Multiple high-regional-incidence cardiac disease diagnosis with deep learning and its potential to elevate cardiologist performance

Yunqing Liu, Chengjin Qin, Chengliang Liu, Jinlei Liu, Yanrui Jin, Zhiyuan Li, Liqun Zhao

qinchengjin@sjtu.edu.cn (C.Q.)
chlliu@sjtu.edu.cn (C.L.)

Highlights
A large-scale clinical dataset containing multiple cardiac diseases is constructed

The performance of our DL model surpasses that of medical cardiologists

Cardiologists with DL results as reference can improve diagnostic performance

Our approach provides solutions for the future AI-aided diagnosis clinical system

Liu et al., iScience 25, 105434
November 18, 2022 © 2022
The Author(s)
https://doi.org/10.1016/j.isci.2022.105434
Multiple high-regional-incidence cardiac disease diagnosis with deep learning and its potential to elevate cardiologist performance

Yunqing Liu,1,2,4 Chengjin Qin,1,2,4,5,* Chengliang Liu,1,2,* Jinlei Liu,1,2 Yanrui Jin,1,2 Zhiyuan Li,1,2 and Liqun Zhao3

SUMMARY
Currently, due to lack of large-scale datasets containing multiple arrhythmias and acute coronary syndrome-related diseases, AI-aided diagnosis for cardiac diseases is limited in clinical scenarios. Whether AI-based ECG diagnosis can assist cardiologists to improve performance has not been reported. We constructed a large-scale dataset containing multiple high-regional-incidence arrhythmias and ACS-related diseases, including 162,622 12-lead ECGs collected between January 2018 and March 2021. We presented a deep learning model for clinical ECG diagnosis of multiple cardiac diseases. Results show that our model for diagnosing 15 cardiac abnormalities achieved 88.216% accuracy, and its average AUC ROC score reached 0.961. On the board-certified re-annotated dataset, its performance surpasses that of cardiologists in non-reference group. Moreover, with aid of labels given by our model, accuracy and efficiency for cardiologist increased by 13.5% and 69.9% than non-reference group. Our approach provides solutions for AI-aided diagnosis systems of cardiac diseases in applications.

INTRODUCTION
According to the statistics of World Health Organization, approximately 17.9 million people died from cardiovascular diseases (CVDs) in 2019, rendering them the leading cause of death globally.1 If CVDs can be detected and intervened as early as possible, and corresponding measures are actively taken to prevent them, the risk of the life-threatening diseases will be greatly reduced.2–6 Around billions recorded annually and commonly utilized in clinical settings from health screening to intensive care units, electrocardiogram (ECG) is a principal tool to diagnose CVDs in early stages.7–9

As a tool frequently employed in clinical scenarios, 12-lead ECGs provide a large number of spatiotemporal features and comprehensive diagnostic assessment of cardiac activity.10 Although 12-lead ECGs contain abundant information about CVDs, the high-dimensional and information-intensive characteristics make manual ECG interpretation time-consuming and insufficiently credible. Fortunately, automated computerized interpretation of 12-lead ECGs has the potential to reach or even surpass the performance of physicians, and provides a significant reference for physicians’ diagnosis. If years of ECG records worldwide are consolidated into the large-scale databases and fully exploited using artificial intelligence (AI) and big data technology, the performance of automated computerized interpretation can be further greatly improved, making it promising for clinical application.

Benefiting from the specially collected and sorted big-data datasets with more than millions of samples, deep learning has achieved great success in image classification and speech recognition.11–13 In recent years, scholars were trying to make full use of the advantages of AI and big data to promote the development of social medical and health system and the intelligent level of medical diagnosis.14 In research combining AI with human medical health issues, the use of large-scale datasets with diverse multinational cohort could lead to high accuracy, satisfactory sensitivity and specificity in the classification of COVID-19 pneumonia of the public health emergency,15,16 as well as increase the accuracy, consistency, and adoption of lung cancer screening worldwide.17 Also, big data, high-performance computing, and deep learning are increasingly becoming key to precision medicine from identifying disease risks and taking preventive measures, to making diagnoses and personalizing treatment for individuals.18
As for the automated computerized interpretation for CVDs, scholars have recently conducted some preliminary attempts. Hannun et al.\textsuperscript{19} developed a deep neural network (DNN) trained on around 91,000 single-lead ECGs and achieved satisfactory results in 12 arrhythmia types. However, as the authors pointed out in the paper, single-lead ECGs contain limited information compared to 12-lead ECGs, and the model performance applied in 12-leads remains to be explored. Also, single-lead paradigm is incapable of accurately diagnosing some complex cardiac disease types, such as acute coronary syndrome (ACS)-related diseases or ACS-related diseases with arrhythmias. Another study that is relevant to this work is the study by Zhu et al.\textsuperscript{20} that 12-lead ECGs for the arrhythmia diagnosis were applied and successfully distinguished a range of different arrhythmias with single- or multi-type labels, exceeding the average performance of the physicians. However, the authors noted that a few types of rhythms and conduction arrhythmias were not
included due to an absence of patient samples. In addition, ECG abnormalities such as myocardial infarction and atrial or ventricular hypertrophy had not been included in their study. Ribeiro et al. 21 also studied about arrhythmia diagnosis, and based on a large-scale dataset from Brazil, they trained a model which outperformed cardiologists in 6 types of abnormalities. As the authors pointed out, the DNN ability to diagnose other ECG abnormalities, including ACSs or cardiac chamber enlargements, and uncommon arrhythmias, remains to be explored and validated. Based on the Brazilian dataset, a deep learning model was developed by Sangha et al. 22 to realize the diagnosis of 6 heart diseases, which had promising performance on the test dataset. However, the authors also pointed out that, among the 6 clinical labels the study focused on, most are types of arrhythmias, and their models would not apply to other cardiovascular diseases. It should be noted that their algorithm can deal with either ECG images (in paper form) or signals as inputs to predict the probability of various rhythm and conduction disorders. An ML model-based classification approach for patients with ACS was established and evaluated by Zaiti et al., 23 and it was the first study that prospectively validated and tested the performance of ML-based models on two separate cohorts to predict ACS using only the prehospital 12-lead ECG. However, as the authors pointed out, the number of the abstracted ECG features was not proportional to the size of the dataset, which might have affected the performance of the classifiers. Increasing the number of patients would probably lead to increased performance. Second, although manual feature selection had a positive effect on the ML classifiers’ performance, further data-driven techniques for feature selection need to be further investigated. Also, some other studies 24–29 which attempt to analyze 12-lead ECGs achieved promising results. It should
be also noted that many scholars have made efforts on the public-accessible datasets containing a small number of patients and achieved some results on ECG images or signals. However, since these datasets were collected from only dozens of patients (the Massachusetts Institute of Technology-Beth Israel Hospital arrhythmia database: 47 patients, European ST-T database: 79 patients, and St. Petersburg INCART database: 75 patients), the potential for practical clinical application of the above approaches is quite limited.

