Journal of Electrocardiology 67 (2021) 124–132

Contents lists available at ScienceDirect

Journal of Electrocardiology

journal homepage: www.jecgonline.com

Detection and classification of arrhythmia using an explainable deep learning model

Yong-Yeon Jo, PhD [a,1], Joon-myoung Kwon, MD, MS [a,b,c,d,1,*], Ki-Hyun Jeon, MD, MS [b,e], Yong-Hyeon Cho, BS [a], Jae-Hyun Shin, BS [a], Yoon-Ji Lee, BS [a], Min-Seung Jung, BS [a], Jang-Hyeon Ban, MS [d], Kyung-Hee Kim, MD, PhD [b,e], Soo Youn Lee, MD, MS [b,e], Jinsik Park, MD, PhD [e], Byung-Hee Oh, MD, PhD [e]

[a] *Medical Research Team, Medical AI, Co., Seoul, South Korea*
[b] *Artificial Intelligence and Big Data Research Center, Sejong Medical Research Institute, Bucheon, South Korea*
[c] *Department of Critical Care and Emergency Medicine, Mediplex Sejong Hospital, Incheon, South Korea*
[d] *Medical R&D Center, Body Friend, Co., Seoul, South Korea*
[e] *Division of Cardiology Cardiovascular Center, Mediplex Sejong Hospital, Incheon, South Korea*

ARTICLE INFO

Keywords:
Electrocardiography
Artificial intelligence
Arrhythmia
Deep learning

ABSTRACT

Background: Early detection and intervention is the cornerstone for appropriate treatment of arrhythmia and prevention of complications and mortality. Although diverse deep learning models have been developed to detect arrhythmia, they have been criticized due to their unexplainable nature. In this study, we developed an explainable deep learning model (XDM) to classify arrhythmia, and validated its performance using diverse external validation data.
Methods: In this retrospective study, the Sejong dataset comprising 86,802 electrocardiograms (ECGs) was used to develop and internally variate the XDM. The XDM based on a neural network-backed ensemble tree was developed with six feature modules that are able to explain the reasons for its decisions. The model was externally validated using data from 36,961 ECGs from four non-restricted datasets.
Results: During internal and external validation of the XDM, the average area under the receiver operating characteristic curves (AUCs) using a 12-lead ECG for arrhythmia classification were 0.976 and 0.966, respectively. The XDM outperformed a previous simple multi-classification deep learning model that used the same method. During internal and external validation, the AUCs of explainability were 0.925–0.991.
Conclusion: Our XDM successfully classified arrhythmia using diverse formats of ECGs and could effectively describe the reason for the decisions. Therefore, an explainable deep learning methodology could improve accuracy compared to conventional deep learning methods, and that the transparency of XDM can be enhanced for its application in clinical practice.

Introduction

The prevalence of arrhythmia is estimated to be 2%–5% worldwide, and it increases with age [1]. Cardiovascular diseases such as arrhythmia, continue to increase in prevalence at an alarming rate, and contribute a significant healthcare burden and morbidity worldwide. Atrial fibrillation (AF) affects at least 2.3 million people in the United States and is associated with increased risk of stroke and mortality [2–5]. Approximately 90,000 cases of supraventricular tachycardia (SVT) are detected annually in the United States. Moreover, ventricular arrhythmias cause 75%–80% of sudden cardiac deaths, resulting in an estimated 184,000 to 450,000 deaths in the United States annually [6].

Accurate interpretation of electrocardiography (ECG) is a cornerstone for arrhythmia diagnosis. However, arrhythmia is often paroxysmal, and is therefore difficult to detect. Accurate detection of the clinical condition presented by an ECG signal is challenging [7]. Consequently, using out-of-hospital ECG devices is important to detect arrhythmia and prevent irreversible complications and mortality [20]. However, the use of lifestyle ECG devices has been limited to detect arrhythmia given that existing commercial computer-aided interpretation still presents substantial rates of misdiagnoses [8].

Recently, deep learning has demonstrated high accuracy and applicability in computer vision, speech recognition, and signal processing. In particular, in the medical domain, deep learning has been applied to interpret arrhythmia using ECG. In previous studies, deep learning models were able to detect arrhythmia successfully, similar to the

[*] Corresponding author at: Department of Emergency Medicine, Mediplex Sejong Hospital, Incheon, South Korea.
E-mail address: kwonjm@sejongh.co.kr (J. Kwon).
[1] These two authors contributed equally to this work.

https://doi.org/10.1016/j.jelectrocard.2021.06.006
0022-0736/© 2021 Published by Elsevier Inc.

performance of a cardiologist [9–11]. However, the deep learning models developed in previous studies were simply black boxes that did not explain their predictions in a way that humans could understand [12]. Consequently, clinicians could not trust this new technology due to the lack of transparency and interpretability, which are key to promote its use in clinical settings. In this study, we developed and validated an explainable deep learning model (XDM) based on a neural network-backed ensemble tree (NBET), i.e., a highly fashioned artificial intelligence (AI) technology. To the best of our knowledge, this is the first study to develop and validate an explainable AI to detect arrhythmia.

Methods

Study design and population

We conducted a retrospective multicenter cohort study in which an XDM was developed using ECGs. The Sejong ECG dataset from Mediplex Sejong Hospital (MSH) and Sejong General Hospital (SGH) was used for the development and internal validation of the XDM. In the Sejong ECG dataset, we identified patients with at least one standard digital 10-s 12-lead ECG acquired in the supine position within the study period, and at least one outpatient department visit, general-health checkup center visit, or admission to the cardiovascular center of the two aforementioned hospitals. We excluded individuals with missing demographic, electrocardiographic, or medical records relating to diagnosis and intervention procedures, as shown in Fig. 1. Study populations from the Sejong ECG dataset were randomly split into algorithm-development (75%) and internal-validation (25%) datasets. We externally validated the developed XDM using four non-restricted ECG datasets, and 36,961 ECGs that had arrhythmia labels were used as the external validation dataset. The Physikalisch Technische Bundesanstalt (PTB-XL) ECG dataset from Europe contained 18,065 ECGs with a sampling rate of 500 Hz [13]; the Georgia ECG challenge dataset from the United States contained 5541 ECGs with a sampling rate of 500 Hz [14]; the Chapman university ECG database from China contained 9269 ECGs with a sampling rate of 500 Hz [15]; and the China Physiological Signal Challenge (CPSC) ECG dataset from China contained 4086 ECGs with a sampling rate of 500 Hz [16]. Given that the developed XDM can be used with diverse formats of ECGs, we were able to confirm the performance of the deep learning model (DLM) using single-lead ECG (lead I) from validation datasets.

This study was approved by the Institutional Review Boards (IRBs) of SGH (2019–0411) and MSH (2019–083). Clinical data included digitally stored ECGs, medical records, intervention results, and demographic information from both hospitals. Both IRBs waived the need for informed consent due to the retrospective nature of the study, and the fact that only fully anonymized ECGs and health data were used.

