CLINICAL STUDY

Explainable Artificial Intelligence Model for Diagnosis of Atrial Fibrillation Using Holter Electrocardiogram Waveforms

Hirohisa Taniguchi, MD, Tomohiro Takata, PhD, Mineki Takechi, PhD, Asuka Furukawa, MD, Jin Iwasawa, MD, Akio Kawamura, MD, Tadahiro Taniguchi, PhD, and Yuichi Tamura, MD

Summary

Atrial fibrillation is a clinically important arrhythmia. There are some reports on machine learning models for AF diagnosis using electrocardiogram data. However, few reports have proposed an eXplainable Artificial Intelligence (XAI) model to enable physicians to easily understand the machine learning model’s diagnosis results.

We developed and validated an XAI-enabled atrial fibrillation diagnosis model based on a convolutional neural network (CNN) algorithm. We used Holter electrocardiogram monitoring data and the gradient-weighted class activation mapping (Grad-CAM) method.

Electrocardiogram data recorded from patients between January 4, 2016, and October 31, 2019, totaling 57,273 electrocardiogram waveform slots of 30 seconds each with diagnostic information annotated by cardiologists, were used for training our proposed model. Performance metrics of our AI model for AF diagnosis are as follows: sensitivity, 97.1% (95% CI: 0.969-0.972); specificity, 94.5% (95% CI: 0.943-0.946); accuracy, 95.3% (95% CI: 0.929-0.933). The area under the receiver operating characteristic curve for AF detection using our model was 0.988 (95% CI: 0.987-0.988). Furthermore, using the XAI method, 94.5 ± 3.5% of the areas identified as regions of interest using our machine learning model were identified as characteristic sites for AF diagnosis by cardiologists.

AF was accurately diagnosed and favorably explained with Holter ECG waveforms using our proposed CNN-based XAI model. Our study presents another step toward realizing a viable XAI-based detection model for AF diagnoses for use by physicians.

Key words: Convolutional neural network, Machine learning, Holter monitoring, Gradient-weighted class activation mapping

Atrial fibrillation (AF) is a clinically important arrhythmia that can cause palpitations, cerebral infarction, and heart failure. The prevalence of atrial fibrillation is increasing because a large portion of the US population is aging. Consequently, the cost of emergency consultation and treatment for atrial fibrillation is also rising. In particular, 20-40% of cerebral infarctions are caused by an unknown embolic source, and about 40% of these cerebral infarctions are related to atrial fibrillation. Nevertheless, the risk of cardioembolic cerebral infarction can be reduced via anticoagulation therapy. There are clinical guidelines indicating anticoagulation therapy, such as CHADS2 and CHADS-VASC scores. It is essential to detect atrial fibrillation before administering anticoagulation therapies.

Recently, machine learning models have been developed to interpret clinical data in a detailed manner. Also, some studies on detecting atrial fibrillation using neural networks have reported high diagnostic performance of these networks similar to that of human cardiologists. However, it is important for clinicians to understand the reason for a diagnosis made by artificial intelligence (AI) models before suggesting treatments for patients. Therefore, it is insufficient to implement traditional
black-box AI decision algorithms, especially those based on convolutional neural networks (CNNs), for such clinical applications. Thus, to enable AI as a practical diagnostic tool for clinicians, it is important to implement interpretable diagnostic models. Therefore, a technology called eXplainable AI (XAI) has been recently developed; XAI shows a visualization of the AI decision process, enabling physicians and clinicians to better understand the results obtained via machine learning models.

In light of the above, in this study, we developed an AI-based electrocardiogram (ECG) interpretation model for atrial fibrillation diagnosis using a CNN-based Holter monitor and the gradient-weighted class activation mapping (Grad-CAM) method to enable the validation of the obtained results for clinical use.

Methods

Ethics approval: This study was approved by the International University of Health and Welfare Ethical Review Board (No. 5-17-23). This approval followed the “Declaration of Helsinki” and the responsible institutional committee’s ethical standards on human experimentation. Informed consent was obtained via an online opt-out option; patients who opted-out online were excluded from the study.

Preparation of training and test data sets: This study included all patients who underwent Holter monitor ECG testing at the International University of Health and Welfare Mita Hospital between January 4, 2016, and October 31, 2019. All Holter monitor ECGs were first analyzed by a clinical technician and then read by a cardiologist to obtain diagnostic information. To prepare the training and test data sets for our machine learning system, we extracted raw ECG data, anonymized it, and annotated it with diagnostic information based on the basic rhythm in the ECG as either sinus rhythm, atrial fibrillation, or paroxysmal atrial fibrillation.