The use of DNNs for automated computerized 12-lead ECG interpretation remains to be unexplored largely. The current existing large-scale datasets include only a limited number of arrhythmias or ACS-related diseases. That is, at present, there is still no such large datasets containing both multiple arrhythmias and ACS-related diseases simultaneously. However, in many regions, these two multiple/hybrid types of cardiac diseases, ranging from arrhythmias to ACS-related diseases, are highly frequent occurring simultaneously, posing serious threats to human health. A large-scale database containing comprehensive types of diseases will incredibly promote the clinical application of deep learning in ECG disease diagnosis and fill the gap in this field. In addition, computerized classification research for standard (10 s) 12-lead ECGs exists certain limitations, since the actual duration of ECGs measured in clinical settings may not be standard 10 s. The automated computerized interpretation research of the variable-length 12-lead ECGs will yield higher theoretical and clinical practical value.

In this study, we constructed a large-scale dataset containing both multiple arrhythmias and ACS-related diseases with high regional incidence, including 162,622 12-lead ECG records. The proposed DNN-based model for computerized 12-lead ECG interpretation was trained, verified, and tested on this large dataset, realizing the classification detection for not only multiple arrhythmias but also types of ST segment change to aid in the diagnosis of ACS-related diseases in practice. T wave change, ventricular hypertrophy, and other cardiac abnormalities were also involved into this research to achieve a large-scale and comprehensive analysis and diagnosis for cardiac diseases utilizing 12-lead ECGs. This approach can not only deal with the variable-length 12-lead ECGs by generating and processing two-dimensional images but also perform an end-to-end paradigm on the ECGs of compound types. Our developed deep network model outperformed Medical PhDs with 4–5 years of experience interpreting ECGs in the interpretation performance of a range of different ECG types in current test application practice. Moreover, with the results of deep

### Table 1. Performance summary of our deep learning model in ECG interpretation, and F1 scores on the test dataset

| Arrhythmia | Model AUC | ROC | Model sensitivity | Model specificity | Model F1 score |
|------------|-----------|-----|-------------------|------------------|----------------|
| Normal sinus rhythm | 0.920     | 0.965 | 0.785             | 0.940            |
| Sinus tachycardia | 0.990     | 0.884 | 0.993             | 0.888            |
| Sinus bradycardia | 0.993     | 0.922 | 0.993             | 0.905            |
| Sinus rhythm & T wave changes | 0.932     | 0.376 | 0.991             | 0.464            |
| Sinus arrhythmia | 0.896     | 0.333 | 0.994             | 0.418            |
| Sinus rhythm & ST-T segment changes | 0.917     | 0.496 | 0.992             | 0.565            |
| Sinus rhythm & Complete right bundle branch block | 0.985     | 0.899 | 0.999             | 0.915            |
| Sinus rhythm & Premature atrial contraction | 0.954     | 0.570 | 0.995             | 0.586            |
| Atrial fibrillation | 0.982     | 0.712 | 0.998             | 0.750            |
| Sinus rhythm & Left ventricular high voltage | 0.974     | 0.441 | 0.996             | 0.490            |
| Sinus rhythm & Premature ventricular contraction | 0.939     | 0.667 | 0.998             | 0.696            |
| Sinus rhythm & Left ventricular hypertrophy with ST-T segment changes | 0.982     | 0.667 | 0.998             | 0.606            |
| Sinus tachycardia & ST-T segment changes | 0.978     | 0.317 | 0.999             | 0.426            |
| Sinus bradycardia & T wave changes | 0.972     | 0.632 | 0.998             | 0.595            |
| Atrial fibrillation & ST-T segment changes | 0.989     | 0.591 | 0.999             | 0.634            |
| Paced rhythm | 0.978     | 0.835 | 0.999             | 0.871            |

AUC, area under the curve; ROC, receiver operating characteristic.
learning model results as reference, they can achieve remarkable higher accuracy and efficiency, which means that AI will effectively assist and improve the current medical diagnosis system.

RESULTS

Testing and performance evaluation

The model network was elaborately designed as a 6-block deep network, and Section STAR Methods shows the specific details of the network. Neural network weights were initialized with Xavier and the bias was initialized to 0. Adam optimizer with learning rate 0.0001 was utilized, and the CrossEntropy loss function was employed. The training process ran for 80 epochs, and we chose the best-performing deep learning model in the validation dataset as the final model for further research. The deep learning model was accurate on 15,871 (88.216%) of the 17,991 cases in the test dataset. The overall accuracy of our research was better than the average physician accuracy of 70% and the model accuracy of 80% of the research in the top medical journal (LANCET DIGIT HEALTH) by Zhu and colleagues, assuming all the labels were appropriate in the test dataset of this research. Figures 1 and 2 and Table 1 show the performance of our deep learning model. Satisfactory performance metrics are achieved for all ECG abnormalities.

Results on re-annotation dataset by professional cardiologists

We selected a group of ECGs from the above-mentioned test dataset ($C_{TEST}$) to construct a new dataset called re-annotated dataset, $C_{RAT}$. The ECGs were discussed by a committee of cardiologists and the final label was given with a final consensus. To further assess the performance of our deep learning model and how deep learning can aid clinical diagnosis, we set two groups with contrast conditions during the re-annotation process as the reference group and the non-reference group: one group of scientists was provided with our deep learning model results for reference during the re-annotation, and the other group was not provided with any reference. The two groups of annotators have similar medical work backgrounds and experiences: they are Medical PhD in cardiovascular medicine with 4–5 years of clinical experience in interpreting ECG working in top cardiology hospitals in China. See STAR Methods for details.
Our deep learning model was accurate in 670 (78.5%) of 854 cases in the re-annotated dataset, which was much better than the 73.5% accuracy of the cardiologist in the non-reference group (628 cases were accurate, and 7 cases were classified out of the 16 types). However, with our deep learning model result label as reference, the cardiologist in the reference group achieved the 87.0% accuracy (743 cases were accurate, and 1 case was classified out of the 16 types). Figures 3–5 show detailed confusion matrices.