Procedures

ECG data were used as predictor variables. The digitally stored 12-lead Sejong ECG dataset, amounting to 5000 per lead, were recorded over 10 s (500 Hz). We removed 1 s at the beginning and end of each ECG because these areas had more artifacts than other parts. Given four open datasets were used as an external validation dataset, we used only 8 s of ECG data extracted from the middle of each ECG. For example, if the length of the external validation ECG data was 30 s, we used only 8 s of ECG data extracted from the middle of those 30 s; consequently, the length of each ECG was 8 s (4,000 data points). The objective of this study was to classify arrhythmia, defined as normal sinus rhythm (NSR), atrial fibrillation or flutter (AF or AFL), junctional rhythm

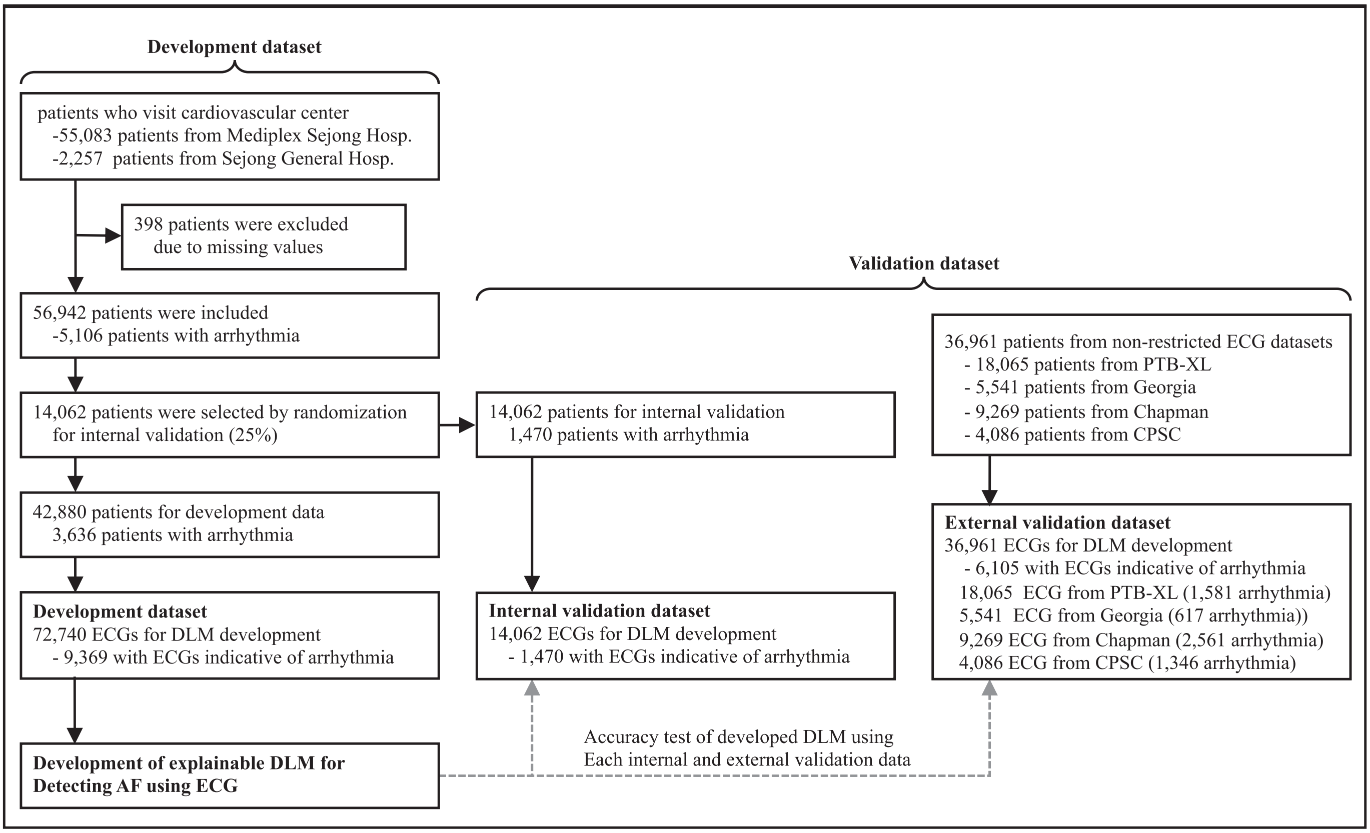

Development dataset
patients who visit cardiovascular center
-55,083 patients from Mediplex Sejong Hosp.
-2,257 patients from Sejong General Hosp.
398 patients were excluded due to missing values
56,942 patients were included
-5,106 patients with arrhythmia
14,062 patients were selected by randomization for internal validation (25%)
42,880 patients for development data
3,636 patients with arrhythmia
Development dataset
72,740 ECGs for DLM development
- 9,369 with ECGs indicative of arrhythmia
Development of explainable DLM for Detecting AF using ECG
Validation dataset
14,062 patients for internal validation
1,470 patients with arrhythmia
Internal validation dataset
14,062 ECGs for DLM development
- 1,470 with ECGs indicative of arrhythmia
36,961 patients from non-restricted ECG datasets
- 18,065 patients from PTB-XL
- 5,541 patients from Georgia
- 9,269 patients from Chapman
- 4,086 patients from CPSC
External validation dataset
36,961 ECGs for DLM development
- 6,105 with ECGs indicative of arrhythmia
18,065 ECG from PTB-XL (1,581 arrhythmia)
5,541 ECG from Georgia (617 arrhythmia))
9,269 ECG from Chapman (2,561 arrhythmia)
4,086 ECG from CPSC (1,346 arrhythmia)
Accuracy test of developed DLM using
Each internal and external validation data

Fig. 1. Study flowchart.
DLM: Deep learning model, ECG: Electrocardiography.

