Canadian Journal of Cardiology 37 (2021) 94–104

Clinical Research

Usefulness of Machine Learning-Based Detection and Classification of Cardiac Arrhythmias With 12-Lead Electrocardiograms

Kuan-Cheng Chang, MD, PhD,[a,b] Po-Hsin Hsieh, MS,[c] Mei-Yao Wu, MD, PhD,[d,e] Yu-Chen Wang, MD, PhD,[a,f,g] Jan-Yow Chen, MD, PhD,[a,b] Fuu-Jen Tsai, MD, PhD,[h] Edward S.C. Shih, PhD,[i] Ming-Jing Hwang, PhD,[i] and Tzung-Chi Huang, PhD[c,j,k]

[a] *Division of Cardiovascular Medicine, China Medical University Hospital, Taichung, Taiwan;* [b] *Graduate Institute of Biomedical Sciences, China Medical University, Taichung, Taiwan;* [c] *Department of Biomedical Imaging and Radiologic Science, China Medical University, Taichung, Taiwan;* [d] *School of Postbaccalaureate Chinese Medicine, China Medical University, Taichung, Taiwan;* [e] *Department of Chinese Medicine, China Medical University Hospital, Taichung, Taiwan;* [f] *Division of Cardiovascular Medicine, Asia University Hospital, Taichung, Taiwan;* [g] *Department of Biotechnology, Asia University, Taichung, Taiwan;* [h] *Department of Medical Research, China Medical University Hospital, Taichung, Taiwan;* [i] *Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan;* [j] *Artificial Intelligence Center, China Medical University Hospital, Taichung, Taiwan;* [k] *Department of Bioinformatics and Medical Engineering, Asia University, Taichung, Taiwan*

See editorial by Zhou et al., pages 17–18 of this issue.

ABSTRACT

Background: Deep-learning algorithms to annotate electrocardiograms (ECGs) and classify different types of cardiac arrhythmias with the use of a single-lead ECG input data set have been developed. It remains to be determined whether these algorithms can be generalized to 12-lead ECG-based rhythm classification.

Methods: We used a long short-term memory (LSTM) model to detect 12 heart rhythm classes with the use of 65,932 digital 12-lead ECG signals from 38,899 patients, using annotations obtained by consensus of 3 board-certified electrophysiologists as the criterion standard.

RÉSUMÉ

Contexte : Des algorithmes d'apprentissage profond conçus pour annoter les électrocardiogrammes (ECG) et classifier différents types d'arythmie cardiaque à partir des données d'un ECG à une seule dérivation ont été mis au point. Nous avons tenté de déterminer si ces algorithmes peuvent être généralisés pour obtenir une classification à partir des données d'un ECG à 12 dérivations.

Méthodologie : Nous avons utilisé un modèle LSTM (*Long Short-Term Memory*) pour reconnaître 12 catégories de rythmes cardiaques à partir de 65 932 signaux numériques d'ECG à 12 dérivations obtenus auprès de 38 899 patients; nous avons utilisé les annotations con-

Machine-learning technology has been widely used to assist image interpretation, speech recognition, item matching, and the presentation of relevant results in searching.[1] Deep learning, which is one type of machine learning, involves an artificial neural network of representation-learning methods with multilayer representation. Deep learning is superior to other machine-learning techniques in image recognition and speech recognition.[2] It has been used to develop automatic interpretation for different types of medical images, eg, from mammography,[3] chest X-ray,[4,5] ultrasound,[6] and magnetic resonance imaging.[7] For cardiovascular images, deep learning has been developed to interpret the results of electrocardiography, echocardiography, coronary computed tomography, and single-photon emission computed tomography for the evaluation of myocardial perfusion.[8]

Electrocardiography is an important noninvasive examination that is widely used to detect various heart diseases, including rhythm disorders, conduction abnormalities, and myocardial ischemia or infarction, by physicians across different specialties.[9] A computer-based automatic interpretation system incorporated in the electrocardiography machine has been developed to assist diagnosis; however, the accuracy rate of diagnosis remains limited and requires improvement.[10] Recently, deep-learning algorithms have been created to read electrocardiograms (ECGs) and detect different types of arrhythmias,[11,12] with variable sensitivity and specificity.[13] Hannun et al. used a deep neural network (DNN) to develop a cardiologist-level arrhythmia detection system for diagnosing 12 types of cardiac rhythms.[14] The sensitivity of arrhythmia detection was improved. However, the study was based on single-lead ECG records, which provide limited signals compared with 12-lead ECG records.

Recurrent neural network (RNN) models, which are often used to process data related to sequence changes, have proved

Received for publication November 6, 2019. Accepted February 26, 2020.

Corresponding author: Dr Kuan-Cheng Chang, Division of Cardiovascular Medicine, China Medical University Hospital, 2, Yude Road, Taichung 40447, Taiwan. Tel.: +886-04-22052121, ext. 4665; fax: +886-4-22065593.
E-mail: kuancheng.chang@gmail.com

See page 10 for disclosure information.

https://doi.org/10.1016/j.cjca.2020.02.096

Results: The accuracy of the LSTM model for the classification of each of the 12 heart rhythms was $\geq$ 0.982 (range 0.982-1.0), with an area under the receiver operating characteristic curve of $\geq$ 0.987 (range 0.987-1.0). The precision and recall ranged from 0.692 to 1 and from 0.625 to 1, respectively, with an F_1 score of $\geq$ 0.777 (range 0.777-1.0). The accuracy of the model (0.90) was superior to the mean accuracies of internists (0.55), emergency physicians (0.73), and cardiologists (0.83).

Conclusions: We demonstrated the feasibility and effectiveness of the deep-learning LSTM model for interpreting 12 common heart rhythms according to 12-lead ECG signals. The findings may have clinical relevance for the early diagnosis of cardiac rhythm disorders.

Résultats : L'exactitude du modèle LSTM utilisé pour classifier les rythmes cardiaques de chacune des 12 catégories s'établissait à $\geq$ 0,982 (plage : de 0,982 à 1,0), l'aire sous la courbe caractéristique de la performance du test étant de $\geq$ 0,987 (plage : de 0,987 à 1,0). La précision et le rappel allaient de 0,692 à 1 et de 0,625 à 1, respectivement, le score F_1 s'établissant à $\geq$ 0,777 (plage : de 0,777 à 1,0). L'exactitude du modèle (0,90) était supérieure à l'exactitude moyenne des internistes (0,55), des urgentologues (0,73) et des cardiologues (0,83).

Conclusions : Nous avons démontré la faisabilité et l'efficacité de l'emploi du modèle d'apprentissage profond LSTM pour l'interprétation de 12 rythmes cardiaques courants à partir des signaux d'un ECG à 12 dérivations. Ces résultats pourraient être utiles sur le plan clinique aux fins du diagnostic précoce des arythmies.

to be effective for difficult machine-learning tasks.[15] The limitation of the RNN model is that it struggles to capture the long-term time correlation, because the simple RNN cannot handle the problem of regression or weighting exponential gradient explosion (vanishing-gradient problem). Long short-term memory (LSTM) is a type of RNN. By combining different LSTM models, the problem of the vanishing gradient can be solved.[16] In the present study, we used LSTM to detect 12 types of cardiac arrhythmias recorded in a digital 12-lead ECG. The diagnostic accuracy of the LSTM-derived algorithm was compared with the ECG classification performance of board-certified doctors from different disciplines—cardiologists, emergency physicians, and internists—to evaluate the feasibility of the machine learning–based model for clinical applications.

