ESC
European Society
of Cardiology

European Heart Journal - Digital Health (2022) **3**, 502–504
https://doi.org/10.1093/ehjdh/ztac063

LETTER TO THE EDITOR

Variational auto-encoders improve explainability over currently employed heatmap methods for deep learning-based interpretation of the electrocardiogram

Rutger R. van de Leur, Rutger J. Hassink, and René van Es *

Department of Cardiology, University Medical Center Utrecht, Internal ref E03.511, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands

Online publish-ahead-of-print 26 October 2022

This is a rebuttal on the earlier letter "Correspondence to the European Heart Journal-digital health in response to the paper by Attia *et al*. 2022", https://doi.org/10.1093/ehjdh/ztac053.

We appreciate the opportunity to address Higaki and Yamaguchi and their detailed commentary on our study.[1] In the referenced study, we show that variational auto-encoders (VAEs), which use deep neural networks (DNNs) to learn the underlying factors of variation in the median beat electrocardiogram (ECG), can be used to provide *improved explainability* over previous attempts to open the 'black box' of ECG-based DNNs using saliency-based heatmaps. There are currently conflicting definitions of explainability and interpretability in the literature and both are used interchangeably. In this work, explainability refers to the concept of providing insight into *why* the algorithm makes a certain decision. Interpretability, on the other hand, refers to *how* the algorithm decides, by providing a direct relation between predictor and outcome.[2]

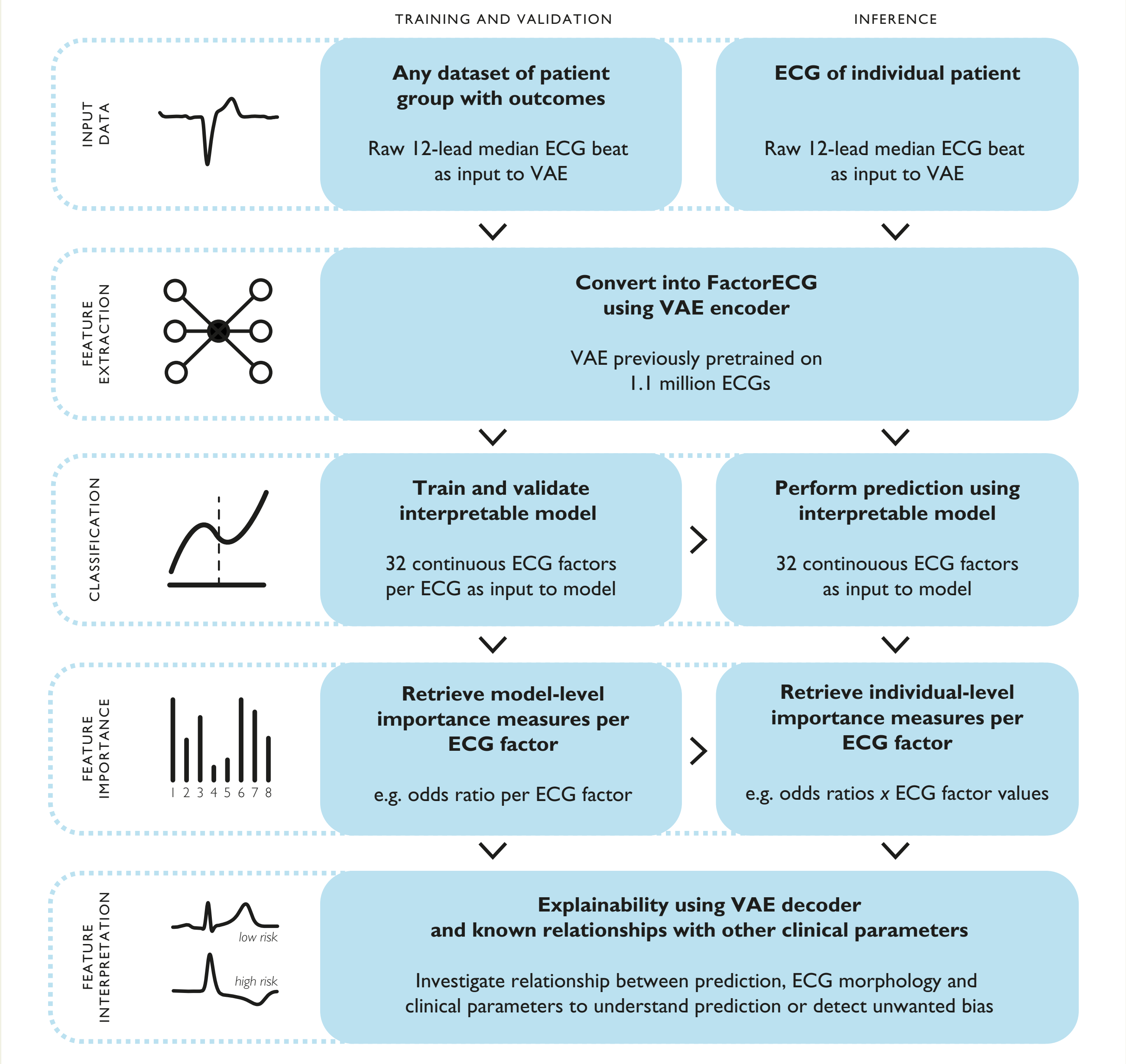
Currently employed explainability techniques for ECGs are usually saliency-based heatmaps, but these techniques have shown to be unreliable and poorly reproducible. For example, Adebayo *et al.*[3] have shown that even untrained DNNs provide heatmaps that look reassuring. Moreover, Hooker *et al.*[4] have shown that when you remove the regions deemed important by many saliency-based methods, performance of the classifier does not decrease after retraining. Our own preliminary experiments have shown similar results for ECGs.