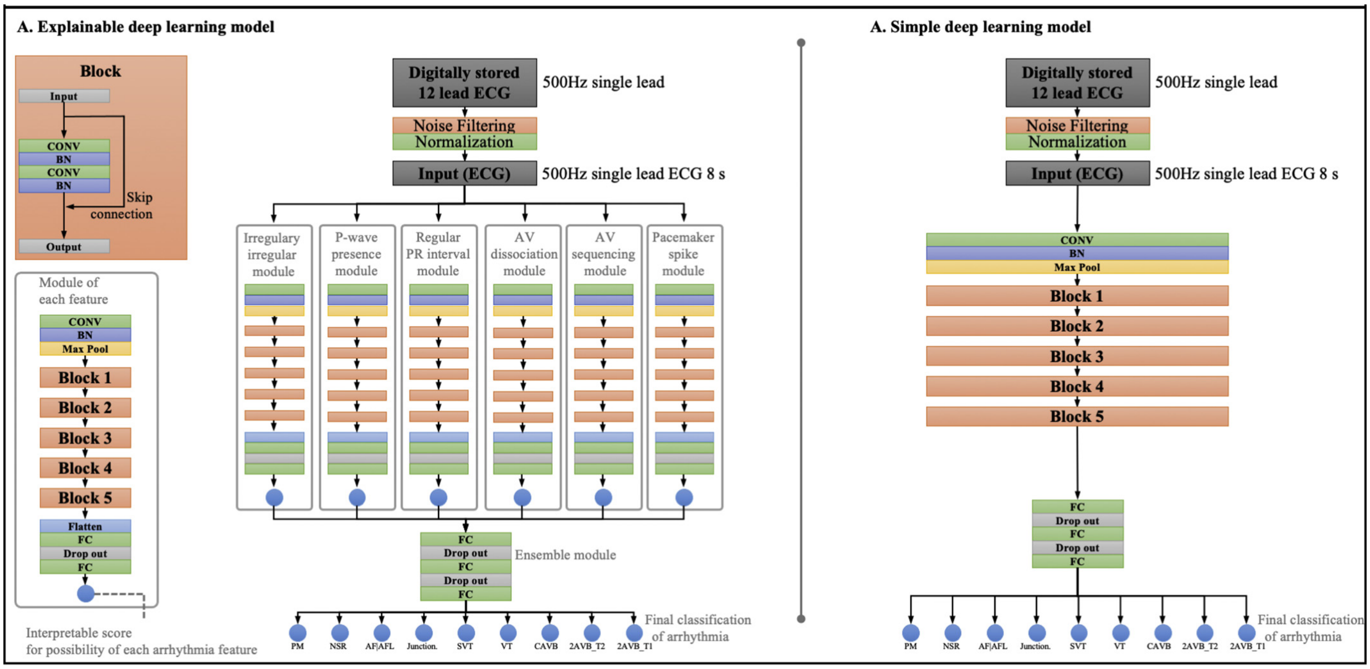


Fig. 2. Architecture of the explainable deep learning model (XDM) for classifying arrhythmia.
AF: Atrial fibrillation, AFL: Atrial flutter, AV: Atrioventricular, BN: Batch normalization layer, CAVB: Complete atrioventricular block, CONV: Convolutional neural network layer, ECG: Electrocardiography, FC: Fully connected layer, JR: Junctional rhythm, NSR: Normal sinus rhythm, PM: Pacemaker rhythm, SDM: Simple multi-classification deep learning model, SVT: Supraventricular tachycardia, VT: Ventricular tachycardia, XDM: Explainable deep learning model, 2AVB_T2: Second degree atrioventricular block Mobitz type II, 2AVB_T1: Second degree atrioventricular block Mobitz type I with Wenckebach phenomenon.

(JR), supraventricular tachycardia (SVT), ventricular tachycardia (VT), complete atrioventricular block (CAVB), second degree atrioventricular block Mobitz type II (2AVB-T2), second degree atrioventricular block Mobitz type I with Wenckebach phenomenon (2AVB-T1), and pacemaker rhythm (PM). Three cardiologists re-labeled each ECG in the external validation datasets. Specifically, the cardiologists labeled the Sejong ECG datasets based on medical records that included progression notes and electrophysiological study reports.

Development of an XDM for arrhythmia classification

To develop an XDM, we developed modules to classify the characteristics of arrhythmia, as opposed to detecting the presence of each possible arrhythmia; we called this method an NBET. We developed six deep learning modules for the features and final ensemble of the XDM using seven labels of each ECG based on supervised learning as shown in Fig. 2. To this end, cardiologists labeled each ECG, not only for the

Table 1
Baseline characteristics.

Characteristics	NSR	Arrhythmia	p-value
n	51,836	5106	
Age, years, mean (SD)	51.92 (18.08)	67.79 (14.22)	<0.001
Male, n, (%)	25,139 (48.5)	2660 (52.1)	<0.001
Heart rate, bpm, mean (SD)	72.34 (17.93)	94.43 (42.78)	<0.001
PR interval, ms, mean (SD)	164.57 (28.21)	176.67 (85.79)	<0.001
QRS duration, ms, mean (SD)	94.14 (14.14)	104.89 (26.68)	<0.001
QT interval, ms, mean (SD)	397.67 (40.15)	396.47 (84.24)	0.077
QTc, ms, mean (SD)	429.90 (30.64)	464.16 (47.75)	<0.001
P axis, mean (SD)	44.50 (27.26)	36.31 (62.29)	<0.001
R axis, mean (SD)	42.11 (39.06)	41.21 (57.00)	0.134
T axis, mean (SD)	39.63 (34.12)	55.80 (81.10)	<0.001
Rhythm (%)			<0.001
NSR	51,836 (100.0)	0 (0.0)	
PM	0 (0.0)	288 (5.6)	
AF or AFL	0 (0.0)	3883 (76.0)	
JR	0 (0.0)	193 (3.8)	
SVT	0 (0.0)	356 (7.0)	
VT	0 (0.0)	45 (0.9)	
CAVB	0 (0.0)	189 (3.7)	
2AVB_T2	0 (0.0)	82 (1.6)	
2AVB_T1	0 (0.0)	70 (1.4)	

AF: Atrial fibrillation, AFL: Atrial flutter, CAVB: Complete atrioventricular block, JR: Junctional rhythm, NSR: Normal sinus rhythm, PM: Pacemaker rhythm, SDM: Simple multi-classification deep learning model, SVT: Supraventricular tachycardia, VT: Ventricular tachycardia, XDM: Explainable deep learning model, 2AVB_T2: Second degree atrioventricular block Mobitz type II, 2AVB_T1: Second degree atrioventricular block Mobitz type I with Wenckebach phenomenon.

Table 2
Performance of XDM and SDM for classifying arrhythmia.

	XDM			SDM		
Arrhythmia	Precision	Recall	F1 score	Precision	Recall	F1 score
Internal validation						
NSR	0.989	0.999	0.994	0.977	0.996	0.987
AF or AFL	0.961	0.966	0.963	0.937	0.922	0.930
JR	0.929	0.718	0.810	0.859	0.459	0.598
SVT	0.965	0.675	0.794	0.782	0.617	0.689
VT	0.842	0.889	0.865	0.632	0.545	0.585
CAVB	0.887	0.846	0.866	0.758	0.443	0.560
2AVB_T2	0.966	0.700	0.812	0.655	0.380	0.481
2AVB_T1	0.923	0.558	0.696	0.577	0.714	0.638
PM	0.959	0.785	0.864	0.851	0.606	0.708
Aggregate accuracy	0.936	0.793	0.852	0.781	0.632	0.686
External validation						
NSR	0.983	0.997	0.990	0.959	0.995	0.977
AF or AFL	0.961	0.949	0.955	0.938	0.882	0.909
SVT	0.928	0.668	0.777	0.919	0.616	0.737
CAVB	0.857	0.800	0.828	0.857	0.161	0.271
PM	0.839	0.559	0.671	0.836	0.358	0.501
Aggregate accuracy	0.914	0.795	0.844	0.902	0.602	0.679

AF: Atrial fibrillation, AFL: Atrial flutter, CAVB: Complete atrioventricular block, JR: Junctional rhythm, NSR: Normal sinus rhythm, PM: Pacemaker rhythm, SDM: Simple multi-classification deep learning model, SVT: Supraventricular tachycardia, VT: Ventricular tachycardia, XDM: Explainable deep learning model, 2AVB_T2 second degree atrioventricular block Mobitz type II, and 2AVB_T1 second degree atrioventricular block Mobitz type I with Wenckebach phenomenon.

classification of arrhythmia, but also for the presence of six features. Cardiologists binarily labeled the ground truths of the features to each ECG. Cardiologist labeled a ground truth of irregularity feature as 1 when the ECG was irregulary irregular, indicating the absence of a regular pattern in an R wave. Irregulary irregular is a character of atrial fibrillation and it means that regularity is never observed in RR intervals. For example, recurrent trigeminy has irregular rhythm but have repeat pattern of RR interval. However atrial fibrillation has no pattern of RR interval at all. Similarly, Cardiologists labeled the ground truth of Regularity PR interval, atrioventricular dissociation, and atrioventricular sequencing as 1 when they observed a regular PR interval in 10 s ECG, no correlation between P wave and R wave, and 1:1 matching between P wave and R wave. First, we developed each module to determine the features of heart rhythm, which were defined as irregularity, presence of P wave, regularity of PR interval, atrioventricular dissociation, atrioventricular sequencing, and pacemaker spike presence. Each module was developed using five residual blocks of the neural network to learn complex hierarchical non-linear representations from the data. In a residual block with four stages, two convolution layers and two batch normalization layers were repeated. We used five residual blocks and two fully connected 1-dimensional (1D) layers to develop each feature model. The second fully connected 1D layer of each module was connected to the output node, which comprised one node. The corresponding values of the output node for six modules represent the probability for each feature of arrhythmia. The corresponding values are described as interpretable scores in Fig. 2. A SoftMax function was used at the output node of each module as an activation function because the output of the SoftMax function ranges between 0 and 1. Finally, we concatenated