Methods

Data description

We collected 65,932 12-lead ECG waveform data signals from 38,899 patients, which were annotated by cardiologists and stored in the digital core ECG laboratory of China Medical University Hospital (CMUH) from 2009 to 2018. All of the 12-lead ECGs were recorded by a GE Marquette MAC 5500 or MAC 3500 ECG recorder (GE Medical Systems, Milwaukee, WI, USA). Using a standardized protocol, a 10-second resting 12-lead ECG was recorded at a sampling frequency of 500 Hz and was digitally transmitted and stored in the MUSE system (GE Marquette) at the core ECG laboratory of CMUH for subsequent analyses. To avoid electromagnetic interferences and other potential noises during ECG recording, we followed the manufacturer's recommendations of filtering. In brief, the low-frequency digital filter cutoff was 0.67 Hz or below and the high-frequency digital filter cutoff was no lower than 150 Hz to reduce artifactual distortion or error measurements in adults.[17] The GE ECG machine recorded the 10-second resting 12-lead ECG signals, which can be read in Extensible Markup Language (XML) format and converted into arrays of numeric values (Fig. 1), which were then fed into the LSTM neural network for learning. The raw data in the XML format comprises 12 types of heart rhythms: atrial fibrillation (AFIB), atrial flutter (AFL), atrial premature beat (APB), ventricular bigeminy (BIGEMINY), complete heart block (CHB), ectopic atrial rhythm (EAR), first-degree atrioventricular (AV) block (FRAV), normal sinus rhythm (NSR), paroxysmal supraventricular tachycardia (PSVT), second-degree AV block (SAV), sinus tachycardia (ST), and ventricular premature beat (VPB). Such end-to-end learning was enabled by deep learning methodologies, and there was no need to extract features from the ECG signals to represent each type of heart rhythms.

The study protocol was reviewed and approved by the Research Ethics Committee of China Medical University Hospital (CMUH107-REC2-134 [AR-1]). All research was performed in accordance with relevant guidelines and regulations. The Research Ethics Committee waived the requirement for the investigator to obtain signed consent forms from the subjects owing to the retrospective database research design of the study.

Algorithm development

Because ECGs represent sequence data, we used a bidirectional LSTM model for sequence-sequence learning tasks.[16] Between the input layer and the output layer, our model had 4 layers of bidirectional LSTM, each with 128 neurons (Fig. 2), and after these 4 layers of bidirectional LSTM we appended a pooling layer and a dense layer (Fig. 1). Conceptually, the input layer took in the 12-lead ECG signals in the form of a stream of numeric values stored in an array; the information was then processed in and propagated through the LSTM neurons and all the hidden layers, and finally the output layer took the output from the last hidden layer (the dense layer) and used a softmax activation function to assign a decimal probability to each of the 12 cardiac rhythm classes. The class with the largest probability value was the class predicted by the model for a given 12-lead ECG input. At each LSTM neuron, an input gate, a forget gate, and an output gate determined whether and how much the inputs from the previous neuron and layer should be used to create and update a memory, and whether and how much the memory should be propagated to the next neuron. The

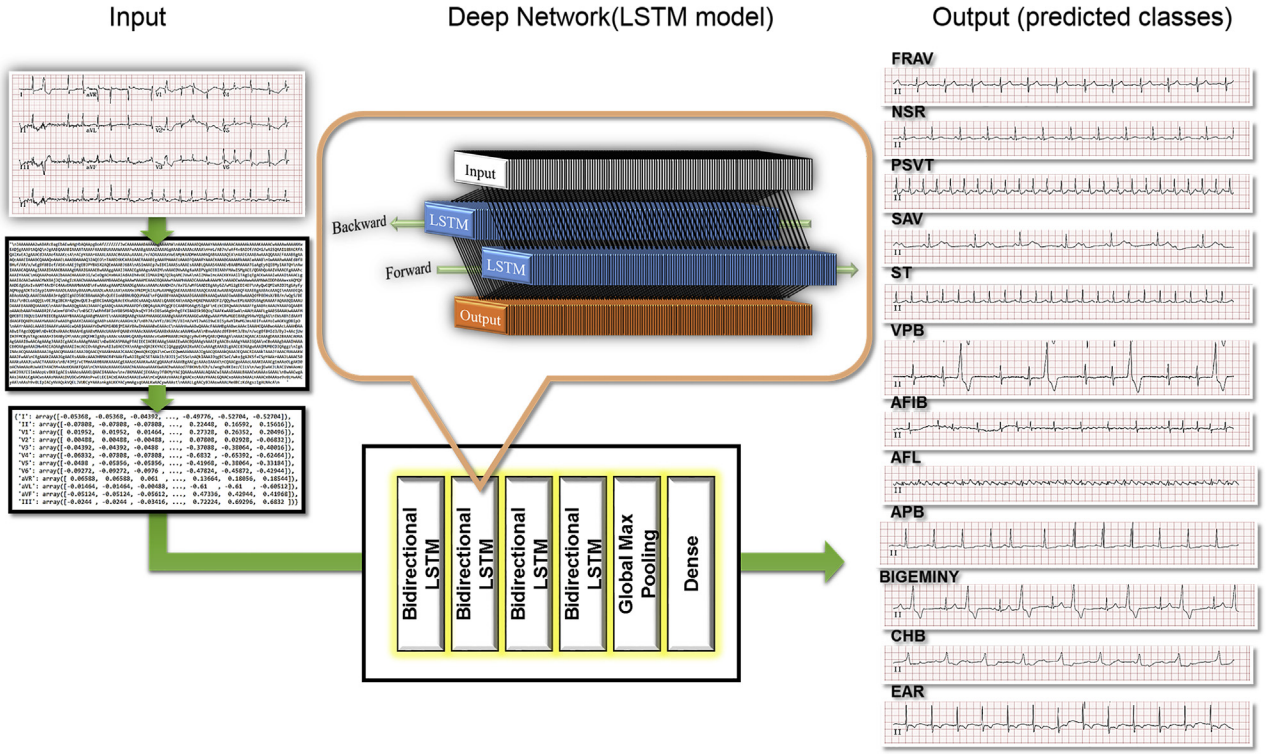

Figure 1. Illustration of max pooling process for the 12-lead electrocardiography (ECG) signals. At the output layer, a softmax activation function was used to generate 12 different outputs, corresponding to the ECG waveform items. AFIB, atrial fibrillation; AFL, atrial flutter; APB, atrial premature beat; BIGEMINY, ventricular bigeminy; CHB, complete heart block; EAR, ectopic atrial rhythm; FRAV, first-degree AV block; LSTM, long short-term memory; NSR, normal sinus rhythm; PSVT, paroxysmal supraventricular tachycardia; SAV, second-degree AV block, type I; ST, sinus tachycardia; VPB, ventricular premature beat.