We compared the F1 scores of our deep learning model with those of the ECG cardiologists. Using the board-certified final labels as the gold standards, F1 scores of the cardiologists increased dramatically as the reference of our deep learning predicted label was involved. The average F1 score was 0.864 for the reference group, and 0.731 for the non-reference group. The mean F1 score of our deep learning model was 0.769, which exceeded the score of the cardiologist without reference. The F1 scores obtained by the reference group for all the 16 types were all higher than the scores of the non-reference group except for sinus arrhythmia. The model had F1 scores ranging from 0.567 (sinus rhythm & T wave changes) to 0.972 (sinus rhythm & complete right bundle branch block); in 11 out of the 16 types, our deep learning model, on average, had higher F1 scores than the cardiologist in the non-reference group. Table 2 shows the details.

The mean AUC ROC score of our deep learning model for interpreting all 16 types was 0.965, with sensitivity 0.757 and specificity 0.985. The lowest AUC ROC score was 0.934 for sinus rhythm & T wave changes excluding normal sinus rhythm (0.914) and the highest was 0.972 for sinus rhythm & complete right bundle branch block. For interpretation of the single-type arrhythmias (sinus tachycardia, sinus bradycardia, sinus arrhythmia, sinus rhythm & premature atrial contraction, atrial fibrillation, and sinus rhythm & premature ventricular contraction), the mean AUC ROC was 0.962. For interpretation of the QRS complex-type abnormalities (sinus rhythm & left ventricular high voltage and sinus rhythm & left ventricular hypertrophy with ST-T segment changes), the mean AUC ROC was 0.985. For interpretation of the ST-T segment-type abnormalities (sinus rhythm & ST-T segment changes, sinus rhythm & left ventricular hypertrophy with ST-T segment changes, sinus tachycardia & ST-T segment changes, and atrial fibrillation & ST-T segment changes), the mean AUC ROC was 0.974. For interpretation of the T wave-related abnormalities (sinus rhythm & T wave changes and sinus bradycardia & T wave changes), the mean AUC ROC was 0.940.
When comparing ROC curves with those of the cardiologists, the reference group performed better comprehensively as shown in Figure 6, except in interpreting sinus rhythm & T wave changes and sinus arrhythmia. Of all the 16 types, our model performed approximately equal to or better than the cardiologist in non-reference group in 11 types such as normal sinus rhythm and sinus rhythm & ST-T segment changes. The confusion matrices of our deep learning model (0.757) showed higher average sensitivity compared with the cardiologist in non-reference group (0.728). The cardiologist group with deep learning labels as reference achieved the highest average sensitivity at 0.853, which means that the missed diagnosis rate was the lowest. The average specificity was 0.991 for the reference group, 0.985 for the proposed model, and 0.983 for the non-reference group, of which the magnitude order is the same as the average sensitivity.

Experimental results also showed that cardiologists in the reference group were more efficient in interpreting ECGs than those in the non-reference group. Cardiologists in the reference group reported that the average time of interpreting all the ECG records was 54.1 s, while those in the non-reference group reported the value was 180.0 s. With labels obtained by the deep learning model for reference, the interpreting efficiency of the cardiologists increased by around 69.9%. Also, we requested cardiologists in the reference group for descriptive evaluations of all labels obtained by our deep learning model. Cardiologists gave positive feedbacks to the 717 records of deep learning labels of all the 854 ECG records, believing that these labels were helpful for their clinical ECG interpretation.

Results for complex ECGs

We randomly selected 200 records in the remaining 41,744 records of the complete dataset to evaluate the model interpretation ability in the face of complex situations. These 41,744 records include ECGs with severe interfering or zero signals, three or more arrhythmias, and other arrhythmia types not included in the above 16 types. At the same time, clinicians of our cooperated hospital sometimes gave specific disease description as the diagnosis label, such as old inferior myocardial infarction (If mapped into our model, the final output may include ST-T segment changes.). The results showed that only 37 (18.5%) board-certified final labels in the 200 ECG records contained new types of heart disease, such as first-degree atrioventricular block, pathological Q wave. 116 (58.0%) output interpretation results of the proposed model were

![Figure 5. Confusion matrix for the predictions of the non-reference group versus the board-certified final result](image)

### Table:

| A | 81 | 3 | 13 | 3 | 3 | 2 | 2 | 1 | 0 | 11 | 0 | 0 | 0 | 0 | 0 | 0 |
|---|---|---|----|---|---|---|---|---|---|----|---|---|---|---|---|---|
| B | 7 | 54 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C | 0 | 0 | 41 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| D | 22 | 1 | 4 | 31 | 0 | 13 | 0 | 1 | 2 | 1 | 0 | 12 | 1 | 9 | 0 | 0 |
| E | 3 | 0 | 1 | 0 | 40 | 0 | 0 | 3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| F | 5 | 1 | 0 | 4 | 0 | 30 | 1 | 1 | 0 | 3 | 0 | 10 | 2 | 2 | 0 | 7 |
| G | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| H | 0 | 0 | 1 | 0 | 2 | 1 | 0 | 43 | 3 | 0 | 7 | 0 | 0 | 0 | 0 | 0 |
| I | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 3 | 55 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |
| J | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| K | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| L | 1 | 3 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| M | 0 | 2 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| N | 0 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 56 |
| O | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 45 |

| Board-certified final label | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P |
|-----------------------------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

*Figure 5. Confusion matrix for the predictions of the non-reference group versus the board-certified final result*
completely or partially consistent with the board-certified labels. For the re-annotation group, there were 141 (70.5%) diagnostic results in the reference group that were completely or partially consistent with the board-certified labels, while 108 (54.0%) diagnostic results in the non-reference group were completely or partially consistent with the board-certified label. It is worth noting that in this stage, the above 16-arrhythmia-type contents were strictly inseparable. For example, if the output or result is atrial fibrillation & ST-T segment changes, and the board-certified final label is sinus rhythm & ST-T segment changes, then this will not be considered as the complete or partial consistency.