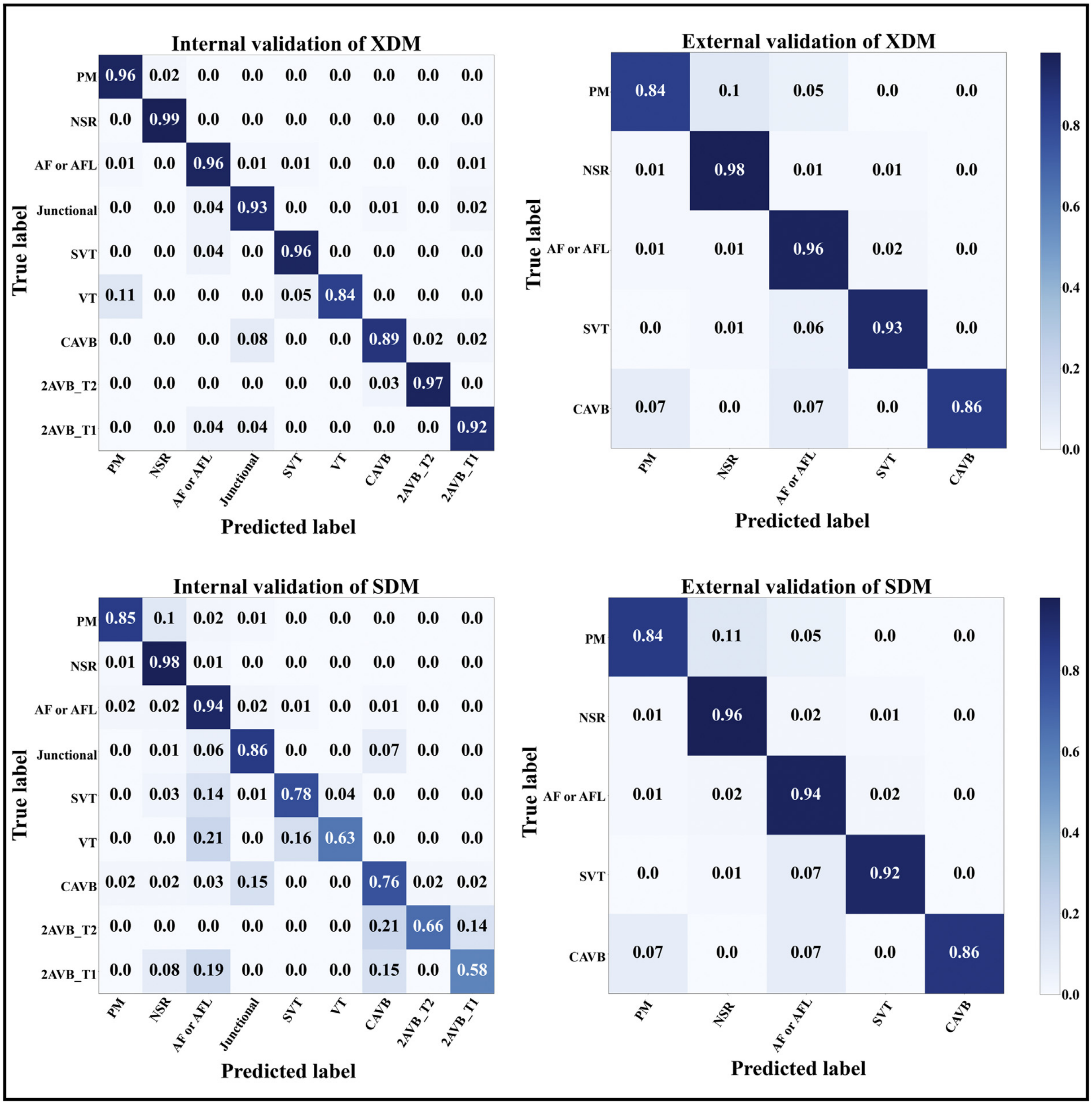
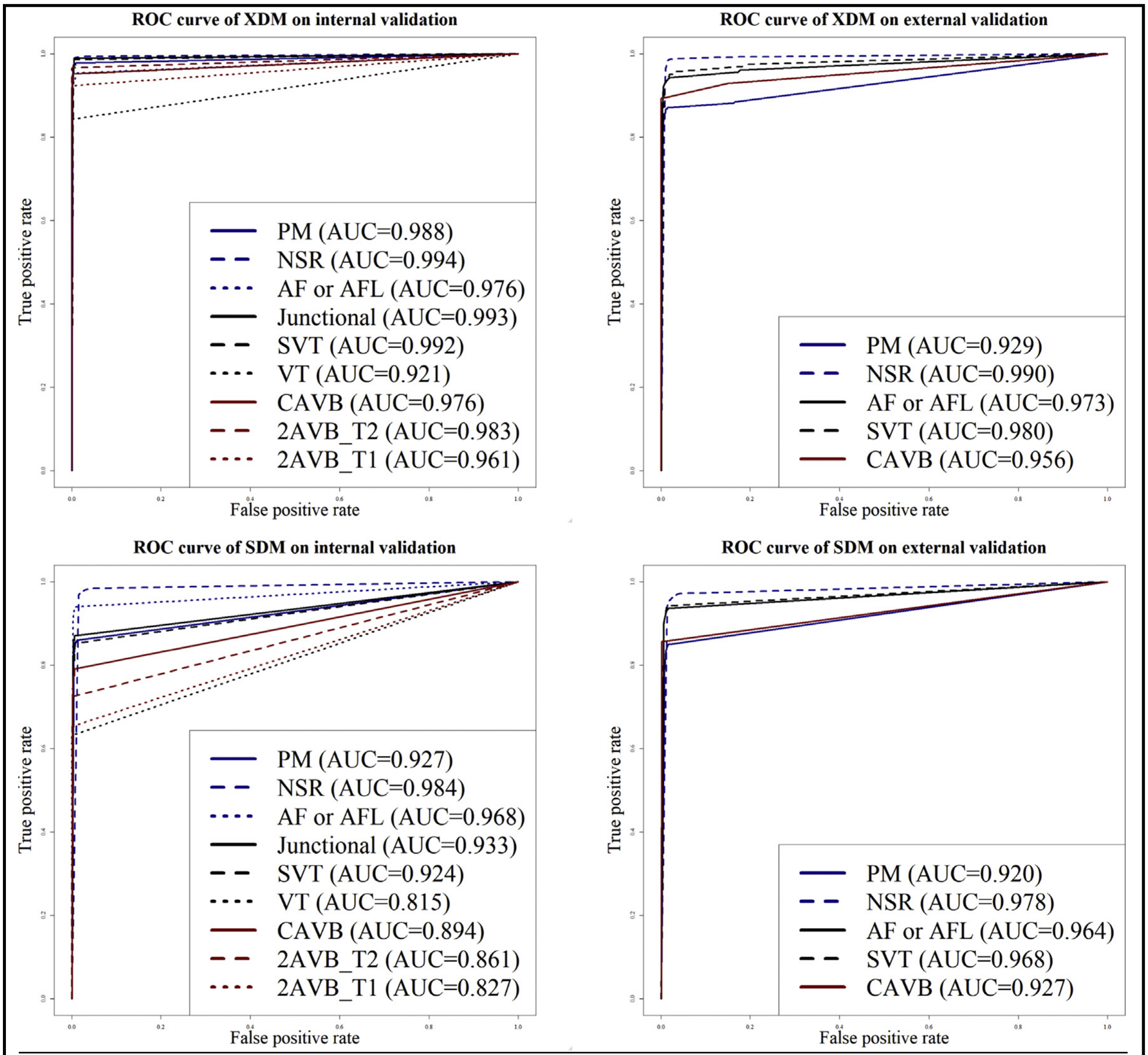


Fig. 3. Confusion matrixes for the explainable deep learning model (XDM) and simple multi-classification deep learning model (SDM) prediction on internal and external validation datasets.