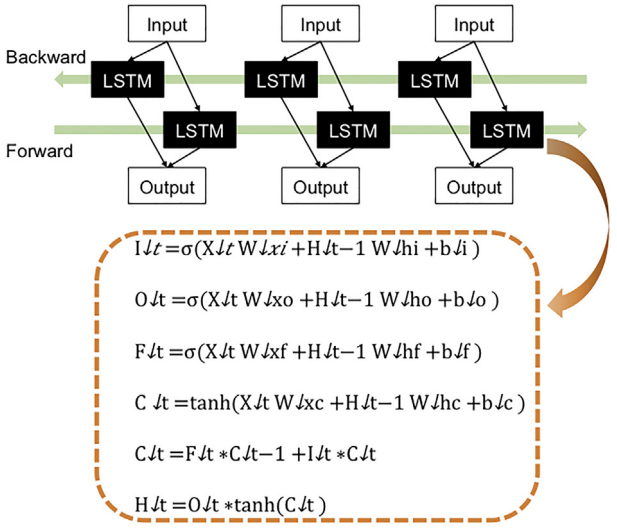

Figure 2. Architecture of the long short-term memory (LSTM) model. A bidirectional 4-layer LSTM model was used. Each layer contained 128 neurons. A single LSTM unit contains input gate, output gate, and forget gate to determine whether the memory is updated, using the equations shown for calculation. I_t, input gate; O_t, output gate; F_t, forget gate; σ, logistic sigmoid function; X_t, input sequence; H_{t-1}, W_{xi}, W_{xo}, W_{xf}, W_{xc}, W_{hi}, the previous block output; W_{ho}, W_{hf}, W_{hc}, weight parameters; b_i, b_o, b_f, b_c, bias parameters; C_t, candidate memory cell similar to the 3 gates but using a tanh activation function; C_{t-1}, the previous LSTM block memory; H_t, the final block output.

pooling layer filtered out less-significant information from the output of the last LSTM layer and reduced the number of parameters; the dense layer was a simple fully connected layer that performed a linear transformation on the output of the pooling layer to produce an information summary on the 12 rhythm classes. In the present work, we separated the data sets of 65,932 12-lead ECG recordings into training and validation sets, with the training set consisting of 90% of the data. By learning from the approximately 50,000 ECG recordings and their labeled arrhythmia classes, that is, the training set, the network was able to determine the parameters that yielded optimal performance. The technical details are shown in the Supplemental Methods.

Details about our LSTM model training, including loss function and initial learning rate, are presented in Supplemental Figure S1.

To test the performance of our LSTM model, we also used 12-lead ECG signals as input data from a public source, the China Physiological Signal Challenges (CPSC) data set, for external validation.

Comparison of performance between model and physicians

A separate data set of 116 12-lead ECGs from 116 patients stored in the same MUSE ECG system with the 12 types of heart rhythms was used for testing. We compared the classification accuracy and the labelling time between the algorithm

Table 1. Data set for training, validation, and testing of the long short-term memory model

No.	Type of heart rhythm	Abbreviation	No. of ECGs for training/validation	No. of ECGs for testing
1	Atrial fibrillation	AFIB	18,077	10
2	Atrial flutter	AFL	10,305	10
3	Atrial premature beat	APB	4416	11
4	Ventricular bigeminy	BIGEMINY	2604	10
5	Complete heart block	CHB	589	8
6	Ectopic atrial rhythm	EAR	430	8
7	First-degree AV block	FRAV	1920	10
8	Normal sinus rhythm	NSR	9019	9
9	Paroxysmal supraventricular tachycardia	PSVT	4168	10
10	Second-degree AV block (type I)	SAV	922	10
11	Sinus tachycardia	ST	9212	10
12	Ventricular premature beat	VPB	4270	10
Total			65,932	116

AV, atrioventricular; ECG, electrocardiogram.

and board-certified physicians from different disciplines—10 cardiologists, 8 emergency physicians, and 10 internists—using a web-based digital ECG system for testing. Annotations performed by a consensus committee of 3 board-certified electrophysiologists were used as the criterion standard for correct classification of each of the cardiac rhythms. The committee members discussed each 12-lead ECG record and reached a labelling consensus. The demographic data and medical history of the patients were obtained from the electronic medical records at CMUH.

A closed-domain online platform was set up for doctors to label the 116 12-lead ECGs' data. The platform page was prepared in advance by staffs before the physicians arrived at the designated room. All participating physicians received a full explanation of the rules and demonstration of the web-based testing, if necessary, before starting the test. The 116 test data records were shown in random order to each of the physicians after they entered their personal identification number and clicked a "Get Data" button. The physicians could actively click any index ECG number to retrieve the first 12-lead ECG recording, accompanied by 12 predetermined options of rhythm classification. The next 12-lead ECG jumped out automatically, also randomly, after clicking a "Submit" button when finishing annotation of the first ECG. The starting time was the instant when the first ECG appeared, and the ending time for the same ECG was the instant when the physician completed the selection and clicked the Submit button to transmit the annotation to the database. The difference between the starting time and the ending time was the length of time required for each of the specific ECGs. The total labelling time was the sum of time for completing the annotation of all 116 ECGs.

All 116 testing 12-lead ECGs were presented via a web-based digital ECG system. Therefore, we could compare the diagnostic accuracy and labeling time between model and physicians immediately after the tests were completed. Two study investigators and a 3-member working staff were present to monitor the test and ensure that the whole process was completed smoothly without protocol deviation.

Statistical analysis

The classification performance of the LSTM model for each of the 12 heart rhythms was assessed according to the receiver operating characteristic (ROC) curve and area under the ROC curve (AUC), accuracy, precision, recall, and F_1 score. The annotations by the electrophysiologist committee were used as the criterion standard. The F_1 score was the harmonic mean of the precision and recall. We used confusion matrices to evaluate the heart-rhythm prediction performance of the LSTM model and the cardiologists, emergency physicians, and internists with respect to the labeling rendered by the committee of electrophysiologists. The differences were assessed by means of a generalized linear model (GLM) with Tukey test to adjust for multiple comparisons.

Results

We used the LSTM model to detect 12 heart-rhythm classes from a 12-lead ECG by with the use of 65,932 digital ECG signals from 38,899 patients at China Medical University Hospital (CMUH). The mean age of the patients was 64.4 ± 19.3 years, and 44% were female. Table 1 presents the numbers of 12-lead ECGs used for training/validation and testing for each of the 12 cardiac rhythms. The diagnostic performance of the LSTM model and the ROC curves for classifying the cardiac rhythm using the model are

Table 2. Diagnostic performance of the long short-term memory model for different heart rhythms

Heart rhythm	Accuracy	AUC	F_1	Precision	Recall
AFIB	0.991	0.998	0.947	1.000	0.900
BIGEMINY	0.982	1.000	0.909	0.833	1.000
SAV	0.991	0.999	0.947	1.000	0.900
EAR	0.974	0.987	0.769	1.000	0.625
AFL	0.982	0.999	0.909	0.833	1.000
CHB	0.991	1.000	0.941	0.880	1.000
NSR	0.965	0.992	0.818	0.692	1.000
FRAV	0.965	0.997	0.777	0.875	0.700
VPB	1.000	1.000	1.000	1.000	1.000
APB	0.982	0.98	0.900	1.000	0.818
ST	0.991	1.000	0.947	1.000	0.900
PSVT	0.991	1.000	0.952	0.909	1.000

AUC, area under the receiver operating characteristic curve; AFIB, atrial fibrillation; AFL, atrial flutter; APB, atrial premature beat; BIGEMINY, ventricular bigeminy; CHB, complete heart block; EAR, ectopic atrial rhythm; FRAV, first-degree AV block; NSR, normal sinus rhythm; PSVT, paroxysmal supraventricular tachycardia; SAV, second-degree AV block (type I); ST, sinus tachycardia; VPB, ventricular premature beat.