**DISCUSSION**

An incredibly important AI ethics debate exists: can AI replace human in decision-making? The version of this debate relating to our research can probably be phrased into the following: can the results provided by deep learning replace those by professional physicians in making clinical decisions? The results of deep learning should be used as an auxiliary tool to assist physicians make better interpretation decisions and improve diagnosis accuracy and efficiency. However, the power to dictate the final clinical conclusions should be retained in the hands of professional physicians. Therefore, the ability of deep learning models to aid physicians’ diagnoses in clinical practice is worth exploring, which is a significant point of this research.

In this paper, we demonstrated the effectiveness of automatic end-to-end 12-lead ECG interpretation, and presented a deep learning model for clinical ECG diagnosis of multiple cardiac diseases. The proposed DNN-based model for computerized 12-lead ECG interpretation was trained, verified, and tested on our large dataset, realizing the classification detection for not only multiple arrhythmias but also types of ST segment change to aid in the diagnosis of ACS-related diseases in practice. This approach can not only deal with the variable-length 12-lead ECGs by generating and processing two-dimensional images but also perform an end-to-end paradigm on the ECGs of compound types. We compared the F1 scores, mean AUC ROC score, and ROC curves of our deep learning model with the two groups of cardiologists. Using the board-certified final labels as the gold standards, the reference group performed better comprehensively, except in one or two types of the diseases. Of all the 16 types, our model performed approximately equal to or better than the cardiologist (with 4–5 years of experience in clinical ECG interpretation) in the non-reference group, which shows the promising interpretation ability of our model. Also, cardiologists in the reference group were more efficient in interpreting ECGs than those in the non-reference group.

**Table 2. Performance summary in ECG interpretation, and F1 scores on the re-annotated dataset**

| Proposed deep learning model | Reference group | Non-reference group |
|------------------------------|-----------------|---------------------|
| Model AUC ROC sensitivity specificity F1 score | Model AUC ROC sensitivity specificity F1 score | Model AUC ROC sensitivity specificity F1 score |
| 0.914 0.899 0.928 0.781 | 0.907 0.953 0.836 | 0.705 0.947 0.705 |
| 0.990 0.985 0.986 0.918 | 0.971 0.995 0.957 | 0.794 0.991 0.837 |
| 0.987 0.886 0.976 0.821 | 0.971 0.978 0.877 | 0.586 1.000 0.739 |
| 0.934 0.487 0.989 0.567 | 0.718 0.984 0.700 | 0.795 0.918 0.456 |
| 0.952 0.447 0.996 0.592 | 0.723 0.988 0.747 | 0.851 0.990 0.842 |
| 0.958 0.750 0.978 0.706 | 0.604 0.995 0.716 | 0.625 0.955 0.527 |
| 0.997 0.981 0.998 0.972 | 0.981 1.000 0.990 | 0.926 1.000 0.962 |
| 0.951 0.811 0.975 0.742 | 0.906 0.996 0.923 | 0.827 0.982 0.789 |
| 0.935 0.656 0.991 0.741 | 0.902 0.989 0.880 | 0.902 0.980 0.834 |
| 0.988 0.786 0.994 0.800 | 0.786 0.994 0.800 | 0.393 0.994 0.500 |
| 0.958 0.685 0.992 0.763 | 0.926 0.999 0.952 | 0.870 1.000 0.930 |
| 0.982 0.800 0.991 0.810 | 0.850 0.996 0.883 | 0.300 0.998 0.444 |
| 0.987 0.690 0.996 0.770 | 0.862 0.998 0.893 | 0.786 0.990 0.759 |
| 0.945 0.742 0.982 0.667 | 0.806 0.995 0.833 | 0.645 0.983 0.615 |
| 0.968 0.638 0.993 0.723 | 0.787 0.999 0.871 | 0.783 0.995 0.837 |
| 0.990 0.875 0.999 0.925 | 0.945 1.000 0.972 | 0.865 0.999 0.918 |
The research was based on the 3-year big data of the First People’s Hospital Affiliated to Shanghai Jiao Tong University, and 16 type high-regional-incidence cardiac diseases containing both arrhythmias and ACS-related diseases were involved. These diseases reflected the overall state of heart health in this specific region. Also, we researched on the accompanying diseases which can provide certain guidance for doctors and patients. Among all these 16 ECG types, a quarter of abnormal ECG types contained two abnormalities, namely sinus rhythm & left ventricular hypertrophy with ST-T segment changes, sinus tachycardia & ST-T segment changes, sinus bradycardia & T wave changes, and atrial fibrillation & ST-T segment changes. It reflected the high association inherent between diseases. This research can provide some pathogeny evidences, and this automatic end-to-end ECG classification system can help to detect...
erroneous diagnoses and improve healthcare delivery in low- and middle-income countries since the cardiologists would only concentrate on more complex cases in this system.

This research demonstrated the deep learning paradigm can effectively interpret multiple compound clinical 12-lead ECG abnormalities apart from a large number of normal ECGs, including ST-T segment changes (one possible evidence for myocardial infarction), ventricular hypertrophy, and other abnormal ECG patterns, which opened up a range of prospects for future research and clinical applications. Through our research, when using deep learning tools for ECG-assisted diagnosis, the deep learning performance could surpass experienced cardiologists with 4–5 years interpreting ECGs to an extent, and the AI model was also able to give these cardiologists significant reference and notably improve their diagnostic accuracy and efficiency. It is also worth noting that the network backbones of the deep learning can be improved by further optimization, so the cardiologists standing on the shoulders of giants (deep AI models learned from big data) will also achieve better diagnostic results.

