	Data type	Description
General metadata	PersonID	Varchar	De-identified patient identifier
	Gender	Varchar	Male, female
	Age	Float	Age at recording in years
	AcquisitionDate	Date	ECG recording date
	AcquisitionTime	Time	ECG recording time
	Device_manuf	Varchar	Acquisition device manufacturer
Analyzed parameters	HeartRate	Float	Speed of heartbeat
	PRInterval	Float	PR interval in msec
	QRSDuration	Float	QRS duration in msec
	QTInterval	Float	QT interval in msec
	QTCorrected	Float	Corrected QT interval in msec
	PAxis	Float	P axis
	RAxis	Float	R axis
	TAxis	Float	T axis
Diagnosis statements and standard	Statement	Varchar	Automatically interpreted diagnosis statement
	Concept_ID	Varchar	The unique identifier of OHDSI OMOP-CDM vocabulary for each concept of ECG findings or disorders derived from SNOMED-CT
	Concept_Name	Varchar	The name of Concept_ID in OHDSI OMOP-CDM vocabulary
	SNOMED_Code	Varchar	The corresponding SNOMED-CT identifier for each concept of OHDSI OMOP-CDM vocabulary
	SNOMED_Name	Varchar	The name of SNOMED_Code in SNOMED-CT
	Minnesota	Categorical	The category corresponding to the Minnesota code classification system (2009) for each concept
	Abnormality	Categorical	The category corresponding to the Minnesota code's abnormality classification system for each concept
Waveform data	Wave_I	Varchar	Waveform data of lead I
	Wave_II	Varchar	Waveform data of lead II
	Wave_III	Varchar	Waveform data of lead III
	Wave_aVR	Varchar	Waveform data of aVR
	Wave_aVL	Varchar	Waveform data of aVL
	Wave_aVF	Varchar	Waveform data of aVF
	Wave_V1	Varchar	Waveform data of V1
	Wave_V2	Varchar	Waveform data of V2
	Wave_V3	Varchar	Waveform data of V3
	Wave_V4	Varchar	Waveform data of V4
	Wave_V5	Varchar	Waveform data of V5
	Wave_V6	Varchar	Waveform data of V6

ECG: electrocardiography, OHDSI: Observational Health Data Sciences and Informatics, OMOP: Observational Medical Outcomes Partnership, CDM: common data model, SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.

clinical patients, indicating its potential for future clinical application. Since the ECG waveform contains important information on the function of the heart, it is medical data that can be effectively used in AI research on heart-related diseases [29]. A 12-lead ECG measures the heart's electrical activity by attaching 10 electrodes to the limbs and chest.

