

Counterfactual Explanations for Myocardial Infarction Classification from ECG Signals

Mario Refoyo, Sandra Calderón, David Luengo

Universidad Politécnica de Madrid, Spain

Machine Learning/Deep Learning (ML/DL) methods, which constitute the algorithmic basis of Artificial Intelligence (AI), are increasingly used in biomedicine. In electrocardiography, ML/DL techniques are used to perform ECG delineation and segmentation, diagnose pathologies, and predict future heart conditions. Unfortunately, AI methods are typically black-box in nature, meaning that the way in which they take decisions is not transparent for physicians. In order to build trust in AI and enhance the use of ML/DL methods for critical medical tasks, the field of eXplainable AI (XAI) attempts to develop either intrinsically interpretable models or *post-hoc* explainability algorithms.

In this work, we adapt a *post-hoc* XAI method developed by the authors, Multi-SpaCE [1], to explain the decisions of an ECG-based detector of myocardial infarction. Firstly, we train a state-of-the-art multivariate time series classifier (InceptionTime [2]) to discriminate between healthy patients and patients with myocardial infarction. Then, we use Multi-SpaCE, a multi-objective method designed to generate high-quality counterfactual explanations (i.e., valid, sparse, plausible and concentrated explanations) for multivariate time series, to explain the classifier's decisions to cardiologists.

Preliminary experimental results, using the PTB database, show the promising results of the proposed approach. First, an InceptionTime classifier is trained, obtaining an accuracy equal to 0.9536. Then, Multi-SpaCE is applied to explain all the instances in the test set. The resulting counterfactuals (CFs) are always valid (i.e., 100% of the generated CFs belong to the alternative class), showing a good plausibility (normalized score equal to 0.62 ± 0.25 obtained using an auxiliary Auto Encoder) and a high probability (0.84 ± 0.10) of belonging to the alternative class. Furthermore, the explanations are typically concentrated in 3–5 locations of the ECG, facilitating their interpretability.

References

- [1] M. Refoyo & D. Luengo (2024). “Multi-SpaCE: Multi-objective subsequence-based sparse counterfactual explanations for multivariate time series classification.” *arXiv preprint* (arXiv:2501.04009).
- [2] H. I. Fawaz *et al.* (2020). “InceptionTime: Finding AlexNet for time series classification.” *Data Mining and Knowledge Discovery* 34(6), 1936–1962.