AF: Atrial fibrillation, AFL: Atrial flutter, AV: Atrioventricular, CAVB: Complete atrioventricular block, JR: Junctional rhythm, NSR: Normal sinus rhythm, PM: Pacemaker rhythm, SVT: Supraventricular tachycardia, VT: Ventricular tachycardia, 2AVB_T2: Second degree atrioventricular block Mobitz type II, 2AVB_T1: Second degree atrioventricular block Mobitz type I with Wenckebach phenomenon.



Internal validation dataset

Performance of XDM	AUC (95%CI)	SEN (95%CI)	SPE (95%CI)	PPV (95%CI)	NPV (95%CI)
PM	0.988 (0.978-0.997)	0.973 (0.951-0.994)	0.996 (0.995-0.997)	0.749 (0.699-0.799)	1.000 (0.999-1.000)
NSR	0.994 (0.993-0.996)	0.990 (0.988-0.991)	0.994 (0.991-0.997)	0.999 (0.999-1.000)	0.937 (0.928-0.947)
AF or AFL	0.976 (0.971-0.981)	0.953 (0.943-0.963)	0.988 (0.986-0.989)	0.894 (0.880-0.908)	0.995 (0.994-0.996)
Junctional	0.993 (0.982-1.000)	0.988 (0.965-1.000)	0.994 (0.993-0.995)	0.459 (0.387-0.531)	1.000 (1.000-1.000)
SVT	0.992 (0.982-1.000)	0.986 (0.967-1.000)	0.995 (0.994-0.996)	0.614 (0.551-0.677)	1.000 (1.000-1.000)
VT	0.921 (0.837-1.000)	0.842 (0.678-1.000)	1.000 (1.000-1.000)	0.889 (0.744-1.034)	1.000 (1.000-1.000)
CAVB	0.976 (0.949-1.000)	0.952 (0.898-1.000)	0.997 (0.997-0.998)	0.567 (0.472-0.663)	1.000 (1.000-1.000)
2AVB_T2	0.983 (0.949-1.000)	0.966 (0.899-1.000)	0.999 (0.999-1.000)	0.718 (0.577-0.859)	1.000 (1.000-1.000)
2AVB_T1	0.961 (0.909-1.000)	0.923 (0.821-1.00)	0.999 (0.998-0.999)	0.558 (0.410-0.707)	1.000 (1.000-1.000)

Performance of SDM	AUC (95%CI)	SEN (95%CI)	SPE (95%CI)	PPV (95%CI)	NPV (95%CI)
PM	0.927 (0.904-0.950)	0.855 (0.809-0.902)	0.993 (0.992-0.994)	0.610 (0.555-0.664)	0.998 (0.998-0.999)
NSR	0.984 (0.981-0.986)	0.976 (0.974-0.978)	0.980 (0.975-0.986)	0.997 (0.996-0.998)	0.862 (0.849-0.875)
AF or AFL	0.968 (0.962-0.974)	0.939 (0.928-0.951)	0.992 (0.990-0.993)	0.925 (0.912-0.937)	0.993 (0.992-0.995)
Junctional	0.933 (0.897-0.969)	0.871 (0.799-0.942)	0.994 (0.992-0.995)	0.396 (0.326-0.466)	0.999 (0.999-1.000)
SVT	0.924 (0.894-0.953)	0.852 (0.794-0.911)	0.992 (0.991-0.993)	0.467 (0.406-0.528)	0.999 (0.998-0.999)
VT	0.815 (0.704-0.927)	0.632 (0.415-0.848)	0.999 (0.999-1.000)	0.545 (0.337-0.754)	1.000 (0.999-1.000)
CAVB	0.894 (0.842-0.945)	0.790 (0.689-0.892)	0.994 (0.993-0.995)	0.322 (0.248-0.397)	0.999 (0.999-1.000)
2AVB_T2	0.861 (0.778-0.944)	0.724 (0.561-0.887)	0.998 (0.997-0.999)	0.362 (0.238-0.486)	1.000 (0.999-1.000)
2AVB_T1	0.827 (0.733-0.920)	0.654 (0.471-0.837)	0.999 (0.998-0.999)	0.415 (0.264-0.565)	0.999 (0.999-1.000)

External validation dataset

Performance of XDM	AUC (95%CI)	SEN (95%CI)	SPE (95%CI)	PPV (95%CI)	NPV (95%CI)
PM	0.929 (0.908-0.951)	0.871 (0.832-0.910)	0.986 (0.985-0.987)	0.323 (0.290-0.356)	0.999 (0.999-0.999)
NSR	0.990 (0.989-0.992)	0.983 (0.982-0.985)	0.986 (0.983-0.989)	0.997 (0.997-0.998)	0.920 (0.913-0.927)
AF or AFL	0.973 (0.970-0.976)	0.940 (0.934-0.947)	0.983 (0.981-0.984)	0.898 (0.890-0.906)	0.990 (0.989-0.991)
SVT	0.980 (0.973-0.987)	0.951 (0.934-0.968)	0.981 (0.980-0.983)	0.465 (0.438-0.493)	0.999 (0.999-0.999)
CAVB	0.956 (0.900-1.000)	0.893 (0.778-1.007)	1.000 (1.000-1.000)	0.806 (0.667-0.946)	1.000 (1.000-1.000)

Performance of SDM	AUC (95%CI)	SEN (95%CI)	SPE (95%CI)	PPV (95%CI)	NPV (95%CI)
PM	0.920 (0.899-0.941)	0.850 (0.808-0.891)	0.985 (0.983-0.986)	0.303 (0.271-0.334)	0.999 (0.998-0.999)
NSR	0.978 (0.976-0.980)	0.960 (0.958-0.963)	0.977 (0.973-0.981)	0.995 (0.994-0.996)	0.829 (0.821-0.838)
AF or AFL	0.964 (0.961-0.967)	0.933 (0.927-0.940)	0.986 (0.985-0.987)	0.914 (0.907-0.922)	0.989 (0.988-0.990)
SVT	0.968 (0.958-0.977)	0.941 (0.923-0.959)	0.987 (0.986-0.988)	0.559 (0.529-0.589)	0.999 (0.999-0.999)
CAVB	0.927 (0.861-0.994)	0.857 (0.728-0.987)	0.998 (0.998-0.998)	0.242 (0.158-0.327)	1.000 (1.000-1.000)

Fig. 4. Performances of the XDM and SDM on internal and external validation datasets.
AF: Atrial fibrillation, AFL: Atrial flutter, AUC: Area under the receiver operating characteristic curve, AV: Atrioventricular, CAVB: Complete atrioventricular block, CI: Confidence interval, JR: Junctional rhythm, NSR: Normal sinus rhythm, NPV: Negative predictive value, PM: Pacemaker rhythm, PPV: Positive predictive value, ROC: Receiver operating characteristics curve, SDM: simple multi-classification deep learning model, SEN: Sensitivity, SPE: Specificity, SVT: Supraventricular tachycardia, VT: Ventricular tachycardia, XDM: explainable deep learning model, 2AVB_T2: Second degree atrioventricular block Mobitz type II, 2AVB_T1: Second degree atrioventricular block Mobitz type I with Wenckebach phenomenon.

six feature modules using a multi-layer perceptron architecture to produce the final arrhythmia classification. The multi-layer perceptron included three fully connected 1D layer and two dropout layers. The third fully connected 1D layer of multi-layer perceptron was connected to the final nine output nodes. A SoftMax function was used at the nine output nodes, and the nine output nodes represented the probability for NSR, AF or AFL, JR, SVT, VT, CAVB, 2AVB-T2, 2AVB-T1, and PM. Given that we could evaluate the output values of each module, we could determine the underlying reasons for the final decision of the XDM.

