A Counterfactual-based Approach for Interpreting Deep Learning Models in Electrocardiogram Analysis

Hak Seung Lee, MD, MS[1,2], Yeji Lee, MD[1,3*], Jong-Hwan Jang, PhD[1*], Min Sung Lee, MD, MS[1,2], Yong-Yeon Jo, PhD[1] and Joon-myoung Kwon, MD, MS[1,2]*

[1]Medical AI Co., Ltd. Seoul, South Korea

[2]Artificial Intelligence and Big Data Research Center, Sejong Medical Research Institute, Bucheon, South Korea

[3]Columbia University Mailman School of Public Health, New York, United States

** these authors contributed equally*

Background: A deep learning shows outstanding performance in various medical fields, including electrocardiogram (ECG) analysis. Model interpretation, however, still relies on gradient-based methods such as Grad-CAM, which visualizes influential input regions as heatmaps. Researchers qualitatively evaluate the reasonableness of these heatmaps. However, in ECG analysis, assessing validity of heatmaps is challenging due to the lack of morphological finding on a target disease. If the interpretation results could reveal insights regarding both where the influential regions are and what morphological aspects affect the prediction, it would significantly enhance the model's reliability and applicability in clinical practice.

Methods: We suggested applying a counterfactual-based interpretation method to ECG classifier to gain morphological insights. A counterfactual-based method is a technique which can explain the model's decision-making rationale by contrasting the generated ECGs with different probability. This method requires to develop ECG generator using generative adversarial neural network structure. When given a condition probability, the generator can generate ECGs that are predicted as the given probability by the classifier. As a proof-of-concept, we applied the proposed method to a classifier for hyperkalemia, since previous studies have revealed a lot of evidence for morphological ECG patterns. This allows us to show the validity and utility of our method by comparing its results with established clinical knowledge. We utilized a hyperkalemia binary classifier with an AUROC of 0.95 and AUPRC of 0.3 on a dataset with a prevalence of 2% (N=300,000). The potassium level distribution for data with non-hyperkalemia and hyperkalemia labels has median values of 4.1 mmol/L (range: 3.8-4.4 mmol/L) and 6.0 mmol/L (range: 5.7-6.3 mmol/L), respectively. We used the generator to create ECGs that the classifier predicted with a probability of 0.9 and transformed them into counterfactual ECGs with lower probabilities to analyze morphological changes.

Results: We show results using both Grad-CAM and our method simultaneously for model interpretation. Grad-CAM visualizes where the model deems important, while our method provides information on what the model considers significant, thereby offering more concrete evidence in ECG analysis. When examining generated ECG with a 0.9 probability through Grad-CAM, the QRS complex and T wave were the main factors contributing to the decision. In a figure, analyzing the counterfactual ECG lead II with probability 0.1 revealed that the classifier relied on morphological changes such as the flattening and disappearance of P waves, the widening of QRS complexes accompanied by a diminished R wave amplitude, minimal ST segment depression, and the presence of prominent, peaked T waves. These findings are consistent with previous ECG research on hyperkalemia, supporting the reasonableness of the classifier's decision-making process and the potential of counterfactual-based method.

Conclusion: We suggested a counterfactual-based approach for interpreting deep learning models in ECG analysis. By generating counterfactual ECG, it can provide more explicit and concrete interpretation. Comparing results of hyperkalemia case with established clinical findings, it enables clinicians to gain clinically relevant insights and evaluate the reasonableness of the model's decisions. Suggested approach serves solely as a method to interpret the decision of a classifier, and the clinical validation of this interpretation remains within a rigorous clinical investigation. Nevertheless, counterfactual-based approach has the potential to offer a foundation and valuable insights for subsequent ECG research. Moreover, this is not confined to hyperkalemia applications; it is a flexible methodology applicable to various domain and multi-class classifiers. Future work will involve validating the proposed method in other ECG-related tasks and exploring its potential in other medical domains.