Construction of the CNN-based atrial fibrillation diagnosis model: We extracted waveform data at all time points from the annotated ECG database and sampled them at 125 Hz. ECG data were recorded for approximately 24 hours per case, and then 30-second slots of data were extracted from it. A cardiologist manually confirmed annotations of the basic cardiac rhythm for the extracted data slots. Next, to learn the morphological characteristics of the extracted ECG waveforms, the corresponding waveform data were transformed into two-dimensional image data, and the related diagnostic annotations were prepared for supervised learning.

Waveform data for atrial fibrillation from subjects with persistent atrial fibrillation and sinus rhythm, respectively, were extracted as training data for our proposed CNN-based atrial fibrillation diagnosis model. Figure 1 shows the architecture of our proposed model. A deep learning framework, called Chainer, was used in this study. Our proposed CNN model primarily consists of two convolutional layers and two fully connected layers. We also included normalization layers (such as local response normalization and dropout) in our model to avoid overfitting.

Model validation and statistical analysis: To evaluate our constructed machine learning diagnosis model’s performance, we performed a validation study using a validation data set. Data from cases that were not used for model training were randomly selected to construct the validation data set. Like the training data, the validation data were composed of 30-second single-lead ECG data annotated with diagnostic information subsequently converted to image data. The validation data were inputted to the constructed machine learning model, and the outputs for atrial fibrillation diagnosis were evaluated based on the annotations. We considered 95% confidence intervals (CI) to evaluate the performance of our model. All statistical analyses were performed using the IBM SPSS Statistics software version 25 or higher.

Grad-CAM: Grad-CAM can be used to produce a visual prediction map to highlight important regions in an image. The key principle on which Grad-CAM is based can be described as follows: For a given input image, we consider channel $k$ of a feature map $A^i$ that is outputted by a convolutional layer and compute the gradient of the score
Table. Performance Evaluation of the Machine Learning Model for Atrial Fibrillation Detection

| The number of predictions for Af by the machine learning model | The number of subjects with |
|---------------------------------------------------------------|----------------------------|
| Af rhythm | non Af rhythm |
| Af rhythm | 18079 | 2135 |
| Non Af rhythm | 543 | 36516 |

Af indicates atrial fibrillation.

Results

Preparation of training and validation data sets from the original database: ECG waveforms from 1883 cases of Holter ECG at our institution between January 4, 2016, and October 31, 2019, were included in our study. All waveform data were stored in a database for model training; these waveform data included 24-hour Holter ECGs annotated with diagnostic information for arrhythmia by cardiologists. We sequentially inputted ECG data of patients with sinus rhythm and atrial fibrillation into a CNN for training from this data set. We continued training the CNN with 45 patients, each with sinus rhythm and atrial fibrillation, leading to a total of 255,478 slots of waveform data.

Furthermore, we evaluated the performance of our CNN-based machine learning model using validation data. In particular, 57,273 slots of waveform data from the database were used to validate the performance of the model; these data were extracted from 10 patients with sinus rhythms, 5 with persistent atrial fibrillation, and 5 with paroxysmal atrial fibrillation. Of these 57,273 slots, 18,622 slots were annotated as atrial fibrillation data, wherein 14,400 and 4,222 slots were for persistent and paroxysmal atrial fibrillation, respectively. The remaining 38,651 slots included non-atrial fibrillation data annotated as sinus rhythm waveforms, with 28,757 and 9,894 slots belonging to subjects with sinus rhythm and those with non-transient rhythms (paroxysmal atrial fibrillation), respectively (Supplemental Figure 1).

Validation of the proposed machine learning model for atrial fibrillation diagnosis: To validate our proposed machine learning model’s performance, we inputted the validation data into the model to obtain atrial fibrillation diagnoses. Table lists the corresponding performance results. The performance metrics obtained for our machine learning model for atrial fibrillation are as follows: sensitivity: 97.1% (95% CI: 0.969-0.972); specificity: 94.5% (95% CI: 0.943-0.946); accuracy: 95.3% (95% CI: 0.952-0.955); positive predictive value: 89.3% (95% CI: 0.892-0.897); and F-value: 93.1% (95% CI: 0.929-0.933). Figure 2 shows the receiver operating characteristics (ROC) curve for atrial fibrillation diagnoses using our machine learning model. The area under the ROC curve (AUC) for atrial fibrillation detection using validation data was 0.988 (95% CI: 0.987-0.988), which indicates the high discrimination ability of our model for the detection of atrial fibrillation.

XAI images of the focused lesions obtained via Grad-CAM: To obtain XAI images, we converted the waveform data of atrial fibrillation and non-atrial fibrillation to image data. Then we evaluated using Grad-CAM to obtain regions of interest, which were then superimposed on the
image data. In particular, typical heat map images superimposed over the original waveform images are shown in Figure 3A.

