




Improving explainability of deep neural network-based electrocardiogram interpretation using variational auto-encoders

Rutger R. van de Leur^{1,2,†}, Max N. Bos^{1,3,†}, Karim Taha^{1,2}, Arjan Sammani ¹, Ming Wai Yeung ⁴, Stefan van Duijvenboden⁵, Pier D. Lambiase⁵, Rutger J. Hassink¹, Pim van der Harst ¹, Pieter A. Doevendans^{1,2,6}, Deepak K. Gupta³, and René van Es^{1,*}

¹Department of Cardiology, University Medical Center Utrecht, Internal ref E03.511, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands; ²Netherlands Heart Institute, Moreelsepark 1, 3511 EP Utrecht, The Netherlands; ³Informatics Institute, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands; ⁴Department of Cardiology, University Medical Center Groningen, Hanzeplein 1. 9713 GZ Groningen, The Netherlands; ⁵Institute of Cardiovascular Science, University College London, 62 Huntley St, London Wc1E 6Dd, UK; and ⁶Central Military Hospital, Lundlaan 1, 3584 Utrecht, The Netherlands

Received 22 February 2022; revised 16 June 2022; online publish-ahead-of-print 25 July 2022

Aims

Deep neural networks (DNNs) perform excellently in interpreting electrocardiograms (ECGs), both for conventional ECG interpretation and for novel applications such as detection of reduced ejection fraction (EF). Despite these promising developments, implementation is hampered by the lack of trustworthy techniques to explain the algorithms to clinicians. Especially, currently employed heatmap-based methods have shown to be inaccurate.

Methods and results

We present a novel pipeline consisting of a variational auto-encoder (VAE) to learn the underlying factors of variation of the median beat ECG morphology (the FactorECG), which are subsequently used in common and interpretable prediction models. As the ECG factors can be made explainable by generating and visualizing ECGs on both the model and individual level, the pipeline provides improved explainability over heatmap-based methods. By training on a database with 1.1 million ECGs, the VAE can compress the ECG into 21 generative ECG factors, most of which are associated with physiologically valid underlying processes. Performance of the explainable pipeline was similar to 'black box' DNNs in conventional ECG interpretation [area under the receiver operating curve (AUROC) 0.94 vs. 0.96], detection of reduced EF (AUROC 0.90 vs. 0.91), and prediction of 1-year mortality (AUROC 0.76 vs. 0.75). Contrary to the 'black box' DNNs, our pipeline provided explainability on which morphological ECG changes were important for prediction. Results were confirmed in a population-based external validation dataset.

Conclusions

Future studies on DNNs for ECGs should employ pipelines that are explainable to facilitate clinical implementation by gaining confidence in artificial intelligence and making it possible to identify biased models.

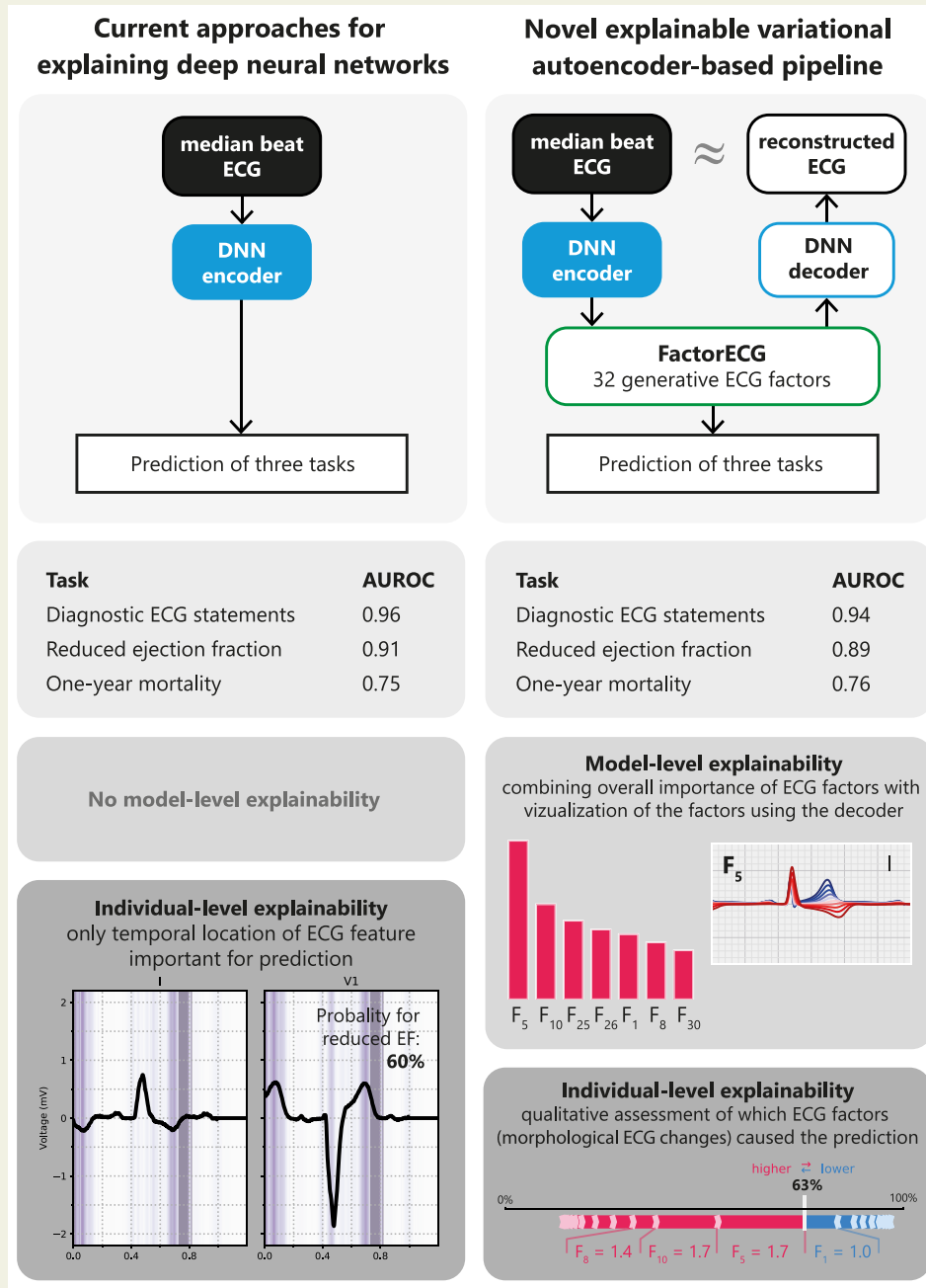
* Corresponding author. Tel: +31 88 757 3453, Fax: +31 88 757 3453, Email: r.vanes-2@umcutrecht.nl

[†]These authors contributed equally to this work.

