

Artificial intelligence using electrocardiography: strengths and pitfalls

Joon-myung Kwon ^{1,2,3,4,*†}, Yong-Yeon Jo ^{1†}, Soo Youn Lee,^{2,5} and Kyung-Hee Kim ^{2,5}

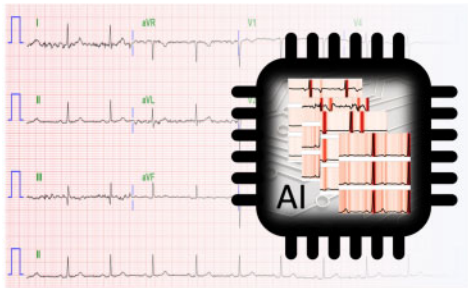
¹Medical research team, Medical AI Co., Seoul, South Korea; ²Artificial Intelligence and Big Data Research Center, Sejong Medical Research Institute, Bucheon, South Korea;

³Department of Critical Care and Emergency Medicine, Mediplex Sejong Hospital, Incheon, South Korea; ⁴Medical R&D center, Bodyfriend Co., Seoul, South Korea; and

⁵Division of Cardiology Cardiovascular Center, Mediplex Sejong Hospital, Incheon, South Korea

online publish-ahead-of-print 22 March 2021

This editorial refers to ‘Electrocardiogram screening for aortic valve stenosis using artificial intelligence’, by M. Cohen-Shelly et al., doi:10.1093/eurheartj/ehab153.



Recent studies related to artificial intelligence using ECG

Diagnosis

- Left ventricular systolic dysfunction
- Heart failure with preserved ejection fraction
- Aortic valve stenosis
- Mitral valve regurgitation
- Pulmonary hypertension
- Left ventricular hypertrophy
- Myocardial infarction with or without ST elevation
- Arrhythmia
- Hyperkalemia
- Anemia

Prediction

- Paroxysmal atrial fibrillation
- Patient deterioration and cardiac arrest
- Aortic valve stenosis
- Heart failure with preserved ejection fraction
- Mitral valve regurgitation

Graphical abstract Recent studies related to artificial intelligence using ECG.

Artificial intelligence (AI) is being applied in various fields of cardiology. In particular, deep learning (DL), a subset of machine learning (ML) in AI, enables the diagnosis and prediction of cardiac diseases using neural networks with more neurons at their layers as well as their interconnectivity. The primary advantage of DL is its ability to discover features of certain data that cannot be discovered from a human perspective.¹

Conventional ML models require meticulous feature engineering with domain expertise to derive features from images or signals for their input. Meanwhile, DL automatically discovers representations and extracts the features from raw data. Therefore, DL requires minimal engineering by hand for development, and it is not restricted by human prejudice when extracting features from data.

† These authors contributed to this work.

The opinions expressed in this article are not necessarily those of the Editors of the *European Heart Journal* or of the European Society of Cardiology.

* Corresponding author. Artificial Intelligence and Big Data Research Center, 20 Gyeongmunhwa-ro, 21080 Incheon, Republic of Korea. Tel: +82 32 240 8129, Fax: +82 32 240 8094, Email: kwonjm@sejongh.co.kr

© The Author(s) 2021. Published by Oxford University Press on behalf of the European Society of Cardiology.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Cohen-Shelly *et al.* developed and validated a DL model for detecting aortic stenosis (AS) using electrocardiography (ECG), and their results are published in this issue of the *European Heart Journal*.² The authors have shown that AI using ECG can identify patients with moderate or severe AS, and might be able to predict developing AS by comparing false-positive and true-negative groups in subgroup analysis. Through learning an implicit representation, the DL model is effective in discovering diverse features based on subtle changes in ECG and creating an algorithm from complex and non-linear ECG data. Since 2019, AI using ECG has been investigated to enable the diagnosis of diseases not possible through conventional ECG (Graphical Abstract). Recent studies have shown that AI-enabled ECG can be used to detect heart failure, pulmonary hypertension, hyperkalaemia, and anaemia, as well as to predict the development of atrial fibrillation and cardiac arrest.^{3–8} Various technologies based on DL, such as the generation of precordial six-lead ECGs from limb six-lead ECGs, are being introduced to detect myocardial infarction.⁹