In conclusion, we constructed a large-scale dataset containing both multiple arrhythmias and ACS-related diseases with high regional incidence, including 162,622 12-lead ECG records. Based on the dataset, we developed an end-to-end deep learning model capable of accurately interpreting 16 high-regional-incidence cardiac diseases that consist of normal sinus rhythm, sinus tachycardia, sinus bradycardia, sinus rhythm & T wave changes, sinus arrhythmia, sinus rhythm & ST-T segment changes, sinus rhythm & complete right bundle branch block, sinus rhythm & premature atrial contraction, atrial fibrillation, sinus rhythm & left ventricular high voltage, sinus rhythm & premature ventricular contraction, sinus rhythm & left ventricular hypertrophy with ST-T segment changes, sinus tachycardia & ST-T segment changes, sinus bradycardia & T wave changes, atrial fibrillation & ST-T segment changes, and paced rhythm. On the test dataset using the final conclusions given by the hospital, our deep learning approach for diagnosing cardiac abnormalities matched perfectly in 15,871 (88.216%) of the 17,991 ECGs, and the average AUC ROC score reached 0.961. On the board-certified re-annotated dataset, the deep learning model was accurate in 670 (78.5%) of 854 cases, which was better than the 73.5% accuracy of the cardiologist in the non-reference group (628 cases were accurate). With our deep learning model result label as reference, the cardiologist in the reference group achieved the 87.0% accuracy (743 cases were accurate) and the interpreting efficiency increased by around 69.9% relative to the non-reference group. Results suggest that the performance of our deep learning model-based 12-lead ECG interpretation paradigm surpasses that of Medical PhD cardiologists with 4–5 years of experience in clinical ECG interpretation, and those cardiologists with our deep learning results as reference can incredibly improve their diagnostic accuracy and efficiency. Our approach provides a solution for the future AI-aided diagnosis systems in clinical applications. Developments in this research could lead to advances in automated medical diagnosis and this makes it possible for deep learning technology to promote social progress.

Limitations of the study
This is a preliminary study using ECG data from China. Whether our method can better apply to people in more regions remains to be studied. At the same time, there are still some types of heart diseases, such as preexcitation syndrome, which are not included in the study since the number of these types in our database is limited.

It is also worth noting that in our research, the 12-lead ECG images were not real scanned but artificially generated, so whether our method can show similar results on real scanned ECG remains to be explored. In the future, research using scanned ECGs to interpret cardiac diseases also has practical significance.

STAR METHODS
Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Experimental model
ACKNOWLEDGMENTS

The authors gratefully thank the Natural Science Foundation of China (NSFC) (Grant No. 52005326) and Shanghai Municipal Science and Technology Major Project (Grant No. 2021SHZDZX0102) for the financial supports.

AUTHOR CONTRIBUTIONS

C.Q. and C.L. conceived the study. Y.L. and C.Q. performed the coding and generated and interpreted the results. J.L., Y.J., and Z.L. provided data support and L.Z. advised on the study. Y.L. and C.Q. wrote the original manuscript with input from all authors. All authors reviewed and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: July 6, 2022
Revised: August 10, 2022
Accepted: October 20, 2022
Published: November 18, 2022

REFERENCES

1. World Health Organization (2021). Cardiovascular Diseases (CVDs) [EB/OL]. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

2. Zeleznik, R., Foldyna, B., Eslami, P., Weiss, J., Alexander, J., Taron, J., Parmar, C., Alvi, R.M., Banerji, D., Uno, M., et al. (2021). Deep convolutional neural networks to predict cardiovascular risk from computed tomography. Nat. Commun. 12, 715. https://doi.org/10.1038/s41467-021-20966-2.

3. Attia, Z.I., Kapa, S., Lopez-Jimenez, F., McKie, P.M., Ladewig, D.J., Satam, G., Pelikka, P.A., Enriquez-Sarano, M., Noseworthy, P.A., Munguer, T.M., et al. (2019). Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nat. Med. 25, 70–74. https://doi.org/10.1038/s41591-018-0240-2.

4. Ulloa-Cerna, A.E., Jing, L., Pfeifer, J.M., Raghunath, S., Ruhl, J.A., Rodha, D.B., Leader, J.B., Zimmerman, N., Lee, G., Steinhubl, S.R., et al. eECHoMmend: an ECG-based machine-learning approach for identifying patients at high-risk of undiagnosed structural heart disease detectable by echocardiography. Circulation. 10.1161/CIRCULATIONAHA.121.057869.

5. Khurshid, S., Friedman, S., Reeder, C., Achille, P., Di, Diamant, N., Singh, P., Harrington, L.X., Wang, X., Al-Alusi, M.A., Sarma, G., et al. (2022). ECG-based deep learning and clinical risk factors to predict atrial fibrillation. Circulation. 145, 122–133. https://doi.org/10.1161/CIRCULATIONAHA.121.07480.

6. Libby, P., Ridker, P.M., and Maseri, A. (2002). Inflammation and atherosclerosis. Circulation. 105, 1135–1143. https://doi.org/10.1161/hc0902.104353.

7. Goto, S., Mahara, K., Beussink-Nelson, L., Ikura, H., Katsumata, Y., Endo, J., Gaggin, H.K., Shah, S.J., Itabashi, Y., MacRae, C.A., et al. (2021). Artificial intelligence-enabled fully automated detection of cardiac amyloidosis using electrocardiograms and echocardiograms. Nat. Commun. 12, 2726. https://doi.org/10.1038/s41591-021-22877-8.

8. Yao, X., Rushlow, D.R., Inselman, J.W., McCoy, R.G., Thacher, T.D., Behnken, E.M., Bernard, M.E., Rosas, S.L., Akfaly, A., Misra, A., et al. (2021). Artificial intelligence–enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. Nat. Med. 27, 815–819. https://doi.org/10.1038/s41591-021-01335-4.

9. Dahlof, B., Devereux, R.B., Kjeldsen, S.E., Julius, S., Beevers, G., de Faire, U., Fyhquist, F., Ibsen, H., Kristansson, K., Lederballe-Pedersen, O., et al. (2002). Cardiovascular morbidity and mortality in the Losartan Intervention for Endpoint reduction in hypertension study (LIFE), a randomised trial against atenolol. Lancet. 359, 995–1003. https://doi.org/10.1016/S0140-6736(02)08089-3.

10. Raghunath, S., Ulloa Cerna, A.E., Jing, L., vanMaanen, D.P., Stough, J., Hartzel, D.N., Leader, J.B., Kirchner, H.L., Stumpe, M.C., Hafez, A., et al. (2020). Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. Nat. Med. 26, 886–891. https://doi.org/10.1038/s41591-020-0870-z.

11. LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature. 521, 436–444. https://doi.org/10.1038/nature14539.

12. Shao, H., Xia, M., Han, G., Zhang, Y., and Wan, J. (2021). Intelligent fault diagnosis of rotor-bearing system under varying working conditions with modified transfer convolutional neural network and thermal images. IEEE Trans. Ind. Informatics. 17, 3488–3496. https://doi.org/10.1109/TII.2020.3005965.
13. Shoukat, M.S., Foers, A.D., Woodmansey, S., Evans, S.C., Fowler, A., and Soilleux, E.J. (2021). Use of machine learning to identify a T cell response to SARS-CoV-2. Cell Reports Med. 2, 100192. https://doi.org/10.1016/j.xcrm.2021.100192.

14. Topal, E.J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. 25, 44–56. https://doi.org/10.1038/s41591-018-0300-7.

15. Harmon, S.A., Sanford, T.H., Xu, S., Turkbey, E.B., Roth, H., Xu, Z., Yang, D., Myronenko, A., Anderson, V., Amalou, A., et al. (2020). Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. Nature. Commun. 11, 4080. https://doi.org/10.1038/s41467-020-17971-2.

16. Harmon, S.A., Sanford, T.H., Xu, S., Turkbey, E.B., Roth, H., Xu, Z., Yang, D., Myronenko, A., Anderson, V., Amalou, A., et al. (2020). Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. Nat. Commun. 11, 4080. https://doi.org/10.1038/s41467-020-17971-2.

17. Ardila, D., Kiraly, A.P., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., Tse, D., Ettemadi, M., Ye, W., Corrado, S., et al. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat. Med. 25, 954–961. https://doi.org/10.1038/s41591-019-0447-x.

18. Prosperi, M., Guo, Y., Sperini, M., Koopman, J.S., Min, J.S., He, Y., Rich, S., Wang, M., Buchan, I.E., and Bian, J. (2020). Causal inference and counterfactual prediction in machine learning for actionable health care. Nat. Mach. Intell. 2, 369–375. https://doi.org/10.1038/s42256-020-0197-y.

19. Hannun, A.Y., Rajurkar, P., Haghnami, M., Tison, G.H., Bourn, C., Turakhia, M.P., and Ng, A.Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat. Med. 25, 65+. https://doi.org/10.1038/s41591-018-0268-3.

20. Zhu, H., Cheng, C., Yin, H., Li, X., Zuo, P., Ding, J., Lin, F., Wang, J., Zhou, B., Li, Y., et al. (2020). Automatic multilead electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning: a cohort study. Lancet Digit. Heal. 2, e348–e357. https://doi.org/10.1016/S2589-7500(20)30107-2.

21. Ribeiro, A.H., Ribeiro, M.H., Paixão, G.M.M., Oliveira, D.M., Gomes, P.R., Canazart, J.A., Ferreira, M.P.S., Andersson, C.R., Macfarlane, P.W., Meira, W., Jr., et al. (2020). Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat. Commun. 11, 1760. https://doi.org/10.1038/s41467-020-15432-4.

22. Sangha, V., Mortazavi, B.J., Haimovich, A.D., Ribeiro, A.H., Brandt, C.A., Jacoby, D.L., Schulz, W.L., Krumholz, H.M., Ribeiro, A.L.P., and Khera, R. (2022). Automated multilabel diagnosis on electrocardiographic images and signals. Nat. Commun. 13, 1583. https://doi.org/10.1038/s41467-022-29153-3.

23. Al-Zaiti, S., Besomi, L., Bouzid, Z., Faramand, Z., Frisch, S., Martin-Gill, C., Gregg, R., Saba, S., Callaway, C., and Sejdic, E. (2020). Machine learning-based prediction of acute coronary syndrome using only the pre-hospital 12-lead electrocardiogram. Nat. Commun. 11, 3966. https://doi.org/10.1038/s41467-020-17804-2.

24. Yao, Q., Wang, R., Fan, X., Liu, J., and Li, Y. (2020). Multi-class Arrhythmia detection from 12-lead varied-length ECG using attention-based time-incremental convolutional neural network. Inf. Fusion 53, 174–182. https://doi.org/10.1016/j.inffus.2019.06.004.

25. Chen, Y.-J., Liu, C.-L., Tseng, V.S., Hu, Y.-F., and Chen, S.-A. (2019). Large-scale Classification of 12-lead ECG with Deep Learning. In 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), pp. 1–4. https://doi.org/10.1109/BHI.2019.8834668.

26. Zhang, D., Yang, S., Yuan, X., and Zhang, P. (2021). Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram. Science 24, 102373. https://doi.org/10.1126/sciadv.20213273.

27. Lai, C.X., Zhou, S.J., and Trayanova, N.A. (2021). Optimal ECG-lead selection increases generalizability of deep learning on ECG abnormality classification. Philos. Trans. R. Soc. A-MATH. PHYS. ENG. SCI. 379, 20210258. https://doi.org/10.1098/rsta.2020.0258.

28. Cai, W., Chen, Y., Guo, J., Han, B., Shi, Y., Ji, L., Wang, J., Zhang, G., and Luo, J. (2020). Accurate detection of atrial fibrillation from 12-lead ECG using deep neural network. Comput. Biomed. Res. 116, 103378. https://doi.org/10.1016/j.cbr.2020.103378.

29. Cai, W., Chen, Y., Guo, J., Han, B., Shi, Y., Ji, L., Wang, J., Zhang, G., and Luo, J. (2020). Accurate detection of atrial fibrillation from 12-lead ECG using deep neural network. Comput. Biomed. Res. 116, 103378. https://doi.org/10.1016/j.cbr.2020.103378.

30. Kiranyaz, S., Ince, T., and Gabbouj, M. (2016). Real-time patient-specific ECG classification by 1-D convolutional neural networks. IEEE Trans. Biomed. Eng. 63, 664–675. https://doi.org/10.1109/TBME.2015.2468569.

31. Zahid, M.U., Kiranyaz, S., Ince, T., Devecioglu, O.C., Chowdhury, M.E.H., Khandakar, A., Tahir, A., and Gabbouj, M. (2022). Robust R-peak detection in low-quality holter ECGs using 1D convolutional neural network. IEEE Trans. Biomed. Eng. 69, 119–128. https://doi.org/10.1109/TBME.2021.3088218.