© The Author(s) 2022. 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-NonCommercial License (<https://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

Graphical Abstract



Downloaded from https://academic.oup.com/ehjdh/article/3/3/390/6649795 by guest on 09 March 2023

Keywords

Deep neural network • Deep learning • Explainable • Interpretable • Artificial intelligence • Electrocardiogram

Introduction

The use of deep neural networks (DNNs) has led to tremendous improvements in automated interpretation of electrocardiograms (ECGs).¹ Recent studies have shown that DNNs achieve similar performance as cardiologists in tasks such as arrhythmia recognition and

triage of ECGs.^{2,3} Even more striking, DNNs have been shown to diagnose disorders that were not yet recognized on the ECG, such as reduced ejection fraction (EF) and 1-year mortality.^{4,5} Despite these promising developments, clinical implementation is severely hampered by the lack of trustworthy techniques to explain the decisions of the algorithm to clinicians.^{6,7} Due to the 'black box' nature of

Table 1 Glossary of terms used throughout the manuscript

Term	Definition
Decoder	The decoder is a part of the VAE and can be used to construct a median beat ECG from any combination of values in the FactorECG.
Deep neural network (DNN)	A deep neural network is an artificial intelligence algorithm that uses many layers with neurons to learn features from the input for prediction. In the case of ECG, a convolutional neural network is used, where the network learns features from the raw ECG signal itself.
Diagnostic ECG statement	Diagnostic statement given to an ECG by the overreading physician, e.g. sinus tachycardia, left bundle branch block, or early repolarization.
ECG factor	An ECG factor is one of the 21 values in the FactorECG and is a continuous value that can be used in any prediction model or for interpretability.
ECG feature	An ECG feature is a distinct morphological change to the ECG, such as a Q-wave or absent P-wave.
ECG measurement	ECG measurements are automated measurements of the intervals and axis of an ECG, such as PR interval and R-wave axis.
Encoder	The encoder is a part of the VAE and can be used to convert any median beat ECG into its respective FactorECG.
Explainable pipeline	The explainable pipeline in this work consists of three parts: firstly, the ECGs are encoded in its FactorECG using the pretrained VAE encoder, then the 21 significant ECG factors are entered into interpretable statistical models to perform the prediction or diagnosis task, and finally the pretrained VAE decoder is used to visualize the ECG factors that were deemed important for a specific task by the statistical model.
Factor traversal	The factor traversal is a method to visualize what ECG morphology a single ECG factor represents. This is done by keeping all ECG factor values at 0, while varying the factor of interest between -5 and 5 and construction and plotting ECGs using the decoder.
FactorECG	The latent space of the VAE proposed here is called the FactorECG and consists of 21 continuous normally distributed factors.
One-year all-cause mortality model	This model is trained to predict which individuals die from any cause within 1 year.
Reduced left ventricular ejection fraction model	The ejection fraction is the fraction of blood ejected from the left ventricle (chamber) of the heart with each contraction. An ejection fraction below 40% is a sign of heart failure with reduced ejection fraction. This model is trained to detect which patients have an ejection fraction below 40% as measured by echocardiography.
Variational auto-encoder (VAE)	The variational auto-encoder consists of three parts, an encoder DNN to compress the raw ECG data into a reduced set of continuous values, the latent space, and a decoder to reconstruct that same ECG from these values. It is trained in an unsupervised manner by learning to reconstruct many ECGs from the latent space.

most algorithms, and the limitations of current *post hoc* explainability methods, the association between input and output remains unexplainable to humans.⁸ The lack of interpretability makes it difficult for clinicians to gain enough confidence to make clinical decisions based on these algorithms, and more importantly, impossible to identify biased or inaccurate models. These issues have already been acknowledged by the new European Union's General Data Protection Regulation, which requires a 'right to explanation' for AI algorithms.⁹

To improve explainability, several *post hoc* explainability methods have been proposed, usually by providing heatmaps on top of the ECG. However, a major limitation of these methods is that they only provide the temporal location of ECG features important in making the diagnosis, but do not indicate the actual feature (e.g. when the QRS-complex is highlighted the feature could be R-wave height, QRS shape, or something else completely).^{5,10,11} This makes that heatmaps are of limited explainable value for showing which morphological ECG changes were important for a specific prediction. Moreover, heatmap-based methods are only able to provide explainability on the level of an individual ECG, but not for the whole model. This combination makes them susceptible to confirmation

bias, as we assume that the feature we think is important is also the one that was used in the few examples that were observed.⁶ Finally, recent studies have shown that saliency-based methods can be very unreliable in providing consequent annotations and can also show reassuring saliency maps when a model is completely untrained, stressing the need for better approaches to explain the output of DNNs.^{8,12,13} Therefore, instead of explaining the 'black box' after it was trained, the preferred way for algorithms to produce trustworthy explanations is to develop pipelines that are explainable by design.⁸

We hypothesized that an ECG can be explained by a few underlying anatomical and (patho)physiological factors of variation. Variational auto-encoders (VAEs) are generative networks that use the power of DNNs to learn to compress any ECG into a selected number of explanatory and independent factors. Moreover, they can reconstruct the ECG from these factors.^{14,15} In this study, we aimed to use a VAE to identify the underlying factors of variation in the ECG morphology and use them to develop an explainable pipeline for the interpretation of ECGs. Firstly, we investigate the underlying generative process of the learned factors by relating

them to known ECG parameters and the most common conventional diagnostic ECG statements. Secondly, we train and internally and externally validate the explainable pipeline for use in the novel ECG use cases, detection of reduced EF and prediction of 1-year mortality, and perform a comparison with current state-of-the-art 'black box' DNNs and conventional ECG algorithms.

Methods

Study participants

The dataset consisted of all patients between 18 and 85 years of age with at least one ECG acquired in the University Medical Center Utrecht (UMCU) between July 1991 and August 2020. All data were de-identified in accordance with the EU General Data Protection Regulation and written informed consent was not required by the UMCU ethical committee.