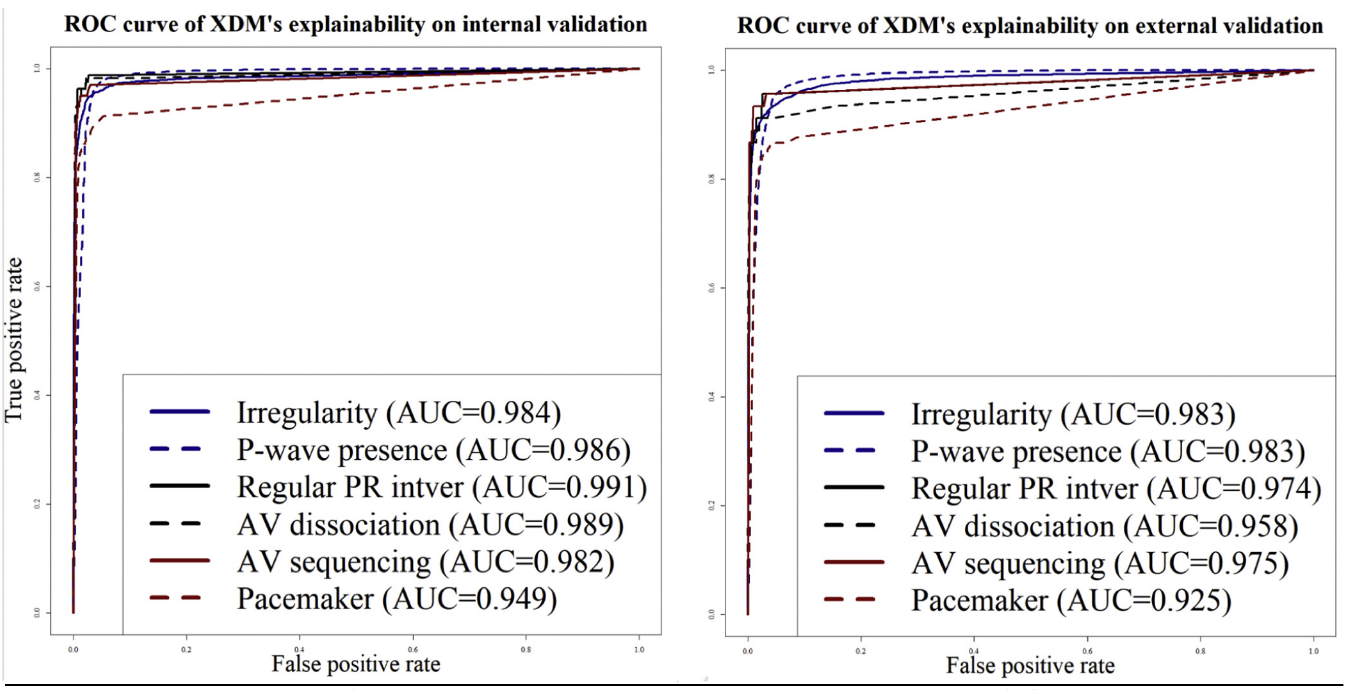
As a comparative method, we developed a simple multi-classification deep learning model (SDM) which had five residual blocks and three fully connected layers. The SDM is a conventional method that has been used in previous studies to detect arrhythmias via ECG. The architecture of the XDM and SDM were confirmed by a grid search.

Statistical analysis

Continuous variables are presented as mean values (applying standard deviation [SD]) and compared using the unpaired Student's *t*-test or Mann-Whitney *U* test. Categorical variables are expressed as frequencies and percentages, and were compared using the χ^2 test.

We confirmed the overlap between the prediction of the XDM and the ground-truth label confirmed by cardiologists using a normalized confusion matrix plot. For each ECG, the XDM produced a final prediction result that was compared against the ground truth of the ECG. We confirmed the F1 score, precision, and recall on multi-classification. We aimed to evaluate the detection performance for each arrhythmia using the area under the receiver operating characteristic curve (AUC). The receiver operating characteristic curve was created by plotting the true positive rate against the false positive rate. The XDM output the probability for each arrhythmia, which was compared against the corresponding ground-truth label to obtain the AUC. We applied the cutoff point to the validation data to calculate the sensitivity, specificity, negative predictive value, and positive predictive value. The sensitivity, specificity, PPV, and NPV were confirmed at the operating point from Youden J statistics in the development data [17], and the performance of the SDM was confirmed in the same manner. We then compared the performance of the SDM with that of the XDM.

We verified the explainability of the DLM through further analyses. To verify the performance of each feature module, we compared the module-calculated probability with the ground-truth feature information provided by cardiologists. Exact 95% confidence intervals (CIs) were used for all of the metrics of diagnostic performance, except for the AUC. The CIs of the AUCs were determined according to Sun and Su's optimization of the De-long method using the pROC package by R (The R Foundation for Statistical Computing, Vienna, Austria).



	Internal validation dataset					External validation dataset				
Feature	AUC (95%CI)	SEN (95%CI)	SPE (95%CI)	PPV (95%CI)	NPV (95%CI)	AUC (95%CI)	SEN (95%CI)	SPE (95%CI)	PPV (95%CI)	NPV (95%CI)
Irregularity	0.984 (0.981-0.988)	0.947 (0.936-0.957)	0.974 (0.972-0.977)	0.796 (0.779-0.814)	0.994 (0.993-0.995)	0.983 (0.981-0.985)	0.930 (0.923-0.936)	0.961 (0.959-0.963)	0.798 (0.788-0.807)	0.988 (0.987-0.989)
P-wave presence	0.986 (0.983-0.990)	0.975 (0.973-0.978)	0.953 (0.943-0.962)	0.994 (0.993-0.996)	0.820 (0.804-0.836)	0.983 (0.981-0.985)	0.957 (0.954-0.959)	0.950 (0.945-0.955)	0.990 (0.989-0.991)	0.805 (0.796-0.814)
Regular PR interval	0.991 (0.978-1.000)	0.988 (0.963-1.000)	0.973 (0.971-0.976)	0.157 (0.125-0.188)	1.000 (1.000-1.000)	0.974 (0.942-1.000)	0.956 (0.895-1.000)	0.974 (0.972-0.976)	0.048 (0.034-0.062)	1.000 (1.000-1.000)
AV dissociation	0.989 (0.970-1.000)	0.983 (0.949-1.000)	0.977 (0.975-0.979)	0.136 (0.103-0.168)	1.000 (1.000-1.000)	0.958 (0.916-1.000)	0.911 (0.828-0.994)	0.979 (0.978-0.981)	0.056 (0.039-0.073)	1.000 (1.000-1.000)
AV sequencing	0.982 (0.964-0.999)	0.970 (0.937-1.000)	0.971 (0.968-0.973)	0.174 (0.142-0.205)	1.000 (1.000-1.000)	0.975 (0.943-1.000)	0.956 (0.895-1.000)	0.968 (0.966-0.970)	0.039 (0.028-0.050)	1.000 (1.000-1.000)
Pacemaker	0.949 (0.929-0.968)	0.912 (0.874-0.950)	0.947 (0.944-0.951)	0.172 (0.150-0.194)	0.999 (0.998-0.999)	0.925 (0.906-0.944)	0.860 (0.822-0.898)	0.962 (0.960-0.964)	0.156 (0.139-0.173)	0.999 (0.998-0.999)

Fig. 5. Performance of the XDM on the internal and external validation datasets.
AUC: Area under the receiver operating characteristic curve, AV: Atrioventricular, SEN: Sensitivity, SPE: Specificity, NPV: Negative predictive value, PPV: Positive predictive value. XDM explainable deep learning model.