DL enables a model to be created using only data, i.e. without the restrictions of human ideas. Furthermore, new insights can be acquired by comparing findings obtained using DL from data only with existing medical knowledge. Using a saliency map from an AI technology developed, Cohen-Shelly *et al.* showed that the TP interval and U waves in the right precordial leads were weighted the most heavily for determining the presence of AS.² Typical ECG findings for left ventricular hypertrophy were not weighted in the developed AI. Although those findings and the methodology involved will inspire many researchers, further research is needed to understand the exact meaning of the former.

One limitation of DL is overfitting. Using only data without human engineering is both a disadvantage and an advantage of DL. DL is merely a method for developing an algorithm with the best accuracy limited to certain data, and the risk of overfitting exists. For example, if a DL model that identifies cats and dogs on an island is developed, where all cats are white and all dogs are black, then the developed DL model will distinguish cats and dogs using only both black and white features. Furthermore, the developed DL model will demonstrate poor accuracy in environments other than the island on which the model was developed. In another example, because suspicious skin lesions are often routinely marked with gentian violet surgical skin markers, Winkler *et al.* demonstrated that skin marking at the periphery of dermoscopic images was significantly associated with the DL model detection of skin cancer.¹⁰ Therefore, to guarantee real-world performance, an external validation with isolated data from a different environment is required in all DL research studies.

An external validation implies performing testing using data that differ completely from those for the internal validation used to develop the AI model. In most cases of DL-based AI models, the number of parameters is significant, and occasionally exceeds the number of study subjects. For example, ResNet-152, a popular DL model with outstanding performance for image classification, comprises 60 million parameters.¹¹ Hence, the DL model might overfit the training data during internal validation; if data extracted from a certain patient belong to both training and test data for the internal validation, then the developed DL model will identify the patient rather than detecting target disease, thereby resulting in

an overestimated performance—this is not guaranteed in real-world applications.

Conducting an external validation implies not only separating data for the internal validation, but also confirming the performance for data in a different environment. Wolpert and Macready explained the ‘no free lunch’ theorem: if AI is optimized for a specific situation, then it cannot yield favourable results in a different situation.¹² For an accurate validation, the data should be split by hospital or region. Although the populations investigated by Cohen-Shelly *et al.* were from Minnesota, Arizona, and Florida, the data were mixed before they were assigned to training and validation data.² Absence of external validation might result in an overestimated performance because the training and test data were not distinctly different. Hence, further studies are needed for external validation such that the developed AI model can be applied across regions and hospitals.

The other disadvantage of DL is that, currently, it cannot unveil the DL decision process, i.e. the black box. In other words, although a DL model can be developed by fitting each coefficient, we cannot specifically interpret the decision process of the model. Based on the study by Cohen-Shelly *et al.*, although we can infer that the TP interval is important through a saliency map, characteristics of the TP interval that are related to AS could not be identified.² Moreover, we could not determine why the DL model did not use the ECG features of left ventricular hypertrophy for detecting AS. As the DL model might make an unreasonable decision, the lack of interpretability of the DL model hinders its clinical use significantly. Because the process and reason related to the wrong decision of the DL model could not be determined, we could not monitor or rectify the model risk that might cause medical errors. Because of this, a safety net is required when using DL in clinical applications. To detect critical errors of the DL model, conventional methods and DL models must be used simultaneously. For example, when we used AI-enabled ECG to screen for AS, conventional methods, such as detailed history taking, careful auscultation, and cardiologist consultation, were needed. Recently, several studies have been conducted to understand the decision-making process of DL, and explainable AI in the field of medicine would continue to evolve.¹³

AI was introduced in the 1950s; since then, two AI ‘winter’ periods of reduced funding and interest in AI research have occurred.¹⁴ These winter periods were due to disappointment from unsatisfactory real-world performances following extravagant endorsements of the idea that AI can solve all problems. Furthermore, unreasonable and unexplainable AI decisions contributed to the recurrence of these winter periods. It is clear that AI exhibits significant potential in the field of medicine; it can improve diagnostic accuracy and support clinical decisions for many diseases. However, the disadvantages of AI should be identified and efforts should be expended to overcome its limitations. This would enable us to continue developing AI technology for medical applications, e.g. AI based on DL for improving the early diagnosis and prevention of irreversible cardiovascular disease progression.