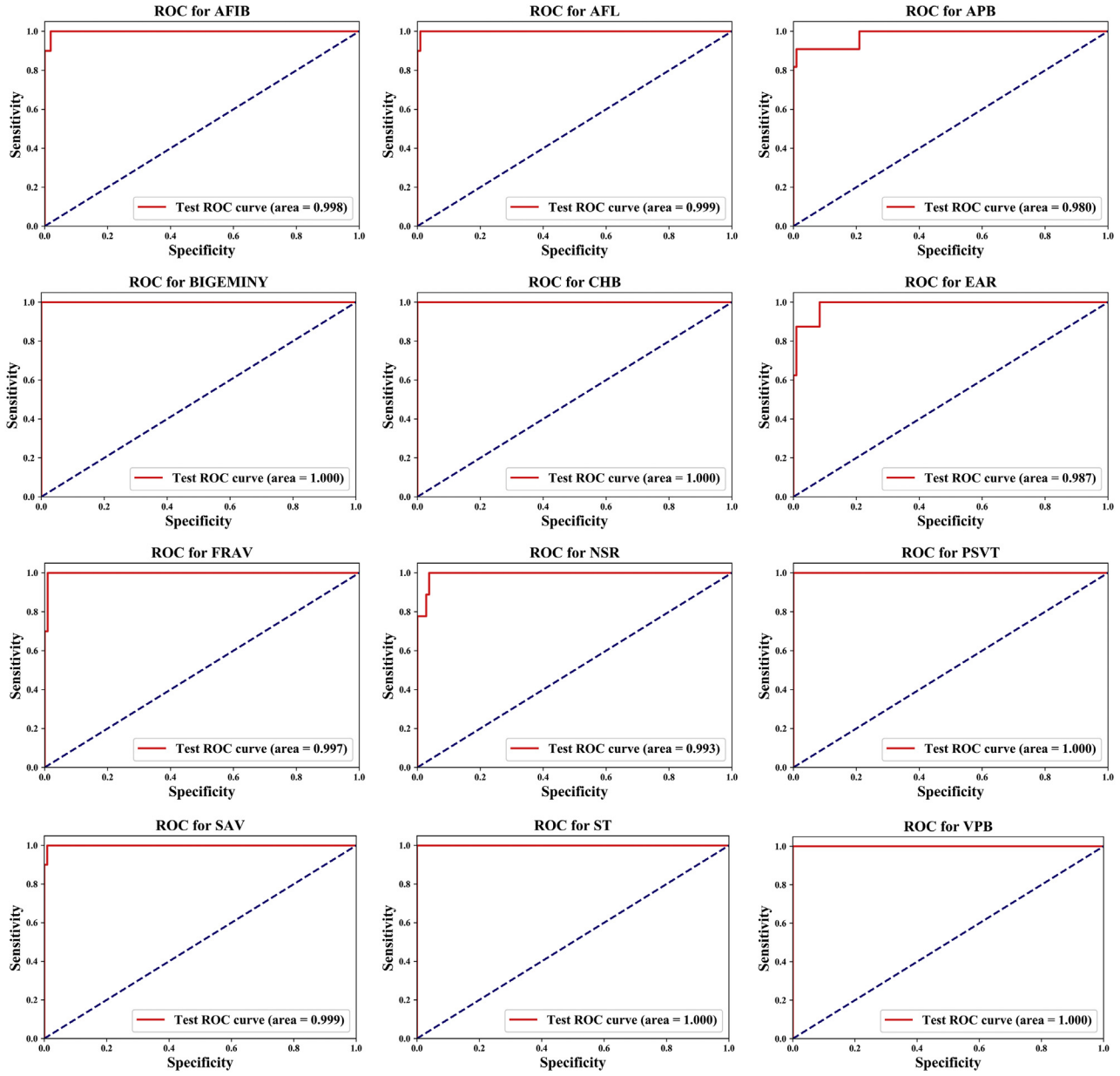

Figure 3. Receiver operating characteristic (ROC) curves for 12 heart-rhythm predictions, using the long short-term memory model. ROC curves were calculated at the sequence level for 12 heart rhythms. AFIB, atrial fibrillation; AFL, atrial flutter; APB, atrial premature beat; BIGEMINY, ventricular bigeminy; CHB, complete heart block; EAR, ectopic atrial rhythm; FRAV, first-degree AV block; NSR, normal sinus rhythm; PSVT, paroxysmal supraventricular tachycardia; SAV, second-degree AV block, type I; ST, sinus tachycardia; VPB, ventricular premature beat.

presented in Table 2 and Figure 3, respectively. The accuracies of the LSTM model for the classification of the 12 heart rhythms were all $>$ 0.982 (range 0.982-1.0), and the model achieved an AUC of $>$ 0.987 for all rhythm classes (range 0.987-1.0). The precision and recall ranged from 0.692 to 1 and from 0.625 to 1, respectively. The F_1 scores, which represented the harmonic mean of precision and recall, ranged from 0.777 to 1.0 for the LSTM model.

The performance of our model in classifying the 5 rhythm classes (AFIB, APB, FRAV, NSR, and VPB) in CPSC common to our study is presented in Table 3. The accuracy against the electrophysiologists' criterion annotations of the CPSC data set was 0.854 for AFIB, 0.689 for APB, 0.733 for FRAV, 0.930 for NSR, and 0.657 for VPB, respectively.

To compare the performance of the LSTM model with that of different board-certified physicians, we computed the classification accuracy and labelling time needed for testing 116 ECG records for 10 cardiologists, 8 emergency physicians, 10 internists, and the LSTM model. Table 4 presents the classification accuracy and labelling times for the LSTM model and the different groups of board-certified doctors. The overall accuracy of the model (0.90) was superior to the mean accuracies of the internists (0.55), emergency physicians (0.73), and cardiologists (0.83). The labelling time was only

Table 3. The performance of the long short-term memory model with the use of the CPSC data set

CPSC/model			CPSC dataset		Mostayed et al.
	Type	Abbreviation	Matched/total	Accuracy*	Accuracy†
1	Normal/normal sinus rhythm	NSR	896/963	0.930	0.82
2	Atrial fibrillation	AF/AFIB	974/1140	0.854	0.74
3	First-degree atrioventricular block/first-degree AV block	1-AVB/FRAV	533/727	0.733	0.70
4	Premature atrial contraction/atrial premature beat	PAC/APB	412/598	0.689	0.57
5	Premature ventricular contraction/ventricular premature beat	PVC/VPB	448/682	0.657	0.88

CPSC, China Physiological Signal Challenges.
*Combination of CPSC training set and validation set.
†Only CPSC validation set.

5.7 seconds for the model, which was far shorter than the time needed by the physicians to complete the annotations.

Table 5 presents the classification accuracies of the physicians and the LSTM model for the 12 heart rhythms. As expected, for each of the 12 cardiac rhythms, the annotations performed by the cardiologists was more accurate than those performed by the internists, and accuracy rates for the cardiologists were equivalent to or higher than those of the emergency physicians. Notably, the model outperformed the cardiologists in 11 of the 12 rhythm classes, with the exception of EAR (0.62 vs 0.66). The LSTM model classified the majority of the 12 testing rhythms with an accuracy rate from 0.82 to 1. The accuracy rates of the LSTM model for predicting EAR and FRAV were 0.62 and 0.7, respectively, which were similar to those of the cardiologists (0.66 and 0.5, respectively). Interestingly, none of the internists correctly annotated the EAR. The labelling time ranged from 20 to 71 minutes for the internists, from 29 to 71 minutes for the emergency physicians, and from 22 to 42 minutes for the cardiologists (Supplemental Table S1).