A significant difference in patient characteristics was defined as a two-sided p-value <0.001. Statistical analyses were computed using R software, version 3.4.2. In addition, we used PyTorch's open-source software library at the backend and Python (version 3.6.11) for the analyses.

Visualizing the developed XDM for interpretation

To understand the model and compare it with existing medical knowledge, it was important to identify which regions had a significant effect on the decision made by the XDM. To this end, we employed a sensitivity map using a saliency method. The map was computed using the first-order gradients of the classifier probabilities with respect to the input signals. If the probability of a classifier was sensitive to a specific region of the signal, the region was considered significant in the model [18,19]. We used a gradient-class activation map as a sensitivity map and a guided gradient backpropagation method. The XDM showed the sensitivity map from each feature module.

Results

The Sejong ECG dataset for development and internal validation dataset included patients who visited MSH (March 1, 2017 to March 31, 2020) and SGH (October 1, 2019 to December 31, 2019). A total of 55,083 patients at MSH and 2257 patients at SGH were eligible for inclusion. We excluded 387 patients at MSH and 11 patients at SGH because of missing values, as shown in Fig. 1. The development dataset from the Sejong ECG dataset included 72,740 ECGs of 42,880 patients (Table 1). The performance of the algorithm was confirmed using 14,062 ECGs from 14,062 patients in the internal validation dataset from the Sejong ECG dataset. The DLM performance was externally validated using 9269, 4086, 5541, and 18,065 ECGs from the Chapman, CPSC, Georgia, and PTB-XL ECG datasets, respectively.

Table 2 shows the performance of the XDM on the internal and external validation datasets for multi-variable classification. The XDM's F1 score of NSR, AF or AFL, JR, SVT, VT, 2AVB-T1, 2AVB-T2, CAVB, and PM, was 0.989, 0.961, 0.929, 0.965, 0.842, 0.887, 0.966, 0.923, and 0.959 on the internal validation dataset, respectively. The XDM's F1 score of NSR, AF or AFL, SVT, CAVB, and PM, was 0.990, 0.955, 0.777, 0.828, and 0.671 on the external validation dataset, respectively. Fig. 3 shows a confusion matrix of the XDM and SDM on the internal and external validation datasets. The AUC of the internal and external validation datasets for detecting each arrhythmia is shown in Fig. 4. Moreover, the sensitivity, specificity, PPV, and NPV were confirmed at the operating point from Youden J statistics using the development data. The XDM outperformed the SDM in all measures.

As shown in Fig. 5, the AUC of each irregularity, P-wave presence, regular PR interval, atrioventricular dissociation, atrioventricular sequencing, and pacemaker spike presence was 0.984, 0.986, 0.991, 0.989, 0.982, and 0.949, respectively. To calculate the performance, we compared the interpretable scores of each output node of each feature module ground truth of each module that was labeled by cardiologists. We employed a sensitivity map to visualize the ECG region to detect each ECG feature as shown in Fig. 6. The map reveals that the XDM focused on the part of the ECG related to each module. For example, the module for determining the presence of P-waves focused on P-waves, whereas the module that made decisions on irregularity focused on QRS complexes. Furthermore, the module that was used to determine

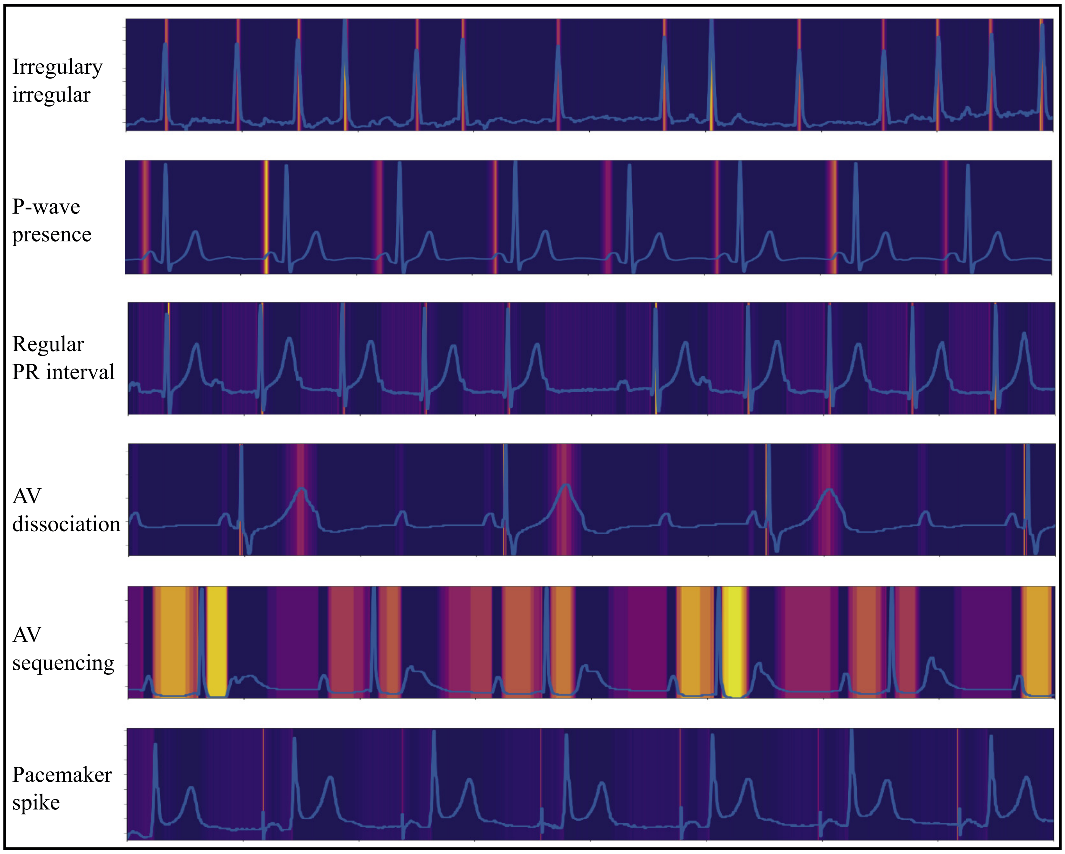


Fig. 6. Sensitivity map of the XDM for detecting each arrhythmia feature.
The sensitivity map shows the region in which the XDM module focused attention for deciding the presence of features. The most important region is in orange and the least important region is in blue. AV: Atrioventricular, XDM: explainable deep learning model.

the presence of a Regular PR interval focused on the PR segment of each beat. The module for determining the presence of AV dissociation focused on peak of the P-, R-, and T-waves. The module that was used to determine the presence of AV sequencing focused on the PR and ST segments, and the module that was used to determine the presence of a pacemaker spike focused on the pacemaker spike signal.