Data acquisition for training and validation of the variational auto-encoder

All resting 12-lead ECGs were exported from the MUSE ECG system (MUSE version 8; GE Healthcare, Chicago, IL, USA) in raw voltage format and converted to median beats as described by van de Leur *et al.*¹⁰ All ECGs that were deemed technically inadequate by either the MUSE 12SL algorithm or interpreting physician were excluded from the analyses. No labels were used in the training of the unsupervised auto-encoder.

Data acquisition for training and validation of the 'black box' deep neural networks and explainable pipelines

For training of the algorithms to detect conventional diagnostic ECG statements, we included a subset of ECGs that were obtained at all non-cardiology departments, as these ECGs were systematically annotated by a physician as part of the regular clinical workflow. We selected the 35 most common diagnostic statements for training [i.e. sinus tachycardia or left bundle branch block (LBBB), a complete overview can be found in the [Supplementary material online, Methods](#)] and used 20% of the patients for hyperparameter optimization. For validation of the ECG interpretation models, an independent dataset comprising 1000 randomly selected ECGs of unique patients was annotated by a panel of five practising electrophysiologists or cardiologists for all diagnostic statements as described by van de Leur *et al.*³ A reduced set of the 35 diagnostic statements was tested, as some abnormalities did not occur in the test dataset. Moreover, the myocardial ischaemia labels in different locations were combined. A glossary of technical terms used throughout the manuscript can be found in [Table 1](#).

To train and validate the algorithms to detect reduced EF (below 40%) and predict 1-year mortality, we selected patients using the same approaches as Attia *et al.*⁴ and Raghunath *et al.*,⁵ respectively. For the reduced EF model, patients with an ECG–echocardiogram pair (acquired within 14 days) were retrieved, the EF was dichotomized at 40% and patients were split in a 75:25 ratio on the patient level. For the test set only the first ECG–echocardiogram pair per patient was used, to avoid the overrepresentation of sicker patients with multiple pairs. For the 1-year mortality model, all patients with at least 1 year of follow-up available for evaluation of all-cause mortality were selected and split in a 60:40 ratio on the patient level. For the test set, we randomly selected one ECG if the patient had multiple ECGs. Importantly, both train-test splits were made on the patient level, ensuring no overlap in patients between the sets. Detailed information on the data acquisition for all three tasks can be found in the [Supplementary material online, Methods](#).

For external validation of the VAE and the performance of the models for detection of reduced EF, we included individuals who underwent both cardiac magnetic resonance (CMR) imaging and 12-lead ECG at the same time at the first imaging visit of the population-based UK Biobank cohort (analysis performed under application number 74 395). All 10 s 12-lead resting ECGs were acquired using a GE CardioSoft device at 500 Hz and converted to median beats by the GE algorithm. Only individuals where the left ventricular EF was determined on the CMR using a manual analysis protocol by Petersen *et al.*^{16,17} were included (UK Biobank return number 2541). Details on the UK Biobank cohort, the CMR protocol, and the manual CMR analysis protocol have been described before.^{17–19}

Training and architecture of the variational auto-encoder

The VAE consists of three parts: the encoder, the latent space (with multiple continuous ECG factors, combined referred to as the FactorECG), and the decoder.¹⁴ The original 12-lead median beat ECG is entered into the encoder that compresses the ECG to its FactorECG with 32 continuous factors. From those same factors, the ECG is reconstructed by the decoder, and the difference between the input and reconstructed ECG was used to train the model. The decoder and encoder are a standard convolutional neural network and the inverse of that neural network, respectively. A specific type of VAE was used, called the β -VAE, where an additional hyperparameter β is included in the loss term to learn disentangled factors, i.e. generative factors of variation that are independent of each other.¹⁵ The two most important hyperparameters in the β -VAE were the number of ECG factors and the β -value. For both, values of 8, 16, 32, 64, and 128 were evaluated. Considering that increasing the β -term results in higher reconstruction errors, we chose a β that resulted in a good trade-off between reconstruction error and adequate disentanglement in significant factors, which was assessed using the factor traversals. Moreover, increasing the number of ECG factors above 32 did not yield an increase in significantly contributing factors (i.e. factors that encode variation), therefore this value was selected. A schematic overview of the technique can be found in [Figure 1](#), while an animation of the approach is included as [Supplementary material online](#). Detailed information on the training and architecture of the VAE can be found in [Supplementary material online, Methods](#).

Training and explainability of the pipeline

To obtain an explainable pipeline for prediction or diagnosis, we combined the following steps: (i) the median beat 12-lead ECGs were encoded in their FactorECG using the pretrained VAE encoder, (ii) the 21 significant ECG factors were entered into common interpretable statistical models to perform the prediction or diagnosis task, and (iii) the pretrained VAE decoder is used to visualize the ECG factors that were deemed important for a specific task by the statistical model.

The explainable pipeline is compared with current state-of-the-art 'black box' DNNs in three tasks: conventional ECG interpretation, detection of reduced EF, and prediction of 1-year mortality. For the conventional ECG interpretation task, we trained binary logistic regression models for each of the 35 diagnostic ECG statements on the FactorECGs, as it provided maximum interpretability. For the detection of reduced EF and prediction of 1-year mortality, as the aim was maximum performance, we trained two extreme gradients boosting decision tree (XGBoost) models.²⁰ For this model, interpretability was obtained using Shapley Additive exPlanations, which can provide feature importance measures for every ECG factor on a model- and individual patient level.²¹ For comparison, a baseline state-of-the-art 'black box' DNN with a similar architecture as the encoder of the VAE and the median