The 2 confusion matrices exhibited a similar pattern of concordance and discordance between the LSTM model and the averages for the cardiologists against the criterion annotations performed by an electrophysiologist consensus committee (Fig. 4). For example, the prediction success was consistently high for both AFIB and BIGEMINY. However, the LSTM model and cardiologists often had difficulty distinguishing between EAR and NSR and between FRAV and NSR.

The interobserver errors of individual rhythm class in each group of the labelling physicians are shown in box and scatter plots in Figure 5. We also calculated the between-group variation of accuracy, which showed significant variations in annotating 9 of the 12 rhythm classes among the 3 groups of physicians (Supplemental Table S2), particularly for EAR ($F = 21.17$), AFL ($F = 8.1$), PSVT ($F = 7.71$), and CHB ($F = 6.14$).

The concordance of ECG annotations from the 3 board-certified electrophysiologists was 98.3%. Only 2 of the 116 testing ECGs showed discordant labelling among the 3 electrophysiologists (Supplemental Table S3), which was resolved by holding a meeting to reach the final consensus. We further checked the model's performance and the cardiologists' performance in labelling the 2 particular ECGs with discordant labelling. For index ECG no. 30 (Fig. 6), the LSTM model and 4 cardiologists correctly classified the rhythm as APB, whereas the remaining 6 cardiologists annotated it as NSR. For index ECG no. 64 (Fig. 7), which was poorly interpreted by cardiologists as well as misclassified by the model, the LSTM model and 9 cardiologists erroneously annotated it as NSR, with only 1 cardiologist correctly classifying it as FRAV. Another example of misidentified ECGs by the model but not by cardiologists is shown in Figure 8. The electrophysiologists' criterion annotation was AFIB; 8 out of 10 cardiologists correctly annotated the rhythm as AFIB, but the LSTM

Table 4. Accuracy and labelling time for the model and doctors

Group	Accuracy	Labelling time
Board-certified doctors		
Internists	55% (0.55 ± 0.14)	46′09″ ± 13′35″
Emergency physicians	73% (0.73 ± 0.08)	35′49″ ± 7′52″
Cardiologists	83% (0.83 ± 0.10)	30′21″ ± 6′02″
Average	70% (0.70 ± 0.17)	37′33″ ± 11′56″
LSTM model	90% (0.90)	5.7″ (49 ms/sample)

Data are presented as % and mean ± SD.
′, minutes; ″, seconds; LSTM, long short-term memory.

Table 5. Average accuracies of the physicians and the LSTM model for the classification of each of the 12 heart rhythms

Heart rhythm	Internists	Emergency physicians	Cardiologists	LSTM model
AFIB	0.65 ± 0.23	0.86 ± 0.23	0.9 ± 0.15	0.90
AFL	0.6 ± 0.17	0.69 ± 0.15	0.89 ± 0.14	1.00
APB	0.37 ± 0.25	0.51 ± 0.20	0.73 ± 0.25	0.82
BIGEMINY	0.65 ± 0.42	0.91 ± 0.23	0.95 ± 0.08	1.00
CHB	0.5 ± 0.23	0.80 ± 0.26	0.83 ± 0.14	1.00
EAR	0.00 ± 0.00	0.23 ± 0.23	0.66 ± 0.30	0.62
FRAV	0.27 ± 0.24	0.48 ± 0.20	0.50 ± 0.15	0.70
NSR	0.96 ± 0.05	0.93 ± 0.08	0.98 ± 0.04	1.00
PSVT	0.47 ± 0.29	0.81 ± 0.20	0.86 ± 0.17	1.00
SAV	0.68 ± 0.20	0.73 ± 0.12	0.88 ± 0.13	0.90
ST	0.67 ± 0.25	0.84 ± 0.15	0.83 ± 0.27	0.90
VPB	0.75 ± 0.26	0.98 ± 0.04	0.94 ± 0.07	1.00
Overall	0.55 ± 0.14	0.73 ± 0.08	0.83 ± 0.10	0.90

Data are presented as mean ± SD.
AFIB, atrial fibrillation; AFL , atrial flutter; APB, atrial premature beat; BIGEMINY, ventricular bigeminy; CHB, complete heart block; EAR, ectopic atrial rhythm; FRAV, first-degree AV block; LSTM, long short-term memory; NSR, normal sinus rhythm; PSVT, paroxysmal supraventricular tachycardia; SAV, second-degree AV block (type I); ST, sinus tachycardia; VPB, ventricular premature beat.

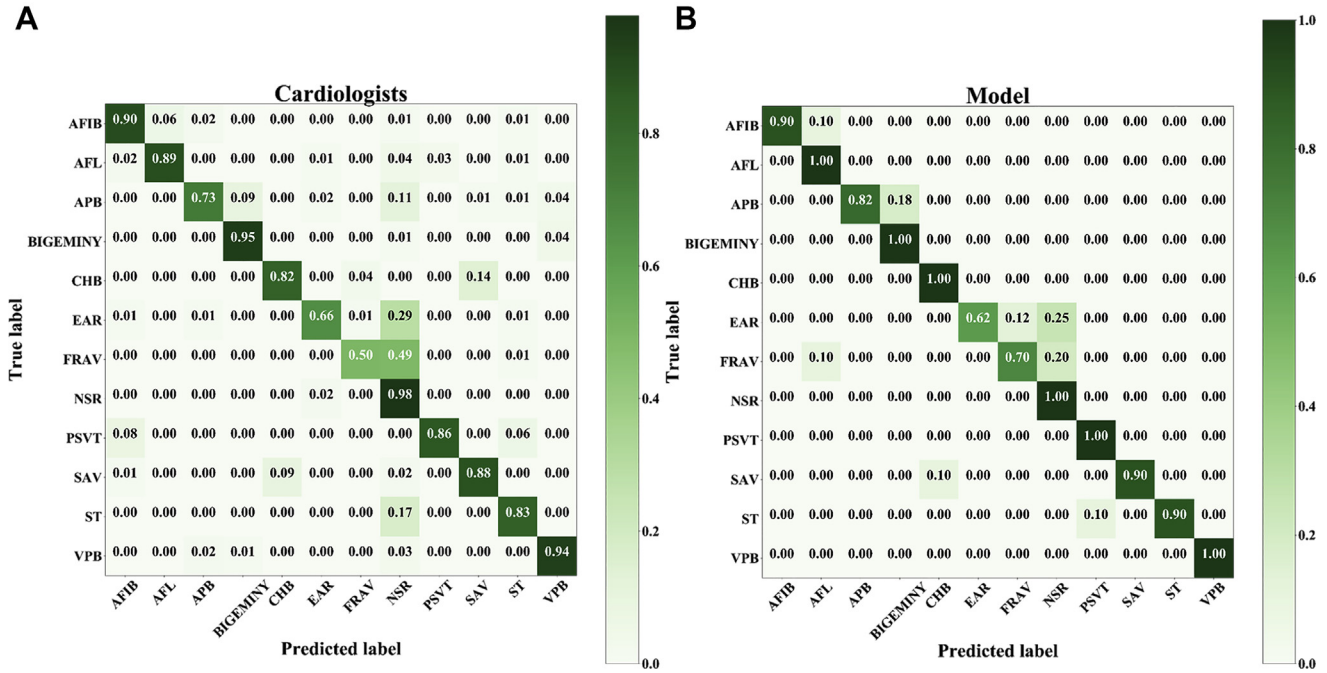

Figure 4. Confusion matrices for the long short-term memory model and different board-certified cardiologists. Confusion matrices for the predictions of (**A**) the cardiologists and (**B**) the LSTM model vs the electrophysiologist committee consensus. The accuracy for each heart rhythm category is displayed on a color gradient scale.

model and 1 cardiologist misclassified this ECG as AFL, and another cardiologist labeled it as ST.