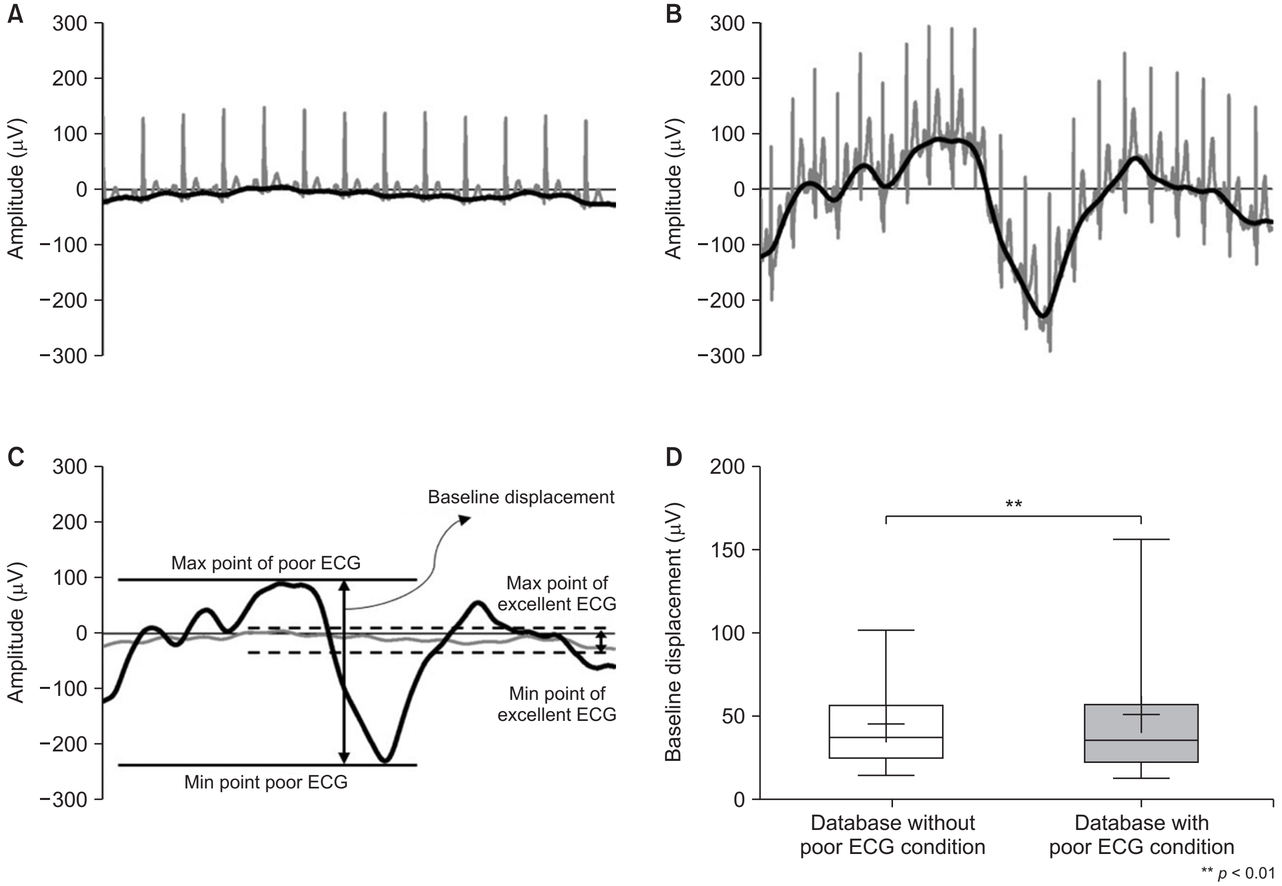

Figure 4. Representative ECG waveform and baseline of (A) an excellent ECG and (B) a poor ECG. (C)The definition of the baseline and baseline displacement of ECG waveforms without poor ECG conditions and with poor ECG conditions. (D) Comparison of the difference in baseline displacement between datasets with or without poor ECG conditions (plus point, median; box, 25%–75% range; whisker, 5th–95th percentiles).

Table 6. Performance of the classification models for seven diagnoses

Category	Accuracy (%)	F1-score	Precision	Recall
Normal sinus rhythm	88.33	0.88	0.91	0.85
Sinus bradycardia	87.90	0.88	0.87	0.88
Left axis deviation	88.84	0.88	0.86	0.90
Atrial fibrillation	90.84	0.91	0.91	0.91
First degree atrioventricular block	87.82	0.88	0.89	0.87
Wolff-Parkinson-White syndrome	90.22	0.95	0.84	0.87
Prolonged QT interval	82.25	0.82	0.83	0.82

The types of noise that occur during this process include AC interference, muscle tremors, baseline wander, and motion artifacts. As this noise is not generated by the heart function, it can reduce the accuracy of the AI model. In addition, commercial 12-lead ECG machines used in medical institutions have different diagnosis systems provided by the machine depending on the manufacturer; for example, GE refers to normal signals as “normal sinus rhythm” or “normal ECG,” while Philips uses the term “sinus rhythm.” These differences cause difficulties in extracting data in AI research that requires the use of big data, and these challenges must be resolved during the DB construction step. The KURIAS-ECG DB presented in this study applied a standardization strategy using the OMOP-CDM international standard and a high-quality strategy to exclude poor-quality ECGs containing noise. This systematic DB construction process is effective for cardiovascular disease AI research because it can obtain large amounts of data while excluding unsuitable data. ECG waveforms stored in a tree-type XML format can be difficult to use directly in research due to their data format. However, the KURIAS-ECG approach in this study extracts and manages tree-type data in cell units, making it efficient in terms of storage space utilization and easy to convert to 5,000 frames, which correspond to the original format. In the future, this approach will be a solution that can combine distributed ECG data and public DBs into a common DB through Concept_ID assignment, which is used as the standard system in KURIAS-ECG. This is expected to provide an environment for promoting research on cardiac disease using AI. The KURIAS-ECG DB is meaningful in that it is a high-quality DB of 12-lead ECGs, but there are limitations in its usability, since it was released as a limited DB on an open data platform. Since this problem is shaped by the disclosure policies of the organization providing the data, various hurdles must be overcome to construct an open DB. Therefore, in this study, the construction process of a standardization DB and related code were disclosed so that each institution can build a standardized DB without disclosing data. The construction protocol of the KURIAS-ECG DB was published on PhysioNet, and the code used for database management system (DBMS) construction and noise removal was shared through GitHub (https://github.com/KURIAS).