Based on the regions of interest highlighted via Grad-CAM, it can be deduced that our proposed CNN focuses on fibrillation waves (f-waves) and irregular QRS waveforms in the case of atrial fibrillation. In contrast, in the case of sinus rhythm, the proposed CNN focuses on the presence of P-waves and on the regularity of QRS waveforms; however, it is noteworthy that the regions of interest in the case of sinus rhythm waveforms were smaller compared with those in the case of atrial fibrillation waveforms.

Grad-CAM was able to highlight areas of interest for diagnosis. The results of the visualization process provide key diagnostic features to clinicians. The XAI model can become a powerful diagnostic tool for clinicians by presenting the diagnosis itself and the interpretable process from the CNN (Figure 3B). In the meantime, typical false-positive and false-negative waveforms illustrate the examples that this AI suffers poor performance for diagnosis (Supplemental Figure 2), which were mainly caused by fine shakings of baselines.

**Validity assessments of XAI images obtained via Grad-CAM by cardiologists:** To assess the validity of the XAI images obtained via Grad-CAM, new atrial fibrillation waveform data that was not previously used for training were randomly extracted from the original database. The obtained Grad-CAM heat map images were superimposed on the ECG waveforms for 500 consecutive slots that were correctly diagnosed by the proposed AI model and were then evaluated by six cardiologists who were not involved in the original annotation process. The cardiologists identified 94.5 ± 3.5% of the area highlighted via Grad-CAM as characteristic sites for atrial fibrillation. Most of the irrelevant areas highlighted via Grad-CAM included drifted baselines, noisy waveforms, or small heat map displays, as shown in Supplemental Figure 3.

**Discussion**

In this study, our machine learning model, based on Holter ECG waveform data, showed acceptable accuracy for atrial fibrillation diagnosis. It also established an interpretation model using Grad-CAM as an XAI approach; our results indicated that the features assumed to be considered by our proposed CNN for the diagnosis of atrial fibrillation matched those perceived by cardiologists with high accuracy.

In particular, the visualization of the regions of interest considered by our CNN-based machine learning model for atrial fibrillation using Grad-CAM is an important feature of this study. First, the diagnostic accuracy of our machine learning model for atrial fibrillation presented significant improvement. Then, we showed that the machine learning model could suggest f-waves and irregularities in RR intervals with an accuracy similar to that of human cardiologists. Thus far, detection algorithms have not been quite successful at detecting P-waves than detecting QRS waves due to the former’s small amplitude. As an alternative method for P-wave detection in previous works, researchers attempted to locate P-waves based on their positional relationship with QRS waves by detecting the maximum and minimum points of the second derivative curve of a p-wave. In a previous study, Suzuki et al. reported that the correct response rate for their automatic atrial fibrillation diagnosis algorithm using Holter ECG data was about 80%. In contrast, in the present study,
the detection performance of our CNN-based algorithm for atrial fibrillation diagnosis was 95.3%, 97.1%, and 94.5% in terms of accuracy, sensitivity, and specificity. Also, the positive diagnosis rate of our algorithm was higher than that proposed by Suzuki et al.22) Thus, our proposed model enables the detection of atrial fibrillation in ECG data, which has been difficult to achieve using existing diagnostic methods; furthermore, our results indicate that the use of CNN’s can provide a breakthrough in terms of detection accuracy compared with conventional automatic ECG diagnosis approaches.

Another important feature of this study is an explainable and interpretable model for atrial fibrillation diagnosis. The “black-box” that exists from input to output in the diagnosis process using AI-based methods poses a problem for physicians in understanding the results and the diagnosis process using AI-based methods poses a problem for physicians in understanding the results and consequently hampers the reliability of those results.23) The most important advantage of using XAI in this study is that the XAI system can be used to visually highlight important diagnostic features in the ECG waveforms that primarily contribute to the prediction result of our machine learning model. By identifying the regions of interest in the ECG waveforms considered by the machine learning model for atrial fibrillation diagnosis, physicians will understand the output results of the model better, thus improving the reliability of the diagnosis results.

**Limitations of the study:** First, our XAI framework does not necessarily reflect the internal analysis performed by the machine learning model; on the contrary, it is just an approximation of those analysis results. Thus, we can identify the parts of the ECG that were considered important in the machine learning model using Grad-CAM, but we cannot determine the reason for their selection or the analysis methodology employed by the model.

Second, this study was a retrospective study, and our results have not been prospectively validated; thus, our proposed machine learning model needs to be validated in a prospective observational study. Also, the effect of ethnic differences in the ECG waveforms cannot be entirely ignored because the participants in this study only included individuals from the Asian ethnicity. As future work, we will improve our CNN-based atrial fibrillation diagnosis model’s robustness using training data obtained from diverse groups of subjects.

**Conclusion**

In conclusion, atrial fibrillation was accurately diagnosed from Holter ECG waveforms using our proposed CNN-based machine learning model. Furthermore, Grad-CAM enabled the visualization of the CNN diagnosis approach, thus providing an explanation model to better understand the working of the CNN diagnoses model, which has not been attempted in any similar study to the best of our knowledge.