Discussion

To our knowledge, this is the first study to demonstrate the feasibility of a deep-learning LSTM model for interpreting the 12 common heart rhythms with the use of a large number of 12-lead ECG signals. The usefulness of the model was confirmed by comparing its rhythm-classification performance with that of board-certified physicians from different specialties. The rhythm-classification accuracy of the LSTM model was superior to that of emergency physicians and internists and outperformed the cardiologists in 11 of the 12 rhythm classes. Because of its ultrafast rhythm classification and performance superior to physicians (who had longer annotation times and inferior accuracy), the proposed model may have clinical relevance for the early diagnosis of cardiac-rhythm disorders.

Recently, Mostayed et al.[18] used a bidirectional LSTM network to classify 9 ECG types based on the 12-lead ECG signals in a CPSC data set. Our LSTM model differs from that of Mostayed et al. in details of the neural network architecture as well as in the inputs and outputs of 12-lead ECG data (Supplemental Table S4). The input data set of the present study was from 65,932 12-lead ECG recordings compared with only 6877 ECG waveforms used in the Mostayed et al. study. Mostayed et al. trained their model on the basis of segments, with each of them consisting of 4 cardiac beats from a given ECG record, whereas we trained our model according to full ECG waveforms without dividing them up into a certain number of cardiac beats. For the classification outputs, our model was designed to classify 12 rhythm types, whereas Mostayed et al. predicted 9 ECG classes in their model, with 5 classes (AFIB, APB, FRAV, NSR, and VPB) common between the 2 models. Our model outperformed in 4 of the 5 rhythm classes common to the 2 models, as presented in Table 3. We noticed that although our model did extremely well on our own data set (Table 2), its performance on VPB classification was worse than that of Mostayed et al. on the CPSC data set. It is possible that the discrepancy might be attributed to the differences in data set used or in physicians' labelling for this particular rhythm class. The exact reasons for this inconsistency remain to be determined. It should also be noted that our model was further tested on 116 ECG recordings against the consensus from a committee of 3 board-certified electrophysiologists in a real-world setting, and the model was demonstrated to outperform 10 internists, 8 emergency physicians, and 10 cardiologists in overall performance.

Twelve-lead ECG vs single-lead samples

Because of the complexity of 12-lead ECG signals, previous studies of deep learning for classifying rhythm disorders commonly used single-lead ECG records[12,14,19,20] or 2-lead ECG records.[21] Hannun et al. developed a DNN to classify 12 rhythm classes with the use of 91,232 single-lead ECGs recorded by a single-lead ambulatory ECG device. In their study, the DNN achieved an average ROC of 0.97 and an average F_1 score of 0.837, exceeding the average performance of cardiologists. However, because the input data set comprised single-lead ECG signals with limited signal information compared with standard 12-lead ECG, it remains to be determined whether that algorithm

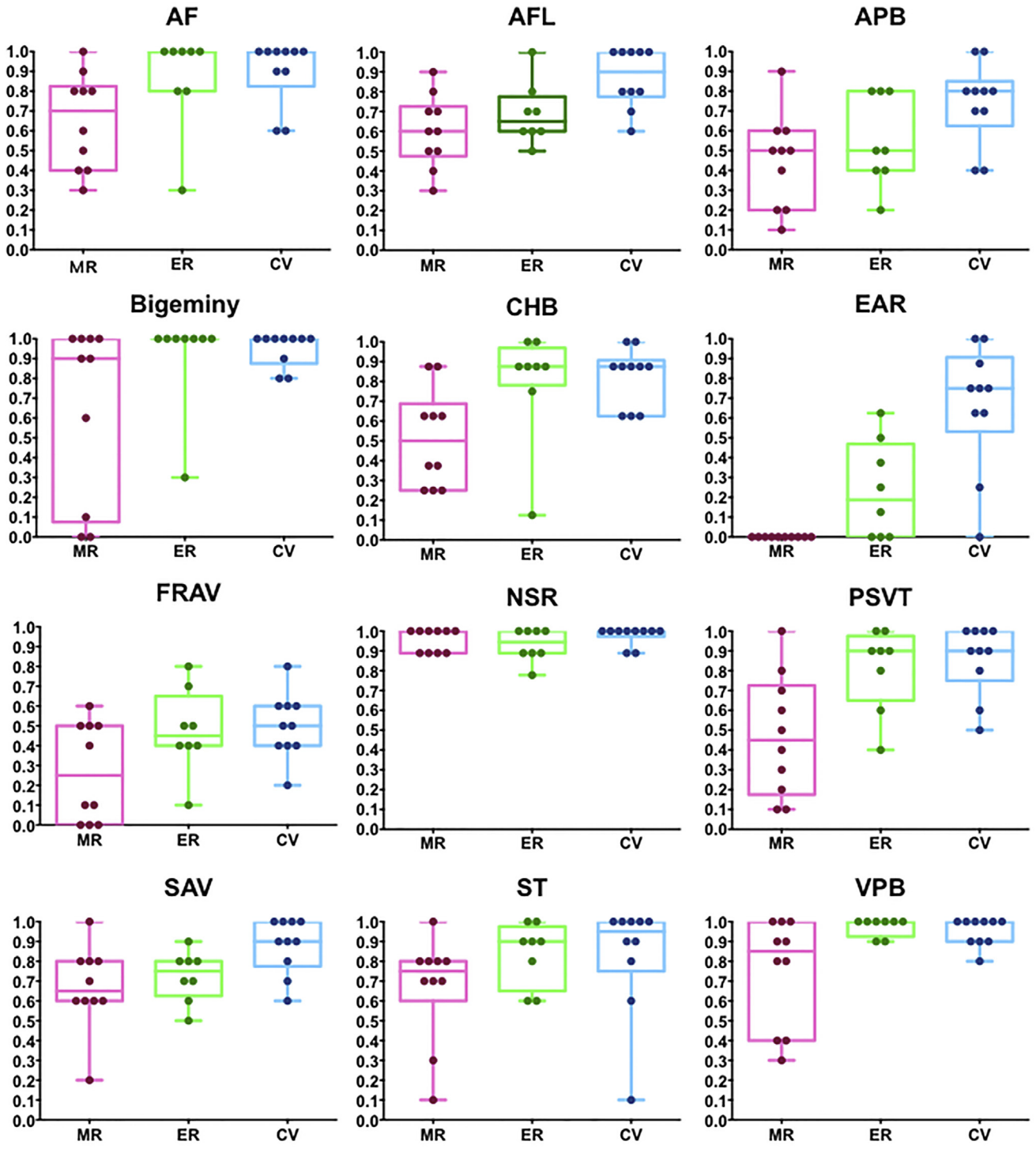

Figure 5. The interobserver error of individual rhythm class in each group of the labeling physicians. The box-scatter plot shows the interobserver variation of 12 heart rhythms among three groups of physicians. AFIB, atrial fibrillation; AFL, atrial flutter; APB, atrial premature beat; BIGEMINY, ventricular bigeminy; CHB, complete heart block; CV, cardiologists; EAR, ectopic atrial rhythm; ER, ER physicians; FRAV, first-degree AV block; MR, internists; NSR, normal sinus rhythm; PSVT, paroxysmal supraventricular tachycardia; SAV, second-degree AV block, type I; ST, sinus tachycardia; VPB, ventricular premature beat.