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2021R1I1A1A01059747).

ORCID

Hakje Yoo (http://orcid.org/0000-0003-4341-5540)
Yunjin Yum (http://orcid.org/0000-0003-3070-3615)
Soo Wan Park (https://orcid.org/0000-0002-1657-3810)
Jeong Moon Lee (https://orcid.org/0000-0003-4020-4561)
Moonjoung Jang (https://orcid.org/0000-0002-6506-4254)
Yoojoong Kim (https://orcid.org/0000-0002-6615-9116)
Jong-Ho Kim (https://orcid.org/0000-0002-1309-0821)
Hyun-Joon Park (https://orcid.org/0000-0002-0394-9030)
Kap Su Han (http://orcid.org/0000-0003-0205-1269)
Jae Hyoung Park (http://orcid.org/0000-0001-8434-0157)
Hyung Joon Joo (http://orcid.org/0000-0003-1846-8464)

References

1. Maron BJ, Friedman RA, Kligfield P, Levine BD, Viskin S, Chaitman BR, et al. Assessment of the 12-lead ECG as a screening test for detection of cardiovascular disease in healthy general populations of young people (12-25 years of age): a scientific statement from the American Heart Association and the American College of Cardiology. Circulation 2014;130(15):1303-34. https://doi.org/10.1161/CIR.0000000000000025
2. Hao P, Gao X, Li Z, Zhang J, Wu F, Bai C. Multi-branch fusion network for Myocardial infarction screening from 12-lead ECG images. Comput Methods Programs Biomed 2020;184:105286. https://doi.org/10.1016/j.cmpb.2019.105286
3. Rajkumar A, Ganesan M, Lavanya R. Arrhythmia classification on ECG using deep learning. Proceedings of 2019 5th International Conference on advanced Computing & Communication Systems (ICACCS); 2019 Mar 15-16; Coimbatore, India. p. 365-9. https://doi.org/10.1109/ICACCS.2019.8728362
4. Liu X, Wang H, Li Z, Qin L. Deep learning in ECG diagnosis: a review. Knowl Based Syst 2021;227:107187. https://doi.org/10.1016/j.knosys.2021.107187
5. Hong S, Zhou Y, Shang J, Xiao C, Sun J. Opportunities and challenges of deep learning methods for electrocardiogram data: a systematic review. Comput Biol Med 2020;122:103801. https://doi.org/10.1016/j.compbiomed.2020.103801
6. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000;101:e215-e220. http://circ.ahajournals.org/cgi/content/full/101/23/e215
7. Taddei A, Distante G, Emdin M, Pisani P, Moody GB, Zeelenberg C, et al. The European ST-T database: standard for evaluating systems for the analysis of ST-T changes in ambulatory electrocardiography. Eur Heart J 1992;13(9):1164-72. https://doi.org/10.1093/oxfordjournals.eurheartj.a060332