Funding

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

Conflict of interest: K.H.K. and S.Y.L. declare that they have no competing interests. J.K. and Y.Y.J. are researchers of Medical AI Co., a medical artificial intelligence company. J.K. is a researcher of Bodyfriend Co. There are no products in development or marketed products to declare.

References

1. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;**521**:436–444.
2. Cohen-Shelly M, Attia ZI, Friedman PA, Ito S, Essayagh BA, Ko WY, Murphree DH, Michelenia HI, Enriquez-Sarano M, Carter RE, Johnson PW, Noseworthy PA, Lopez-Jimenez F, Oh JK. Electrocardiogram screening for aortic valve stenosis using artificial intelligence. *Eur Heart J* 2021;**42**:2885–2895.
3. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, Pellikka PA, Enriquez-Sarano M, Noseworthy PA, Munger TM, Asirvatham SJ, Scott CG, Carter RE, Friedman PA. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med* 2019;**25**:70–74.
4. Kwon J myoung, Kim KH, Medina-Inojosa J, Jeon KH, Park J, Oh BH. Artificial intelligence for early prediction of pulmonary hypertension using electrocardiography. *J Heart Lung Transpl* 2020;**39**:805–814.
5. Galloway CD, Valys A V., Shreibati JB, Treiman DL, Petterson FL, Gundotra VP, Albert DE, Attia ZI, Carter RE, Asirvatham SJ, Ackerman MJ, Noseworthy PA, Dillon JJ, Friedman PA. Development and validation of a deep-learning model to screen for hyperkalemia from the electrocardiogram. *JAMA Cardiol* 2019;**4**:428–436.
6. Kwon JM, Cho Y, Jeon KH, Cho S, Kim KH, Baek SD, Jeung S, Park J, Oh BH. A deep learning algorithm to detect anaemia with ECGs: a retrospective, multi-centre study. *Lancet Digit Health* 2020;**2**:e358–e367.
7. Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, Carter RE, Yao X, Rabinstein AA, Erickson BJ, Kapa S, Friedman PA. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet* 2019;**394**:861–867.
8. Kwon JM, Kim KH, Jeon KH, Lee SY, Park J, Oh BH. Artificial intelligence algorithm for predicting cardiac arrest using electrocardiography. *Scand J Trauma Resusc Emerg Med* 2020;**28**:98.
9. Cho Y, Kwon JM, Kim KH, Medina-Inojosa JR, Jeon KH, Cho S, Lee SY, Park J, Oh BH. Artificial intelligence algorithm for detecting myocardial infarction using six-lead electrocardiography. *Sci Rep* 2020;**10**:20495.
10. Winkler JK, Fink C, Toberer F, Enk A, Deinlein T, Hofmann-Wellenhof R, Thomas L, Lallas A, Blum A, Stolz W, Haenssle HA. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatol* 2019;**155**:1135–1141.
11. Wu Z, Shen C, van den Hengel A. Wider or deeper: revisiting the ResNet model for visual recognition. *Pattern Recognit* 2019;**90**:119–133.
12. Wolpert DH, Macready WG. No free lunch theorems for optimization. *IEEE Trans Evol Comput* 1997;**1**:67–82.
13. Jo YY, Cho Y, Lee SY, Kwon J, Kim KH, Jeon KH, Cho S, Park J, Oh BH. Explainable artificial intelligence to detect atrial fibrillation using electrocardiogram. *Int J Cardiol* 2020; doi: 10.1016/j.ijcard.2020.11.053.
14. Hender J. Avoiding another AI winter. *IEEE Intell Syst* 2008;**23**:2–4.