performance can be generalized to 12-lead ECG—based rhythm classification. Furthermore, most of the previous studies used data from the MIT-BIH arrhythmia database (PhysioNet), which has limitations such as an insufficient number of patients and rhythm strips for a broad spectrum of arrhythmia classification.[22]

One of the strengths of the present study is that we used 65,932 digital 12-lead ECG signals from 38,899 patients that

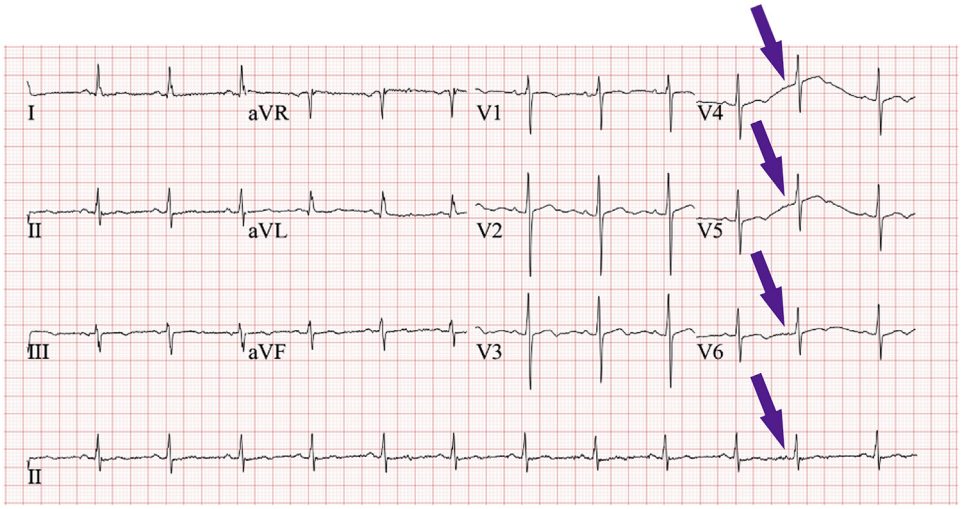

Figure 6. Example 1: Electrocardiogram of discordant labelling by a consensus committee of 3 board-certified electrophysiologists. The **purple arrows** indicate the atrial premature beat (APB). The long short-term memory model and 4 cardiologists correctly classified the rhythm as APB, whereas the remaining 6 cardiologists annotated it as normal sinus rhythm.

were exclusively labelled by 5 experienced cardiologists and then stored in a unique ECG data set according to a standardized protocol to develop the LSTM model. LSTM, which is a type of RNN model, can accept a wide range of time-series structure inputs. This is particularly useful for machine learning—based classification of cardiac arrhythmias with the use of 12-lead ECG recordings.[19] LSTM models can solve complex, artificial, long time-lag tasks[16] and overcome the limitations of conventional RNNs by accessing past and future input features at a given time and applying a 2-way LSTM network, as proposed by Graves and Schmidhuber.[20] Because of these strengths, the LSTM model outperformed the board-certified physicians from 3 specialties in the classification accuracy of the 12 cardiac rhythms. Recently, Yildirim also developed a bidirectional LSTM network-based model for classifying ECG signals[23]; however, the data was based on single-lead ECGs from the PhysioBank MIT-BIH arrhythmia database to classify 5 rhythm types only, using a relatively small number of ECG samples for training, validation, and testing. The present study further expands the use of the deep-learning LSTM model to predict more clinically relevant heart rhythms with the use of a larger number of 12-lead ECG signals.

Testing in the clinically relevant scenario

The 12-lead ECG is a very useful first-line diagnostic tool for the detection of various cardiovascular diseases, including cardiac-rhythm disorders, conduction abnormalities, and myocardial ischemia or infarction, by physicians across different specialties. In hospital settings, emergency physicians or internists are usually the first medical contacts for patients; thus, they must interpret 12-lead ECGs at the scene. We compared the performance not only between the model and cardiologists but also between the model and emergency physicians and internists. This study is the first to demonstrate the evidence-based usefulness of the LSTM model for predicting the 12 common cardiac rhythms. The classification accuracy and ultrafast annotation time of the model were

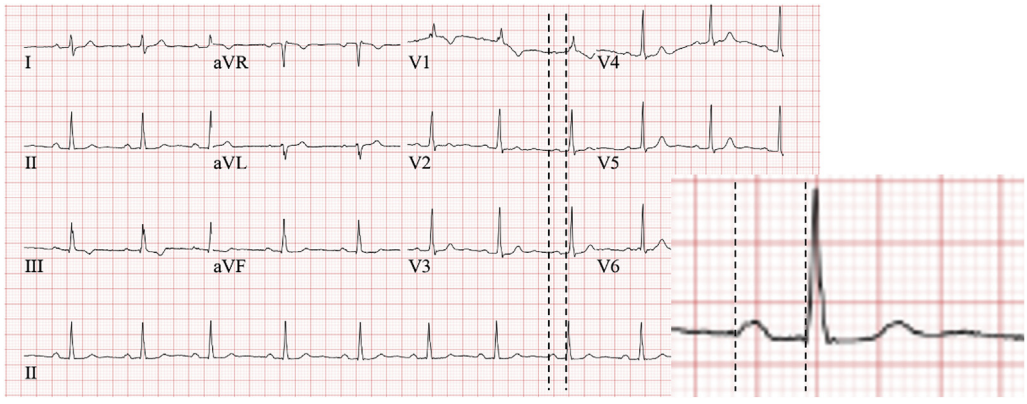

Figure 7. Example 2: Electrocardiogram of discordant labelling by a consensus committee of 3 board-certified electrophysiologists. The interval between the 2 **dotted lines** shows prolongation of the P-R interval. The long short-term memory model and 9 cardiologists erroneously annotated normal sinus rhythm, with only 1 cardiologist correctly classifying it as first-degree atrioventricular block.

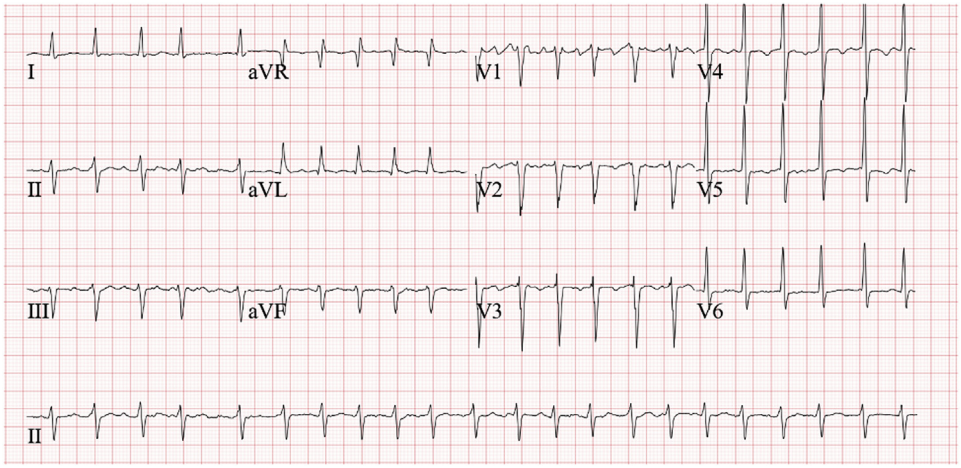

Figure 8. Example of an electrocardiogram misidentified by the long short-term memory (LSTM) model. The electrophysiologists' consensus was atrial fibrillation (AFIB); 8 of the 10 cardiologists correctly annotated the rhythm as AFIB, whereas the LSTM model and 1 cardiologist misclassified it as atrial flutter, and another cardiologist labeled it as sinus tachycardia.

superior to those of all the board-certified physicians, including the cardiologists. We think that this approach may reduce the misdiagnosis rate for computerized 12-lead ECGs and enhance the efficiency of 12-lead ECG interpretation by more precisely triaging or accelerating the decision-making process in patient care.