8. Jager F, Taddei A, Moody GB, Emdin M, Antolic G, Dorn R, et al. Long-term ST database: a reference for the development and evaluation of automated ischaemia detectors and for the study of the dynamics of myocardial ischaemia. Med Biol Eng Comput 2003;41(2):172-82. https://doi.org/10.1007/BF02344885
9. Moody GB, Mark RG. The impact of the MIT-BIH arrhythmia database. IEEE Eng Med Biol Mag 2001;20(3):45-50. https://doi.org/10.1109/51.932724
10. Moody GB, Muldrow W, Mark RG. A noise stress test for arrhythmia detectors. Comput Cardiol 1984;11(3):381-4.
11. Laguna P, Sornmo L. The STAFF III ECG database and its significance for methodological development and evaluation. J Electrocardiol 2014;47(4):408-17. https://doi.org/10.1016/j.jelectrocard.2014.04.018
12. Kreiseler D, Bousseljot R. Automatisierte EKG-Auswertung mit Hilfe der EKG-Signaldatenbank CARDIODAT der PTB. Biomed Tech (Berl) 1995;40(s1):319-20. https://doi.org/10.1515/bmte.1995.40.s1.319
13. Yakushenko E. St Petersburg INCART 12-lead Arrhythmia Database [Internet]. Cambridge (MA): PhysioNet; 2008 [cited at 2023 Mar 30]. Available from: https://physionet.org/content/incartdb/1.0.0/
14. Moody GB. The PhysioNet/Computers in Cardiology challenge 2008: T-wave alternans. Proceedings of 2008 Computers in Cardiology; 2008 Sep 14-17; Bologna, Italy. p. 505-8. https://doi.org/10.1109/CIC.2008.4749089
15. Kalyakulina AI, Yusipov II, Moskalenko VA, Nikolskiy AV, Kosonogov KA, Osipov GV, et al. LUDB: a new open-access validation tool for electrocardiogram delineation algorithms. IEEE Access 2020;8:186181-90. https://doi.org/10.1109/ACCESS.2020.3029211
16. Zheng J, Zhang J, Danioko S, Yao H, Guo H, Rakovski C. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. Sci Data 2020;7(1):48. https://doi.org/10.1038/s41597-020-0386-x
17. Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, et al. PTB-XL, a large publicly available electrocardiography dataset. Sci Data 2020;7(1):154. https://doi.org/10.1038/s41597-020-0495-6
18. Donnelly K. SNOMED-CT: the advanced terminology and coding system for eHealth. Stud Health Technol Inform 2006;121:279-90.
19. Lee D, de Keizer N, Lau F, Cornet R. Literature review of SNOMED CT use. J Am Med Inform Assoc 2014;21(e1):e11-9. https://doi.org/10.1136/amiajnl-2013-001636
20. Prineas RJ, Crow RS, Zhang ZM. The Minnesota code manual of electrocardiographic findings. New York (NY): Springer Science & Business Media; 2009.
21. Willems JL, Abreu-Lima C, Arnaud P, van Bemmel JH, Brohet C, Degani R, et al. The diagnostic performance of computer programs for the interpretation of electrocardiograms. N Engl J Med 1991;325(25):1767-73. https://doi.org/10.1056/NEJM199112193252503
22. Smulyan H. The computerized ECG: friend and foe. Am J Med 2019;132(2):153-60. https://doi.org/10.1016/j.amjmed.2018.08.025
23. Makadia R, Ryan PB. Transforming the premier perspective hospital database into the Observational Medical Outcomes Partnership (OMOP) common data model. EGEMS (Wash DC) 2014;2(1):1110. https://doi.org/10.13063/2327-9214.1110
24. Altay Y, Kremlev A, Zimenko K, Margun A. The effect of filter parameters on the accuracy of ECG signal measurement. Biomed Eng 2019;53(3):176-80. https://doi.org/10.1007/s10527-019-09903-2
25. Sohn J, Yang S, Lee J, Ku Y, Kim HC. Reconstruction of 12-lead electrocardiogram from a three-lead patch-type device using a LSTM network. Sensors (Basel) 2020;20(11):3278. https://doi.org/10.3390/s20113278
26. Gan F, Ruan G, Mo J. Baseline correction by improved iterative polynomial fitting with automatic threshold. Chemometr Intell Lab Syst 2006;82(1-2):59-65. https://doi.org/10.1016/j.chemolab.2005.08.009
27. Baek SJ, Park A, Ahn YJ, Choo J. Baseline correction using asymmetrically reweighted penalized least squares smoothing. Analyst 2015;140(1):250-7. https://doi.org/10.1039/c4an01061b
28. Meek S, Morris F. ABC of clinical electrocardiography. Introduction. I-Leads, rate, rhythm, and cardiac axis. BMJ 2002;324(7334):415-8. https://doi.org/10.1136/bmj.324.7334.415
29. Ebrahimi Z, Loni M, Daneshtalab M, Gharehbaghi A. A review on deep learning methods for ECG arrhythmia classification. Exp Syst Appl 2020;7:100033. https://doi.org/10.1016/j.eswax.2020.100033
30. Yoo H, Chung SH, Lee C-N, Joo HJ. Deep Learning Algorithm of 12-Lead Electrocardiogram for Parkinson Disease Screening. J Parkinsons Dis 2023(Preprint):1-12. https://doi.org/10.3233/JPD-223549