Even when saliency-based methods produce reliable results, the heatmap can only point at temporal locations in the ECG, which does not provide enough explainable value. For example, a highlighted terminal T-wave could mean the QT interval, the T-wave height, the T-wave morphology, or something else.[5] Some researchers have tried to overcome this by entering two-dimensional images of the ECG into the DNN and applying the heatmap on the image.[6] Although this may add some 'voltage-related' information, it will still not provide information on the exact morphology of that feature.

Lastly, next to the individual explanations of decisions by the model, some form of model-level explainability is necessary to gain insight into the overall decision-making process of the model. Especially in big datasets, it is not feasible to inspect all individual heatmaps. Although there have been attempts to translate the individual heatmaps to complete datasets, for example by taking the mean, model-level explainability remains unsatisfactory.[7] A lack of model-level explainability poses the risk of confirmation bias: when there are many possible individual explanations for your complex model, will you just pick the ones that confirm your hypothesis?[2] Many papers show only some example ECGs with their respective heatmaps, and draw conclusions from these examples alone about the workings of the algorithm.[8]

In our study, we demonstrated *improved explainability* over heatmap-based methods for these three major limitations. This is done by intentionally decoupling feature discovery from classification in DNNs using a β-VAE to decompose the ECG into its generative factors (the FactorECG). By combining these learned *explainable* factors with standard *interpretable* models (such as logistic regressions) in a pipeline, we are able to create a fully explainable pipeline (*Figure 1*). This approach greatly improves reproducibility and reliability, as a pretrained VAE will always produce the same FactorECG for a given ECG. Moreover, we are able to show actual changes to ECG morphology instead of just a temporal location in the ECG by using visual inspection of the factor traversals. In the current analysis, we provide additional insight into the factors by showing relationships

* Corresponding author. Tel: +0031 88 757 3453, Fax: +0031 88 757 3453, Email: r.vanes-2@umcutrecht.nl



TRAINING AND VALIDATION
INFERENCE
INPUT DATA
Any dataset of patient group with outcomes
Raw 12-lead median ECG beat as input to VAE
ECG of individual patient
Raw 12-lead median ECG beat as input to VAE
FEATURE EXTRACTION
Convert into FactorECG using VAE encoder
VAE previously pretrained on 1.1 million ECGs
CLASSIFICATION
Train and validate interpretable model
32 continuous ECG factors per ECG as input to model
Perform prediction using interpretable model
32 continouous ECG factors as input to model
FEATURE IMPORTANCE
1 2 3 4 5 6 7 8
Retrieve model-level importance measures per ECG factor
e.g. odds ratio per ECG factor
Retrieve individual-level importance measures per ECG factor
e.g. odds ratios x ECG factor values
FEATURE INTERPRETATION
low risk
high risk
Explainability using VAE decoder and known relationships with other clinical parameters
Investigate relationship between prediction, ECG morphology and clinical parameters to understand prediction or detect unwanted bias

Figure 1 Overview of the novel explainable pipeline. During training and validation (left), any electrocardiogram dataset with available outcomes can be used as input to the pipeline. As the electrocardiograms are converted to only 32 factors, datasets can be relatively small. The variational auto-encoder, which was pretrained on 1.1 million electrocardiograms, is subsequently used to convert every single electrocardiogram into its FactorECG (32 continuous values that represent that electrocardiogram). These 32 factors per electrocardiogram are used in the next step to train an interpretable statistical model for diagnosis or prediction, such as logistic regression. As these models are inherently interpretable, they provide importance values, such as odds ratios, for every electrocardiogram factor individually. As we are able to visualize the influence of the individual electrocardiogram factors on the electrocardiogram morphology using the variational auto-encoder decoder, a direct relationship between electrocardiogram morphology and the prediction can be obtained on the model-level. During inference (right), an individual electrocardiogram can be entered into the pretrained variational auto-encoder. Prediction is performed using the previously trained interpretable, and individual-level importance measures per electrocardiogram factor are obtained. These individual importance measures can subsequently be related to the electrocardiogram morphologies and correlates of each factor, to better understand why the algorithm made this specific prediction. An online tool is provided for other researchers to use the FactorECG in their study (https://encoder.ecgx.ai).

with diagnoses and conventional ECG characteristics (e.g. PR interval), but using solely these characteristics does not lead to comparable performance as using the ECG factors.[9] We completely agree with Higaki and Yamaguchi, however, that associations with echocardiography or genetics are much more interesting, and this is an area of active investigation by our group.