Limitations

This study had several limitations. First, the LSTM model was used only to predict the 12 rhythm classes and not for detecting the ST-T change, which is important for diagnosing acute myocardial infarction.[24] Indeed, our ongoing work is to incorporate prelabelled 12-lead ECG data of different clinically relevant cardiovascular diseases, such as ST-segment-elevation acute myocardial infarction, non−ST-segment-elevation myocardial infarction, ventricular preexcitation, and ventricular tachycardia, for machine learning with the use of our large 12-lead ECG database over a more than 10-year period. Second, we did not add specific ECG noise, including baseline wander, electrode motion artifacts, and muscle artifacts, to the input data for training. Therefore, the proposed model may not be effective in cases of ECG noise. Third, our test data set consisted of only 116 patients; the prediction accuracy for rare rhythms should be investigated by increasing the amounts of training and testing data. Fourth, the current LSTM model was not designed to interpret more complex ECG patterns, such as those patients with AFIB combined with intraventricular conduction disturbances. Therefore, we did not include any ECGs of AFIB combined with intraventricular conduction disturbances for the clinical testing in Table 5. However, we found that, of the 18,077 ECGs labeled as AFIB in the training/validation data set (Table 1), 387 (2.14%) were compounded with intraventricular conduction disturbances, and the LSTM model accurately interpreted them to be AFIB in 29 of the 32 testing cases (0.91) in the same data set. Further studies with improvement of machine-learning technologies are required to circumvent this limitation for interpreting more complex arrhythmias. Finally, complex AV block and differentiation of wide QRS-complex tachycardia are 2 areas where physician interpretation may fail and where the failure is most clinically relevant. In this study, we aimed to test the feasibility and efficacy of using a deep-learning LSTM model for classifying the 12 common heart rhythms with the use of a large number of 12-lead ECG signals. The model is not used to differentiate wide QRS-complex tachycardia mainly because we do not have a sufficient number of wide QRS-complex tachycardia ECGs for machine learning so far. Indeed, the usefulness of our model in differentiating wide QRS-complex tachycardia should be confirmed in future studies.

Conclusion

We demonstrated the feasibility and effectiveness of a deep-learning LSTM model for interpreting the 12 common heart rhythms with the use of a large number of 12-lead ECG signals. The findings may have clinical relevance for expediting the diagnosis of cardiac-rhythm disorders and facilitating decision making in patient management.

Funding Sources

This study was supported in part by the Taiwan Ministry of Health and Welfare (MOHW107-TDU-B-212-123004), the Ministry of Science and Technology (MOST 108-2314-B-039-055, MOST 107-2314-B-039-061, and MOST 106-2314-B-039-033), and the China Medical University Hospital (DMR-CELL-1802, DMR108-180, DMR-108-013, and DMR-109-012). None of these funding sources had a further role in study design; collection, analysis, or interpretation of data; writing the report; or decision to submit the paper for publication.

Disclosures

The authors have no conflicts of interest to disclose.

References

1. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.

2. Lee JG, Jun S, Cho YW, et al. Deep learning in medical imaging: general overview. Korean J Radiol 2017;18:570-84.

3. Becker AS, Marcon M, Ghafoor S, et al. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. Invest Radiol 2017;52:434-40.

4. Becker AS, Bluthgen C, Phi Van VD, et al. Detection of tuberculosis patterns in digital photographs of chest X-ray images using deep learning: feasibility study. Int J Tuberc Lung Dis 2018;22:328-35.

5. Baltruschat IM, Nickisch H, Grass M, Knopp T, Saalbach A. Comparison of deep learning approaches for multi-label chest X-ray classification. Sci Rep 2019;9:6381.

6. Pehrson LM, Lauridsen C, Nielsen MB. Machine learning and deep learning applied in ultrasound. Ultraschall Med 2018;39:379-81.

7. Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ. Deep learning for brain MRI segmentation: state of the art and future directions. J Digit Imaging 2017;30:449-59.

8. Al'Aref SJ, Anchouche K, Singh G, et al. Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. Eur Heart J 2019;40:1975-86.

9. Schlant RC, Adolph RJ, DiMarco JP, et al. Guidelines for electrocardiography. A report of the American College of Cardiology/American Heart Association Task Force on Assessment of Diagnostic and Therapeutic Cardiovascular Procedures (Committee on Electrocardiography). J Am Coll Cardiol 1992;19:473-81.

10. Schlapfer J, Wellens HJ. Computer-interpreted electrocardiograms: benefits and limitations. J Am Coll Cardiol 2017;70:1183-92.

11. Mjahad A, Rosado-Munoz A, Bataller-Mompean M, Frances-Villora JV, Guerrero-Martinez JF. Ventricular fibrillation and tachycardia detection from surface ECG using time-frequency representation images as input dataset for machine learning. Comput Methods Programs Biomed 2017;141:119-27.

12. Mathews SM, Kambhamettu C, Barner KE. A novel application of deep learning for single-lead ECG classification. Comput Biol Med 2018;99:53-62.

13. Guglin ME, Thatai D. Common errors in computer electrocardiogram interpretation. Int J Cardiol 2006;106:232-7.

14. Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 2019;25:65-9.

15. Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions. IEEE Trans Pattern Anal Mach Intell 2017;39:664-76.

16. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9:1735-80.

17. Kligfield P, Gettes LS, Bailey JJ, et al. Recommendations for the standardization and interpretation of the electrocardiogram: part I: the electrocardiogram and its technology: a scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society: endorsed by the International Society for Computerized Electrocardiology. Circulation 2007;115:1306-24.

18. Mostayed A, Luo J, Shu XL, Wee W. Classification of 12-lead ECG signals with bi-directional LSTM network 2018 [e-print]. arXiv:1811.02090.

19. Graves A, Liwicki M, Fernandez S, et al. A novel connectionist system for unconstrained handwriting recognition. IEEE Trans Pattern Anal Mach Intell 2009;31:855-68.

20. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw 2005;18:602-10.

21. Rahhal MMA, Bazi Y, AlHichri H, Alajlan N, Melgani F, Yager RR. Deep learning approach for active classification of electrocardiogram signals. Inf Sci 2016;345:340-54.

22. Moody GB, Mark RG. The impact of the MIT-BIH arrhythmia database. IEEE Eng Med Biol Mag 2001;20:45-50.

23. Yildirim O. A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification. Comput Biol Med 2018;96:189-202.

24. Afsar FA, Arif M, Yang J. Detection of ST segment deviation episodes in ECG using KLT with an ensemble neural classifier. Physiol Meas 2008;29:747-60.

Supplementary Material

To access the supplementary material accompanying this article, visit the online version of the *Canadian Journal of Cardiology* at www.onlinecjc.ca and at https://doi.org/10.1016/j.cjca.2020.02.096.