Discussion

Although several previous studies applied deep learning algorithms to diagnose arrhythmia using ECGs, such algorithms were still black boxes; in other words, we neither understood their decision nor knew the reasons for a particular diagnosis of arrhythmia. Our study group recently adopted the saliency map in ECGs to achieve explainability, although the method did not completely explain how the model made conclusions [21–23]. The saliency map only highlighted the part of the ECG that was important to the decision but did not explain the exact reason for the meaning of the part [18]. For example, when a deep learning algorithm focused on the QRS complex to diagnose a disease in a saliency map, we were unable to determine which particular factor of the QRS complex that the diagnosis was based on.

To overcome these limitations of DLMs, we adopted state-of-the-art explainable AI technologies, i.e., an NBET, in our ECG research. Our key insight was to combine neural networks with decision trees, preserving high-level interpretability while using neural networks for low-level decisions. These NBET models have accuracy that is matched to that of neural networks, while also preserving the interpretability of a decision tree. In this study, we developed six modules for features based on deep learning. Because of this, we not only classified arrhythmia, but also elucidated the underlying reasons for the classification result. As shown in Supplemental material, we described the correlation between features and arrhythmias. Atrial fibrillation and flutter were strongly correlated with features such as presence of P-wave and irregularity, and CAVB exhibited strong correlation with AV dissociation. However, we were unable to elucidate the exact meaning of these correlations because we could not the exploration the process of deep learning. In our next study, we hope to reveal the exact decision process of deep learning architecture. For example, if XLM decided that a normal ECG demonstrated AF, we could determine the reason for the decision; such reasons may include “XLM decided that the P wave was absent (XLM could not find the P wave on input ECG)” or “XLM decided that the rhythm was irregular.” This explainability is vital in determining and editing the error in the model. Doctors could also determine errors if their decisions did not match that of the XLM. For example, if a doctor could not find a small P wave and decided an ECG as AF, XLM could assist the doctor because the XLM could determine the P-wave in the ECG based on the value of the P-wave module and display the focal P-wave with a sensitivity map. In this study, we preserved the accuracy of XDMs by adopting explainability, and the XDM was found to outperform SDM.

There are several limitations to the present study. First, we developed six feature modules to develop XDM. Although we selected ten features based on current medical knowledge, it is possible to enhance the XDM performance using other features of ECG. This is the next research area of our study group. Second, studies related to the clinical significance of the new technology are required for application in clinical practice. In our next study, we will verify the performance and significance of XDM using a prospective study in daily clinical practice.

Conclusion

We developed an XDM for arrhythmia classification and confirmed that the model accurately classifies arrhythmia in diverse formats of ECGs using external validation datasets. The results indicate that the proposed XAI methodology could be used to describe the reasons for the decision made by the XDM in arrhythmia classification with high performance.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1F1A1073791).

Affiliations

JK, KHK, KHJ, SYL, JP, and BHO (Mediplex Sejong Hospital); JK, YYJ, MSJ, YJL, YHC, and JHS (Medical AI Co. Ltd.); JK and JHB (Bodyfriend Co. Ltd.).

Data availability statement

The data used in this study will be shared upon reasonable request to the corresponding author.

Declaration of Competing Interest

KHJ, KHK, SYL, JP, and BHO declare that they have no competing interests. YYJ, JK, YHC, JHS, YJL, and MSJ are researchers of Medical AI Co., a medical artificial intelligence company. JK and JHB are researchers of Body friend Co. There are no products in development or marketed products to declare. This does not alter our adherence to Journal policies.

Acknowledgement

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jelectrocard.2021.06.006.

References

[1] Khurshid S, Choi SH, Weng L-C, Wang EY, Trinquart L, Benjamin EJ, et al. Frequency of cardiac rhythm abnormalities in a half million adults. Circ Arrhythmia Electrophysiol. 2018;11.

[2] Benjamin EJ, Blaha MJ, Chiuve SE, Cushman M, Das SR, Deo R, et al. Heart disease and stroke statistics-2017 update: a report from the American Heart Association. Circulation. 2017;135:e146–603.

[3] Go AS, Hylek EM, Phillips KA, Chang Y, Henault LE, Selby JV, et al. Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the AnTicoagulation and risk factors in atrial fibrillation (ATRIA) study. JAMA. 2001;285:2370–5.

[4] Corley SD, Epstein AE, DiMarco JP, Domanski MJ, Geller N, Greene HL, et al. Relationships between sinus rhythm, treatment, and survival in the Atrial Fibrillation Follow-Up Investigation of Rhythm Management (AFFIRM) study. Circulation. 2004;109:1509–13.

[5] Stewart S, Hart CL, Hole DJ, McMurray JJV. A population-based study of the long-term risks associated with atrial fibrillation: 20-year follow-up of the Renfrew/Paisley study. Am J Med. 2002;113:359–64.

[6] Orejarena LA, Vidaillet H, DeStefano F, Nordstrom DL, Vierkant RA, Smith PN, et al. Paroxysmal supraventricular tachycardia in the general population. J Am Coll Cardiol. 1998;31:150–7.

[7] Mustaqeem A, Anwar SM, Khan AR, Majid M. A statistical analysis based recommender model for heart disease patients. Int J Med Inform. 2017;108:134–45.

[8] Giebel GD, Gissel C. Accuracy of mHealth devices for atrial fibrillation screening: systematic review. JMIR Mhealth Uhealth. 2019;7:e13641.

[9] Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med. 2019;25:65–9.

[10] Ribeiro AH, Ribeiro MH, Paixão GMM, Oliveira DM, Gomes PR, Canazart JA, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat Commun. 2020;11:1760.

[11] van de Leur RR, Blom LJ, Gavves E, Hof IE, van der Heijden JF, Clappers NC, et al. Automatic triage of 12-lead ECGs using deep convolutional neural networks. J Am Heart Assoc. 2020;9:e015138.

[12] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.

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

[14] Perez Alday EA, Gu A, Shah AJ, Robichaux C, Wong A-KI, Liu C, et al. Classification of 12-lead ECGs: the PhysioNet/Computing in cardiology challenge 2020. Physiol Meas. 2021;41(12):124003. https://doi.org/10.1088/1361-6579/abc960.

[15] 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;7:48.

[16] Liu F, Liu C, Zhao L, Zhang X, Wu X, Xu X, et al. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. J Med Imaging Health Inform. 2018;8:1368–73.

[17] Schisterman EF, Perkins NJ, Liu A, Bondell H. Optimal cut-point and its corresponding Youden index to discriminate individuals using pooled blood samples. Epidemiology. 2005;16:73–81.

[18] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision; 2017 p. 1;618–626.

[19] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis. 2020;128:336–59.

[20] Zimetbaum P, Goldman A. Ambulatory arrhythmia monitoring. Circulation. 2010; 122:1629–36.

[21] Kwon J, Cho Y, Jeon K-H, Cho S, Kim K-H, Baek SD, et al. A deep learning algorithm to detect anaemia with ECGs: a retrospective, multicentre study. Lancet Digit Heal. 2020;2:e358–67.

[22] Kwon J, Lee SY, Jeon K, Lee Y, Kim K, Park J, et al. Deep learning–based algorithm for detecting aortic stenosis using electrocardiography. J Am Heart Assoc. 2020;9.

[23] Kwon J-M, Jeon K-H, Kim HM, Kim MJ, Lim SM, Kim K-H, et al. Comparing the performance of artificial intelligence and conventional diagnosis criteria for detecting left ventricular hypertrophy using electrocardiography. Europace. 2020;22(3): 412–9. https://doi.org/10.1093/europace/euz324.