32. Tuncer, T., Dogan, S., Pawlak, P., and Rajendra Acharya, U. (2019). Automated arrhythmia detection using novel hexadecimal local pattern and multilevel wavelet transform with ECG signals. Knowledge-Based Syst. 186, 104923.

33. Kaplan Berkaya, S., Uysal, A.K., Sora Gunal, E., Ergin, S., Gunal, S., and Gulmezoglu, M.B. (2018). A survey on ECG analysis. Biomed. Signal Process Control 43, 216–235.

34. Liu, Y., Qin, C., Liu, J., Jin, Y., Li, Z., and Liu, C. (2022). An efficient neural network-based method for patient-specific information involved arrhythmia detection. Knowledge-Based Syst. 190, 0001.

35. Acharya, U.R., Oh, S.L., Hagwara, Y., Tan, J.H., Adam, M., Gertych, A., and Tan, R.S. (2017). A deep convolutional neural network model to classify heartbeats. Comput. Biol. Med. 89, 389–396.

36. Jin, Y., Qin, C., Liu, J., Lin, K., Shi, H., Huang, Y., and Liu, C. (2020). A novel domain adaptive residual network for automatic atrial fibrillation detection. Knowledge-Based Syst. 203, 106122.

37. Martis, R.J., Acharya, U.R., and Min, L.C. (2018). ECG beat classification using PCA, LDA, ICA and Discrete Wavelet Transform. Biomed. Signal Process Control 8, 437–448.

38. Luz, E.J. da S., Schwartz, W.R., Camara-Chavez, G., and Menotti, D. (2016). ECG-based heartbeat classification for arrhythmia detection: a survey. Comput. Methods Programs Biomed. 127, 144–164.
STAR METHODS

KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---------------------|--------|------------|
| Deposited data      | This paper | N/A        |
| Raw ECG data        | This paper |          |
| Code for model and network training and analysis | This paper | https://jbox.sjtu.edu.cn/l/G1EPlg (password: sgbv) |

Software and algorithms

| Spyder 4.2.5        | International Spyder Community | https://github.com/spyder-ide/spyder/graphs/contributors |
| Python version 3.8.8 | Python Software Foundation | https://www.python.org/ |

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Chengjin Qin (qinchengjin@sjtu.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- ECG data reported in this paper will be shared by the lead contact upon request.
- All original code has been deposited at jbox.sjtu and is publicly available as of the date of publication. DOI is listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Experimental model

We collected a complete dataset containing 162,622 ECG records from Shanghai First People’s Hospital Affiliated to Shanghai Jiao Tong University, and we constructed a large and comprehensive ECG dataset called main-research dataset for 16 high-regional-incidence single or complex cardiac disease types, which contains 119,995 ECGs from 86,254 patients. We obtained the committee authorization for our studies and confirmed that informed consent was obtained from all subjects. Around a quarter of the 16 ECG types in the dataset contains two or more abnormalities. During the model training stage, we utilized the main-research dataset and split the processed dataset into training, validation and test datasets. Among them, the non-test dataset contains 85% of the data. The test dataset consists of the remaining 15% of the main-research dataset. It should be noted that test dataset will not be utilized in the training process at all. The ratio of the validation dataset which was used for hyperparameter tuning to the training dataset was around 17:3.

We trained a DNN to detect both multiple arrhythmias and ACS-related diseases with high regional incidence: normal sinus rhythm, sinus tachycardia, sinus bradycardia, sinus rhythm & T wave changes, sinus arrhythmia, sinus rhythm & ST-T segment changes, sinus rhythm & complete right bundle branch block, sinus rhythm & premature atrial contraction, atrial fibrillation, sinus rhythm & left ventricular high voltage, sinus rhythm & premature ventricular contraction, sinus rhythm & left ventricular hypertrophy with ST-T segment changes, sinus tachycardia & ST-T segment changes, sinus bradycardia & T wave changes, atrial fibrillation & ST-T segment changes, paced rhythm.
We designed a DNN architecture which consisted of multiple stacks of blocks employing small-scale convolutional kernels. Since directly obtaining 1-dimensional ECG data of the ECG machines in practice was difficult, we herein adopted 2-dimensional ECG images as the input to this network architecture for training, validation and test in order to simplify the model application process in clinical scenarios. The parameters of our proposed DNN were tuned based on the training and validation datasets. The test dataset, which was not involved in model training, will be used to evaluate the performance of our proposed DNN model.

Subject details

We obtained the committee authorization for our studies and confirmed that informed consent was obtained from all subjects.

Reporting of the sex and gender, or both for human subjects. In the complete dataset (consisting of all 162,622 subjects), 50.39% of subjects are female, while 49.61% of subjects are male. In the main-research dataset (consisting of 119,995 subjects), 52.02% of subjects are female, while 47.98% of subjects are male.

Reporting of the age or developmental stage of subjects. In the complete dataset (consisting of all 162,622 subjects), the subjects’ average age is 53.83 with a standard deviation of 19.60. In the main-research dataset (consisting of 119,995 subjects), the subjects’ average age is 50.95 with a standard deviation of 18.41.

METHOD DETAILS

Dataset acquisition

All 12-lead ECGs analyzed in this research were collected by the Shanghai First People’s Hospital Affiliated to Shanghai Jiao Tong University. The involved data was recorded from January 2018 to March 2021. The length of ECG sampling points was between 500 and 133500, and the sampling frequency was 500 Hz. In the process of establishing the database, ECG records with the length of less than 1500 and more than 15000 sampling points were excluded, in order to integrate the length of the data of the database. One of the 14 team member physicians from the hospital concluded the final ECG reports that were then reviewed. All the clinical information, ECG recording signals and final ECG reports were included in the database.