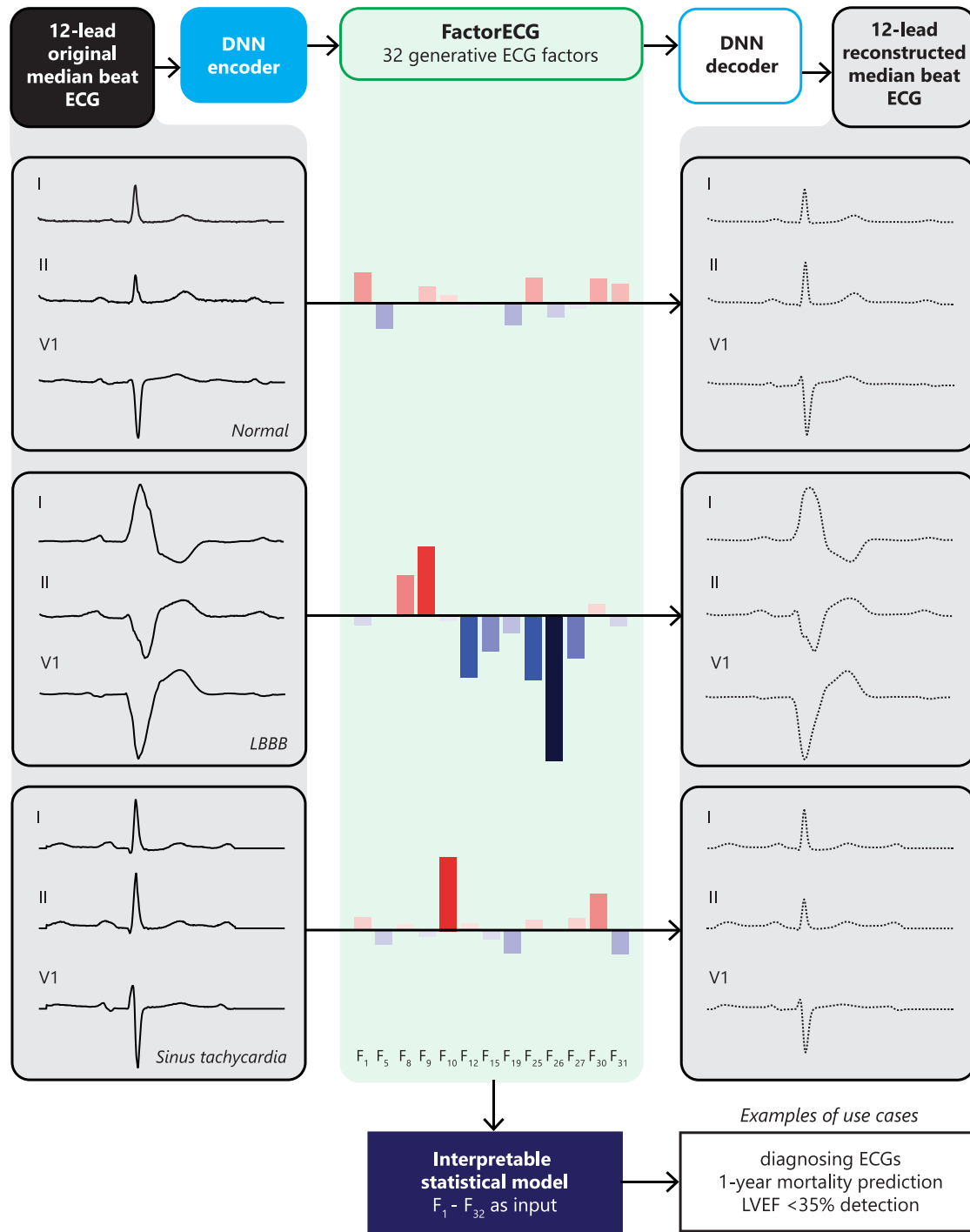


Figure 1 Illustration of the full pipeline: a variational auto-encoder, the FactorECG, and reconstructions. The variational auto-encoder consists of three parts: the encoder, the FactorECG space, and the decoder. An input 12-lead median beat electrocardiogram is entered into the decoder that compresses the electrocardiogram to its FactorECG with 32 continuous factors. From those same factors, the electrocardiogram is reconstructed and the difference between the input and reconstructed electrocardiogram is used to train the model. The electrocardiogram factors are subsequently used in two ways: for development of interpretable classifiers for electrocardiogram diagnostic statements, reduced ejection fraction and 1-year mortality, and for visualization purposes. Electrocardiogram factors can provide both individual patient- and model-level visualizations. Individual visualizations are depicted here, where three median beat electrocardiograms and their reconstructions are represented in the FactorECG. Notably, as dimension 10 encodes ventricular frequency, we see high values for the sinus tachycardia electrocardiogram. Moreover, as dimension 26 inversely encodes left bundle branch conduction delay, we see low values for the left bundle branch block electrocardiogram. The normal electrocardiogram has value around zero for all factors, as the variational auto-encoder is forced to learn factors with zero mean. ECG, electrocardiogram; LVEF, left ventricular ejection fraction; LBBB, left bundle branch block.

beat ECG as input was trained for all three tasks.^{10,22} Additional information on the baseline model and training procedures for the three tasks are available in the [Supplementary material online, Methods](#).

The pipeline can provide explanations on both the model- and individual patient level. On the model-level, ECG factors are visualized by factor traversals using the pretrained VAE decoder: varying the values of an individual factor while decoding and plotting the median ECG beat. Every visualization starts with zeros for all factors, which represents the mean ECG in the training dataset. Then, for every individual factor, values between -5 and 5 are assigned, while keeping the others at zero, and through decoding a new generated ECG is obtained. These reconstructions are subsequently visualized in the same graph. This allows for detailed visualizations of morphological changes. On the individual patient level, explainability is obtained by combining the distinct FactorECG values of that ECG with knowledge on the predictors that were important for a specific task. For example, if the FactorECG of an ECG contains a high value for a specific factor and this factor was associated with the outcome by the interpretable statistical model, this would explain why this specific ECG has a higher risk of the outcome. Other explainability is provided by associating the ECG factors with known ECG parameters (i.e. PR interval or QRS duration) and known ECG diagnoses (i.e. LBBB or sinus tachycardia).

Statistical analysis

All data are presented as mean \pm SD or median with interquartile range, where appropriate. All individual ECG factors were related to the conventional ECG measurements computed by the MUSE algorithm (i.e. ventricular rate, PR, QRS, and Bazett corrected QT duration, and R- and T-axis) using hexagon plots and Pearson correlation coefficients. Discriminatory performance of the models is assessed in the test sets using the c-statistic or area under the receiver operating curve (AUROC) and the area under the precision-recall curve (AUPRC). As all models are weighted for class imbalance, a probability cut-off of 50% was used. Overall, 95% confidence intervals are obtained using 2000 bootstrap samples. The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis Statement for the reporting of prediction models was followed.²³

Results

Development of the variational auto-encoder and explainability of the FactorECG

The dataset for training of the VAE consisted of 1 144 331 12-lead median beat ECGs of 251 473 unique patients. The VAE was able to reconstruct the median beat ECGs excellently with a mean Pearson correlation coefficient of 0.90 ($P < 0.001$) between the original and reconstructed ECG. Reconstructions were most accurate for sinus rhythm, sinus bradycardia, early repolarization, and pericarditis ECGs (mean $r = 0.91$ – 0.92), and least accurate for the rarer ECGs with ST elevation suspected of myocardial infarction and ventricular tachycardia (mean $r = 0.62$ – 0.70). An overview of mean correlation coefficients per diagnostic ECG statements can be found in [Supplementary material online, Table S1](#).