Conversely to the suggestion of Higaki and Yamaguchi, we have designed and extensively described the employed pipeline not to hide the fact that we use simpler interpretable statistical models (such as logistic regression or extreme gradient boosting with shapley explanations) for prediction tasks, but rather as a major strength of the selected methodology. This allows establishing a direct relation between the ECG factors (and their respective influence on the ECG morphology) and the prediction on the individual and model-level, without a loss of predictive performance (*Figure 1*). When logistic regression is used, the odds ratios for each ECG factor provide model-level explainability, while for individual cases the ECG factor values of that specific ECG can be investigated in combination with the odds ratios. Furthermore, due to the dimensionality reduction, it broadens the applicability of DNNs to much smaller datasets. In recent publications, we have shown that the FactorECG is able to predict the risk of life-threatening ventricular arrhythmias in patients with dilated cardiomyopathy and success of cardiac resynchronization therapy.[9,10]

In conclusion, we show that decoupling feature extraction from classification in deep learning-based ECG analysis allows for *improved explainability* over heatmap-based methods. Our pipeline employs the power of deep learning to discover features in the median beat ECG morphology, while also enabling the use of different interpretable classification models. Our experiments show that this decoupling does not lead to a loss in predictive performance, which contradicts a longstanding assumption that the 'black box' nature of the currently applied DNNs was inevitable to achieve impressive performances. Future studies should thus focus on using such explainable pipelines, consisting of a separate feature extraction method (for example a VAE) and interpretable classification method, as they could increase trust in artificial intelligence (AI), allow for bias detection, and broaden the application of AI to many other (rare) diseases.

Data availability

The data underlying this article are available in the article.

References

1. van de Leur RR, Bos MN, Taha K, Sammani A, Yeung MW, van Duijvenboden S, Lambiase PD, Hassink RJ, van der Harst P, Doevendans PA, Gupta DK, van Es R. Improving explainability of deep neural network-based electrocardiogram interpretation using variational auto-encoders. *Eurn Heart J Digital Heal* 2022;**3**:390–404.
2. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 2019;**1**:206–215.
3. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. *Adv Neural Inf Process Syst* 2018;**31**:9505–9515.
4. Hooker S, Erhan D, Kindermans P-J, Kim B. A benchmark for interpretability methods in deep neural networks. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc.; 2019. p9737–9748.
5. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digital Heal* 2021;**3**:e745–e750.
6. Makimoto H, Höckmann M, Lin T, Glöckner D, Gerguri S, Clasen L, Schmidt J, Assadi-Schmidt A, Bejinariu A, Müller P, Angendohr S, Babady M, Brinkmeyer C, Makimoto A, Kelm M. Performance of a convolutional neural network derived from an ECG database in recognizing myocardial infarction. *Sci Rep* 2020;**10**:8445.
7. van de Leur RR, Taha K, Bos MN, van der Heijden JF, Gupta D, Cramer MJ, Hassink RJ, van der Harst P, Doevendans PA, Asselbergs FW, van Es R. Discovering and visualizing disease-specific electrocardiogram features using deep learning: proof-of-concept in phospholamban gene mutation carriers. *Circ Arrhythmia Electrophysiol* 2021;**14**.
8. Hughes JW, Olgin JE, Avram R, Abreau SA, Sittler T, Radia K, Hsia H, Walters T, Lee B, Gonzalez JE, Tison GH. Performance of a convolutional neural network and explainability technique for 12-lead electrocardiogram interpretation. *Jama Cardiol* 2021;**6**:1285–1295.
9. Sammani A, van de Leur RR, Henkens MTHM, Meine M, Loh P, Hassink RJ, Oberski DL, Heymans SRB, Doevendans PA, Asselbergs FW, te Riele ASJM, van Es R. Life-threatening ventricular arrhythmia prediction in patients with dilated cardiomyopathy using explainable electrocardiogram-based deep neural networks. *EP Europace* 2022;**24**:1645–1654.
10. Wouters P, van de Leur R, Vessies M, van Stipdonk A, Ghossein MA, Maass AH, Prinzen FW, Vernooy K, Meine M, Es RV. PO-658–01 explainable deep learning outperforms guideline criteria for prediction of cardiac resynchronization therapy outcome. *Heart Rhythm* 2022;**19**:S274–S275.