Neural network architecture

We designed a convolutional neural network with the input of 2-dimensional images as shown in Figure. All 1-dimensional ECG recordings were converted into 2-dimensional ECG images with the size of 224 × 224. The detailed description of how we process the ECG input is presented below. We employed the raw desensitized 1-dimensional ECG recordings without any denoising process. The distance between each sampling point from the 1-dimensional signal was fixed when the 2-dimensional image was generated, and the maximum length of the sampling point was 15000. When the length of the sampling point was less than 15000, the right side of the image is left blank since we have not performed any additional zero filling or other operations in order to retain the original state of the raw ECG signals. In the longitudinal directional of each image, the ECG images of lead I, II, III, aVR, aVL, aVF, V1, V2, V3, V4, V5, V6 were arranged in sequence, which was consistent with the hospital detecting ECG. The pylab module of Python was used to generate images. The ECG lines were black solid with a width of 0.2. Except for the 12-lead ECG lines, the image does not contain any other information or lines. The dpi of each image was set to 150. These 2-dimensional ECG images were the input to the neural network.
The network is elaborately designed as a 6-block deep network, in which the first two blocks contain 2 convolutional layers, the next three blocks contain 3 convolutional layers and the last block is composed of 3 fully connected layers which employs ReLU activation function and dropout. The outputs of the first five blocks are using Maxpooling for dimensionality reduction. The convolution kernel size is 3 × 3. The representation of the convolution kernel number $n$ in the first five blocks ($n = 1, 2, 3, 4, 5$) is:

$$num_n = \begin{cases} 64 \times 2^{n-1}, & n = 1, 2, 3, 4. \\ 512, & n = 5. \end{cases}$$  \hspace{1cm} (Equation 1)

The fully connected layer of the network includes 2048 neurons. Neural network weights are initialized with Xavier and the bias are initialized to 0. Adam optimizer with learning rate 0.0001 is utilized, and the CrossEntropy loss function is employed. The training process ran for 80 epochs, and we selected the model with the best validation results during optimization as the final model.

**Our deep learning model architecture.** The neural network architecture used for ECG classification

The network is elaborately designed as a 6-block deep network, in which the first two blocks contain 2 convolutional layers, the next three blocks contain 3 convolutional layers and the last block is composed of 3 fully connected layers which employs ReLU activation function and dropout. The outputs of the first five blocks are using Maxpooling for dimensionality reduction. The convolution kernel size is 3 × 3. The representation of the convolution kernel number $n$ in the first five blocks ($n = 1, 2, 3, 4, 5$) is:

$$num_n = \begin{cases} 64 \times 2^{n-1}, & n = 1, 2, 3, 4. \\ 512, & n = 5. \end{cases}$$  \hspace{1cm} (Equation 1)

The fully connected layer of the network includes 2048 neurons. Neural network weights are initialized with Xavier and the bias are initialized to 0. Adam optimizer with learning rate 0.0001 is utilized, and the CrossEntropy loss function is employed. The training process ran for 80 epochs, and we selected the model with the best validation results during optimization as the final model.

**QUANTIFICATION AND STATISTICAL ANALYSIS**

**Performance indicators of test results**

Model performance is assessed through the overall prediction accuracy, sensitivity, specificity and F1 score. The confusion matrix is used to evaluate whether the prediction results are inconsistent with the labels. We use the overall prediction accuracy to evaluate the agreement between the predicted labels and the labels of the physician’s diagnosis conclusion, according to the following expression:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%.$$  \hspace{1cm} (Equation 2)

Precision, sensitivity, specificity and F1 score of each ECG type are related to true positive (TP), true negative (TN), false positive (FP) and false negative (FN), with formulas as follows:

$$\text{Precision} = \frac{TP}{TP + FP}.$$  \hspace{1cm} (Equation 3)
Sensitivity = \frac{TP}{TP + FN} \quad \text{(Equation 4)}

Specificity = \frac{TN}{TN + FP} \quad \text{(Equation 5)}

F1 score = \frac{2 \times Precision \times Sensitivity}{Precision + Sensitivity} \quad \text{(Equation 6)}

We also plotted the receiver operating characteristic (ROC) curves and calculated the area under the curve (AUC) to assess the approach interpretation ability for each ECG type. The curve shows the relationship between sensitivity and specificity.

Re-annotation procedure
We selected a group of ECGs from the above-mentioned test dataset (C_{TEST}) to construct a new dataset called re-annotated dataset, C_{RAT}. It is obvious that the patients’ data in the new dataset C_{RAT} do not overlap with those in the training dataset, since test dataset and training dataset are mutually exclusive, and

C_{RAT} \subset C_{TEST}. \quad \text{(Equation 7)}

The ECGs of the dataset C_{RAT} were re-annotated by a committee of five scholars studying on heart disease-related research. The ECGs were also discussed by a committee of cardiologists and the final label was given by a final consensus.

To further assess the performance of our deep learning model and how deep learning can aid clinical diagnosis, we set two groups with contrast conditions during the re-annotation process as the reference group and the non-reference group: one group of scientists was provided with our deep learning model results for reference during the re-annotation, and the other group was not provided with any reference. In reference group, scientists were asked to evaluate whether our deep learning model reference was helpful to the final diagnosis. Specifically, we provided the scientists in the reference group with the prediction results of the deep learning model for each recording, and they can fully accept the result, partially accept the result, or not accept the result. We performed this work to analyze the extent to which deep learning could improve the diagnostic accuracy and efficiency of medical physicians in future clinical practice. The reference group and the non-reference group were respectively composed of three scientists, one was the annotator, and the other two were responsible for proofreading, recording and questioning. It is worth noting that the two groups of annotators have similar medical work backgrounds and experiences: they are Medical PhD in cardiovascular medicine with 4–5 years of clinical experience in interpreting ECG working in top cardiology hospitals in China.

The re-annotation process was carried out in two stages. In the first stage, two groups of annotators were annotating simultaneously and were asked to provide the time cost required to annotate the ECG. In the second stage, they discussed the ECGs that were inconsistently annotated in the first stage, and were asked to provide an agreed conclusion. Annotated ECG records that cannot be realized a consensus view will be deleted.

Sample selection for complex ECGs
We randomly selected 200 records in the remaining 41,744 records of the complete dataset to evaluate the model interpretation ability in the face of complex situations. These 41,744 records include ECGs with severe interfering or zero signals, three or more arrhythmias and other arrhythmia types not included in the above 16 types. At the same time, clinicians of our cooperated hospital sometimes gave specific disease description as the diagnosis label, such as old inferior myocardial infarction (If mapped into our model, the final output may include ST-T segment changes.). This phenomenon can be seen as different descriptions for the same phenomenon. In this stage, records related to this phenomenon were processed with uniform labels by the two groups of the annotators. All the ECG records of these types will appear in the 200 ECG records.