By analyzing the factor traversals (see [Supplementary material online, Figure S2](#)), only 21 of the 32 factors were found to be necessary to reconstruct the ECG, and the other 11 were not used by the model to encode significant data. Model-level explainability, using

factor traversals, is shown for a subset of the 21 factors in [Figure 2](#). An online tool to visualize the generated ECGs interactively is available via <https://decoder.ecgx.ai>. To further investigate and gain interpretability in the ECG factors, Pearson correlation coefficients were computed between conventional ECG measurements and ECG factors values ([Figure 3](#)). Ventricular rate is mostly correlated to factor 10 ($r = 0.96$, $P < 0.001$), while QRS duration is mostly correlated to factor 25 ($r = -0.47$, $P < 0.001$). PR and QT interval are mostly correlated to factors 8 ($r = 0.62$, $P < 0.001$) and 30 ($r = -0.52$, $P < 0.001$), respectively. The 21 significant ECG factors were independent of each other, with Pearson correlation coefficients ranging between -0.06 and 0.09 (see [Supplementary material online, Figure S3](#)).

Performance and explainability for conventional electrocardiogram interpretation

The dataset for training the algorithms to perform conventional ECG interpretation consisted of 369 216 ECGs of 152 831 patients, while for validation the expert-annotated dataset was used, containing 965 ECGs (of 965 patients) of adequate quality. Three hundred and forty-three (36%) of the ECGs had more than one diagnostic statement and sinus rhythm was the most prevalent (72%), while third-degree AV block was the least prevalent (0.1%, [Table 2](#)). The mean AUROC of the explainable pipeline was 0.94 (95% CI 0.92–0.96), compared with 0.73 (95% CI 0.65–0.81) for the rule-based MUSE algorithm and 0.96 (95% CI 0.94–0.98) for the 'black box' DNN. The explainable pipeline performed similarly for most diagnostic statements but was outperformed for the diagnosis of left ventricular hypertrophy and low QRS voltage by the 'black box' DNN ([Table 2](#)). The conventional MUSE algorithm, that is currently used in clinical practise, performed worst for all diagnostic statements ([Table 2](#)). To understand which ECG factors were important for the pipeline to detect each ECG statement, we used the logistic regression's coefficients as feature importance scores ([Figure 4](#)). The negative (blue) and positive (red) scores from [Figure 4](#) can be related to the generated ECGs in the factor traversals after negative (blue) and positive (red) perturbations in [Figure 2](#) and [Supplementary material online, Figure S2](#).

Performance and explainability for detection of reduced ejection fraction

For the algorithms to detect reduced EF, 39 603 matched ECG–echocardiogram pairs of 22 676 patients were available, of which 25% (5669 unique patients, first pair per patient used) were used for validation. Seven hundred and thirteen patients (13%) in the validation set had an EF below 40%. The explainable pipeline achieved an AUROC and AUPRC of 0.89 (95% CI 0.89–0.91) and 0.66 (95% CI 0.63–0.70), in comparison to 0.91 (95% CI 0.89–0.92) and 0.70 (95% CI 0.68–0.74) for the 'black box' DNN, respectively. The most important model-level ECG factors for detecting reduced EF were high values in factors 5, 10, and 8 and low values in factors 25, 26, 1, and 30 ([Figure 5](#)). These correspond to negative T waves, higher ventricular rate, ST elevation, increased P-wave area and PR-interval, right bundle branch block (RBBB), and LBBB, respectively. [Figure 6](#) shows a model- and individual