**Acknowledgments**

The authors would like to thank Rika Takeyasu and Yui Shiga for data collection.

**Disclosure**

**Conflicts of interest:** Hirohisa Taniguchi, Tomohiro Takata and Mineki Takechi are shareholders of CardioIntelligence, Inc. Yuichi Tamura is the founder and a shareholder of CardioIntelligence, Inc.

**Author contribution:** Hirohisa Taniguchi corrected data. Tomohiro Takata performed the study analyses. Yuichi Tamura conducted the study. All authors contributed to writing and revision of the manuscript.

**Data availability:** The data that support the findings of this study was used under license for the current study, therefore it is not publicly available. On the other hand, data excluding programming code is available from the authors upon reasonable request and with permission of IUHW legal department and Cardio Intelligence Inc.

**References**

1. Wolf PA, Abbott RD, Kannel WB. Atrial fibrillation as an independent risk factor for stroke: the Framingham Study. Stroke 1991; 22: 983-8.
2. Prabhia S, Voskoboinik A, Kaye DM, Kistler PM. Atrial fibrillation and heart failure — cause or effect? Heart Lung Circ 2017; 26: 967-74.
3. Rozen G, Hosseini SM, Kaadan ML, et al. Emergency department visits for atrial fibrillation in the United States: trends in admission rates and economic burden from 2007 to 2014. J Am Heart Assoc 2018; 7: e009024.
4. Sanna T, Diener HC, Passman RS, et al. Cryptogenic stroke and underlying atrial fibrillation. N Engl J Med 2014; 370: 2478-86.
5. Ntaios G, Papavasileiou V, Milionis H, et al. Embolic strokes of undetermined source in the Athens Stroke Registry: an outcome analysis. Stroke 2015; 46: 2087-93.
6. Perera KS, Vanassche T, Bosch J, et al. Embolic strokes of undetermined source: prevalence and patient features in the ESUS Global Registry. Int J Stroke 2016; 11: 526-33.
7. Granger CB, Alexander JH, McMurray JJV, et al. Apixaban versus warfarin in patients with atrial fibrillation. N Engl J Med 2011; 365: 981-92.
8. Patel MR, Mahaffey KW, Garg J, et al. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. N Engl J Med 2011; 365: 883-91.
9. Giugliano RP, Ruff CT, Braunwald E, et al. Edoxaban versus warfarin in patients with atrial fibrillation. N Engl J Med 2013; 369: 2093-104.
10. Agarwal S, Hachamovitch R, Menon V. Current trial-associated outcomes with warfarin in prevention of stroke in patients with nonvalvular atrial fibrillation: a meta-analysis. Arch Intern Med 2012; 172: 623-31; discussion 631.
11. Gage BF, Waterman AD, Shannon W, Boechler M, Rich MW, Radford MJ. Validation of clinical classification schemes for predicting stroke: results from the National Registry of atrial fibrillation. JAMA 2001; 285: 2864-70.
12. Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest 2010; 137: 263-72.
13. Kwon S, Hong J, Choi EK, et al. Deep learning approaches to detect atrial fibrillation using photoplethysmographic signals: algorithms development Study. JMRI MHealth UHealth 2019; 7: e12770.
14. Poh MZ, Poh YC, Chan PH, et al. Diagnostic assessment of a deep learning system for detecting atrial fibrillation in pulse waveforms. Heart 2018; 104: 1921-8.
15. Ivanovic MD, Atanasoski V, Shvilkin A, Hadzievski L, Ma-
luckov A. Deep learning approach for highly specific atrial fibrillation and flutter detection based on RR intervals. Annu Int Conf IEEE Eng Med Biol Soc 2019; 2019: 1780-3.

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

17. Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): towards medical XAI. IEEE Trans Neural Netw Learning Syst 2019. Available at: https://128.84.21.199/abs/1403.3301v1. Accessed July 17, 2019 (in press).

18. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 2020; 58: 82-115.

19. Panigrahy D, Sahu PK. P and T wave detection and delineation of ECG signal using differential evolution (DE) optimization strategy. Australas Phys Eng Sci Med 2018; 41: 225-41.

20. Maršánová L, Němcová A, Směšek R, Vítek M, Smítal L. Advanced P wave detection in ecg signals during pathology: evaluation in different arrhythmia contexts. Sci Rep 2019; 9: 19053.

21. Suzuki T, Hashimoto T. Automated diagnosis of paroxysmal atrial fibrillation using ambulatory electrocardiography. J Showa Med Assoc 1997; 57: 139-48.

22. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov 2019; 9: e1312.

Supplemental Files

Supplemental Figures 1-3

Please see supplemental files; https://doi.org/10.1536/ihj.21-094