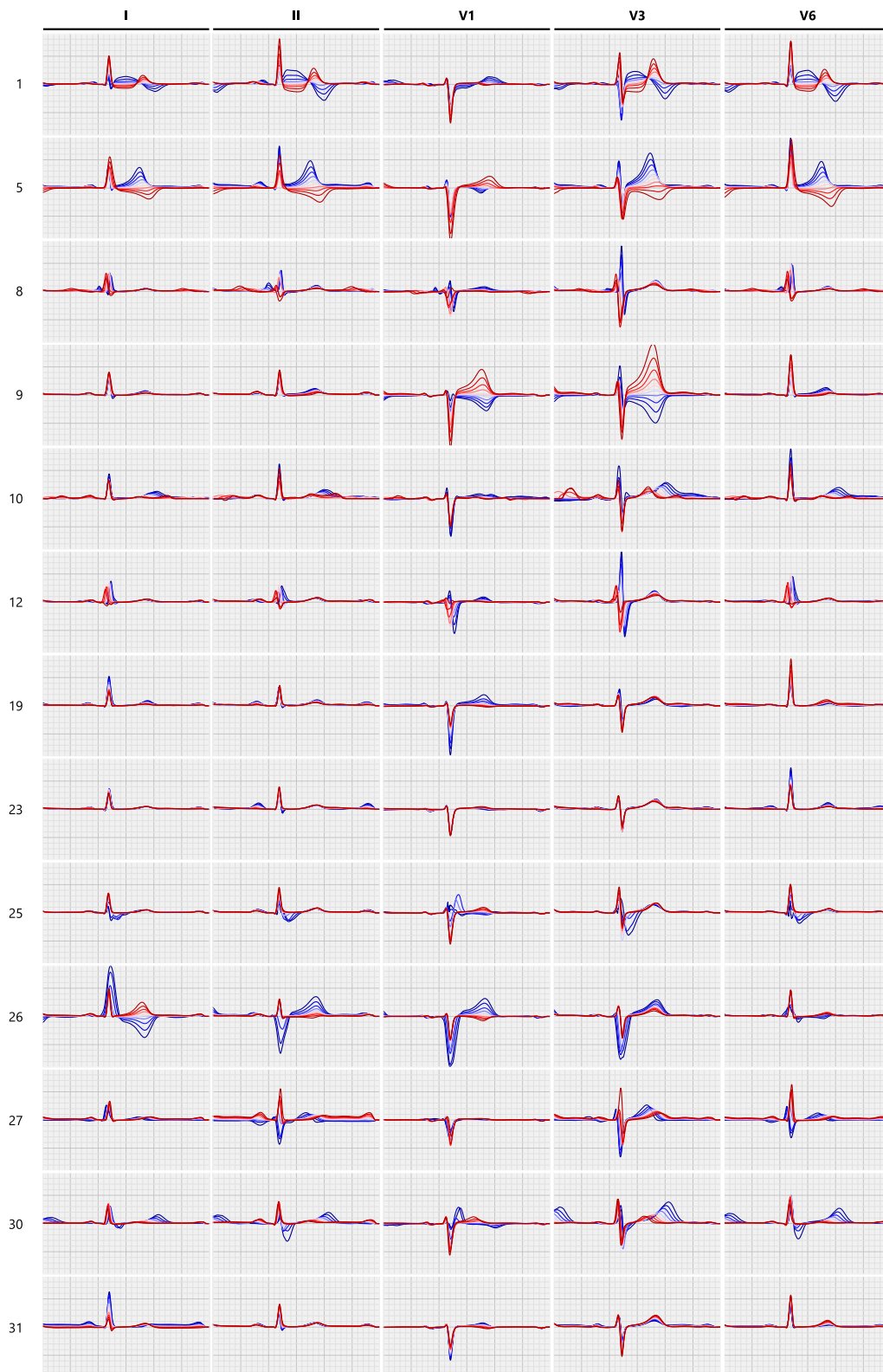


Figure 2 Factor traversals of a subset of the electrocardiogram factors for Leads I, II, V1, V3, V6. Factor traversals of a subset of the 21 electrocardiogram factors that hold significant information for correctly reconstructing electrocardiograms. Each row corresponds to the factor traversal for one electrocardiogram factor and the columns to a subset of the 12 leads. The factor traversal for one row is obtained by starting with a 'mean' FactorECG where all factors are zero and adding offsets for that factor in a range of -5 to 5 . The generated electrocardiograms are then plotted where red represents high values for that factor and blue low values.

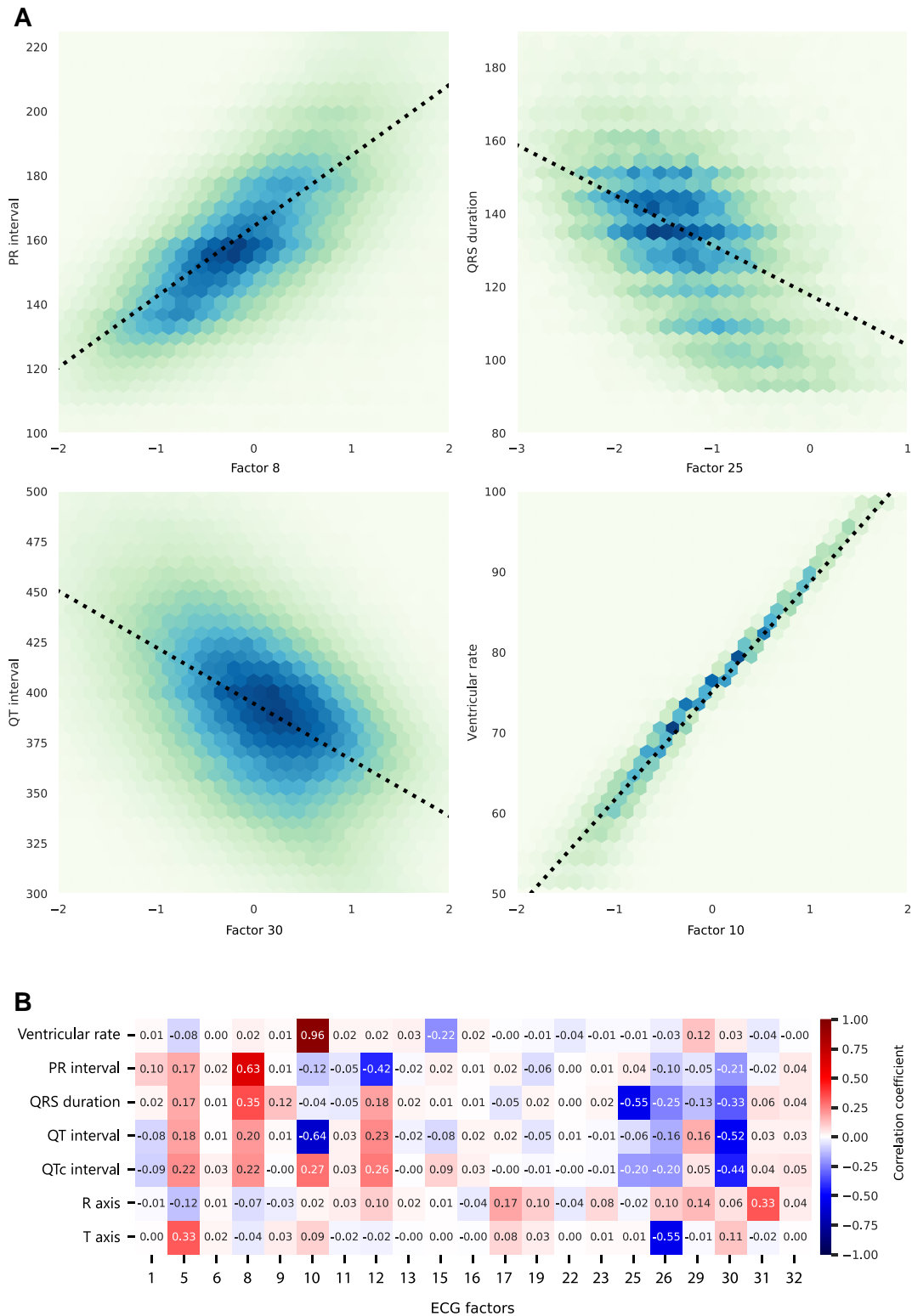


Figure 3 Relationship of the electrocardiogram factors with conventional electrocardiogram measurements. (A) Hexagon plots where datapoints of electrocardiogram factor–electrocardiogram measurements pairs over all samples in the variational auto-encoder dataset are binned into hexagon grids to relate values of Factors 8, 25, 30, and 10 to the PR interval, QRS duration, QT interval, and ventricular rate, respectively. (B) Pearson correlation coefficients between electrocardiogram measures of ventricular rate, PR interval, QRS duration, QT interval, Bazett corrected QT interval, R-axis, and T-axis, and electrocardiogram factor values over all samples in the variational auto-encoder dataset.

Table 2 Diagnostic performance values for the conventional electrocardiogram interpretation task in the expert-annotated test set

Diagnostic statement	Prevalence n (%)	MUSE 12 SL		Explainable pipeline		Black box DNN	
		AUROC (95% CI)	AUPRC	AUROC (95% CI)	AUPRC	AUROC (95% CI)	AUPRC
Sinus rhythm	697 (72)	0.90 (0.88–0.92)	0.96	0.94 (0.92–0.96)	0.96	0.96 (0.95–0.97)	0.98
Sinus bradycardia	30 (3.1)	0.70 (0.61–0.78)	0.09	0.95 (0.92–0.98)	0.39	0.94 (0.87–0.97)	0.37
Sinus tachycardia	91 (9.4)	0.95 (0.92–0.97)	0.75	0.99 (0.98–0.99)	0.81	0.99 (0.99–1.00)	0.94
Atrial fibrillation	90 (9.3)	0.88 (0.84–0.93)	0.73	0.99 (0.98–0.99)	0.78	0.98 (0.97–0.99)	0.86
Atrial flutter	2 (0.2)	0.74 (0.49–1)	0.04	0.98 (0.96–0.99)	0.04	1.00 (0.99–1.00)	0.67
Supraventricular tachycardia	18 (1.9)	0.58 (0.5–0.67)	0.15	0.97 (0.95–0.98)	0.33	0.98 (0.96–0.99)	0.34
Junctional bradycardia	4 (0.4)	0.75 (0.5–1)	0.13	0.99 (0.96–1.00)	0.46	1.00 (0.99–1.00)	0.56
Ventricular tachycardia	2 (0.2)	0.50 (0.5–0.5)	0	0.99 (0.98–1.00)	0.21	1.00 (0.99–1.00)	0.23
Pacemaker rhythm	27 (2.8)	0.92 (0.85–0.98)	0.74	0.97 (0.94–0.98)	0.46	0.97 (0.93–0.99)	0.68
First-degree AV block	57 (5.9)	0.86 (0.8–0.92)	0.66	0.98 (0.97–0.99)	0.68	0.96 (0.94–0.98)	0.71
Third-degree AV block	1 (0.1)	0.5 (0.5–0.5)	0	1.00 (1.00–1.00)	0.31	1.00 (0.99–1.00)	0.14
RBBB	59 (6.1)	0.95 (0.91–0.98)	0.66	0.98 (0.97–0.99)	0.69	0.99 (0.98–1.00)	0.83
LBBB	22 (2.3)	0.88 (0.79–0.97)	0.64	1.00 (0.99–1.00)	0.82	1.00 (1.00–1.00)	0.95
LAFB	71 (2.4)	0.64 (0.59–0.69)	0.29	0.84 (0.79–0.88)	0.28	0.97 (0.96–0.98)	0.62
NICD	14 (1.5)	0.63 (0.53–0.76)	0.09	0.94 (0.92–0.96)	0.12	0.88 (0.73–0.97)	0.3
Myocardial infarction	66 (6.8)	0.6 (0.55–0.65)	0.19	0.77 (0.72–0.82)	0.16	0.77 (0.71–0.82)	0.19
Left ventricular hypertrophy	44 (4.6)	0.79 (0.71–0.86)	0.32	0.82 (0.77–0.87)	0.15	0.97 (0.95–0.98)	0.63
Low QRS voltage	40 (4.2)	0.76 (0.68–0.83)	0.36	0.8 (0.74–0.86)	0.18	0.96 (0.94–0.98)	0.63
Prolonged QT interval	22 (2.3)	0.69 (0.6–0.8)	0.14	0.95 (0.91–0.97)	0.43	0.93 (0.89–0.95)	0.2
Early repolarization	23 (2.4)	0.52 (0.5–0.57)	0.04	0.96 (0.93–0.98)	0.45	0.98 (0.97–0.99)	0.61
Acute pericarditis	7 (0.7)	0.57 (0.5–0.71)	0.15	0.99 (0.99–1.00)	0.49	0.99 (0.96–1.00)	0.61

The AUROC and AUPRC scores per diagnostic statement in the ECG interpretation task for the rule-based MUSE algorithm, explainable pipeline, and 'black box' DNN are shown. A reduced set of the 35 diagnostic statements was tested, as some abnormalities did not occur in the test dataset. Moreover, the myocardial ischaemia labels in different locations were combined. AUROC, area under the receiver operating curve; AUPRC, area under the precision-recall curve; AV, atrioventricular; CI, confidence interval; DNN, deep neural network; LAFB, left anterior fascicular block; LBBB, left bundle branch block; NICD, non-specific intraventricular conduction delay; RBBB, right bundle branch block.

patient-level explanation for the detection of reduced EF using the novel pipeline, in comparison to the *post hoc* explainability methods used up until now.

Performance and explainability for prognosis of 1-year mortality

For the models to predict 1-year mortality, follow-up was available for 909 958 ECGs of 177 448 patients, of which 40% (70 979 unique patients, ECG sampled randomly per patient) was used for validation. Five thousand three hundred and thirty-four patients (7.5%) in the validation set deceased within 1 year. The explainable pipeline achieved an AUROC and AUPRC of 0.76 (95% CI 0.76–0.77) and 0.21 (95% CI 0.20–0.22) compared with 0.75 (95% CI 0.74–0.76) and 0.21 (95% CI 0.20–0.22) for the 'black box' DNN, respectively. In contrast, an XGBoost model that included only age and sex had an AUROC of 0.65 (95% CI 0.64–0.66) and an AUPRC of 0.12 (95% CI 0.12–0.13). The most important global ECG factors for the prediction of 1-year mortality were high values for factors 10, 5, 12, and 11, and low values for factors 1, 30, 9, and 27 (Figure 5). These correspond to an increased risk of 1-year mortality with higher ventricular rate, inferolateral negative T-waves, ST-elevation, prolonged QT interval, and anterior negative T-waves. Table 3 shows a summary of

the ECG morphology of all ECG factors, in combination with the most important associations for each factor.

External validation of the FactorECG pipeline for detection of reduced ejection fraction

Manually analysed CMR imaging and 12-lead ECG recordings were available for 4855 individuals, of which 28 had a reduced EF (0.62%). The VAE, which was trained in the UMC Utrecht dataset, could accurately reconstruct the median beat ECGs from the UK Biobank (mean Pearson correlation coefficient between the original and reconstructed ECG: 0.88, $P < 0.001$). The FactorECG pipeline achieved an AUROC of 0.89 (95% CI 0.84–0.95) and an AUPRC of 0.06 (95% CI 0.03–0.15) for the detection of reduced EF in the external validation dataset. In comparison, the 'black box' DNN achieved an AUROC of 0.86 (95% CI 0.76–0.94) and an AUPRC of 0.12 (95% CI 0.06–0.27).

Discussion

In this study, we demonstrate a novel pipeline that provides improved explainable interpretation of ECGs, which consists of three major components: (i) a generative deep learning model that learned

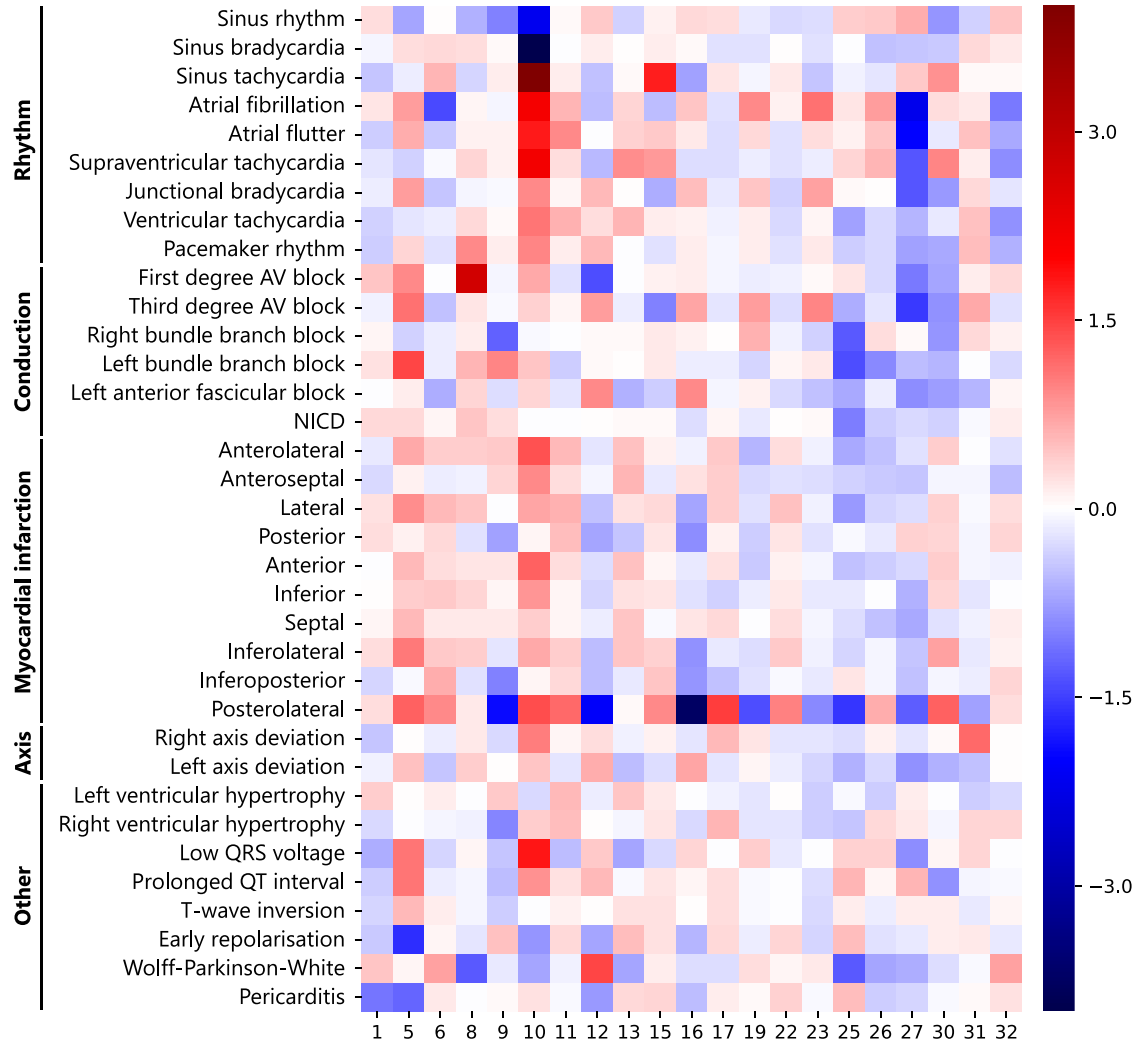


Figure 4 Importance score for each of the 32 factors in predicting 35 diagnostic electrocardiogram statements. Importance scores of each of the 32 factors in the logistic regression for all 35 diagnostic electrocardiogram statements are shown to relate which dimensions are important for diagnosis. High importance values indicate that a high value for the dimension is diagnostic for that abnormality, and vice versa. The negative (red) and positive (blue) scores can be related to the reconstructions after negative (red) and positive (blue) perturbations in Figure 2. Notably, Factor 10 encodes ventricular frequency (as observed in Figures 2 and 3) and therefore has a high value in sinus tachycardia (red) and a low value in sinus bradycardia (blue). NICD, non-specific intraventricular conduction delay.

to summarize the underlying factors of variation of an ECG in 21 factors (the FactorECG), (ii) a visualization technique to provide insight into ECG morphology that these factors encode, and (iii) a common interpretable statistical method to perform diagnosis or prediction using the ECG factors (Figure 1). We investigated the FactorECG using visualizations and associations with conventional ECG measurements and diagnostic ECG statements to show that many of the factors represent valid and relevant generative factors of ECG morphology (Table 3). Moreover, when applying the novel explainable technique for conventional ECG interpretation and recently emerged use cases for the ECG, not only did it perform similarly to the ‘black box’ algorithms for these use cases, but it could also explain which morphological ECG changes were important for prediction or diagnosis. Finally, we showed that FactorECG itself, and the pipeline for detection of reduced EF, generalize

excellently to a completely different population-based cohort. This indicates that inherently explainable deep learning methods should be used to gain confidence in AI for clinical decision making, and more importantly, make it possible to identify biased or inaccurate models.

A longstanding assumption was that the high-dimensional and non-linear ‘black box’ nature of the currently applied DNNs was inevitable to gain the impressive performances shown by these algorithms.^{5,13,22} The major finding of the current study is that a VAE-based approach performs on par with the ‘black box’ algorithms in both conventional and novel tasks (Table 2), while also giving insight in the ECG morphology that explains the prediction. A main advantage of the current approach over previous attempts to open the ‘black box’ of DNNs using *post hoc* explainability methods (i.e. heatmaps) is that we can reliably and quantitatively specify the

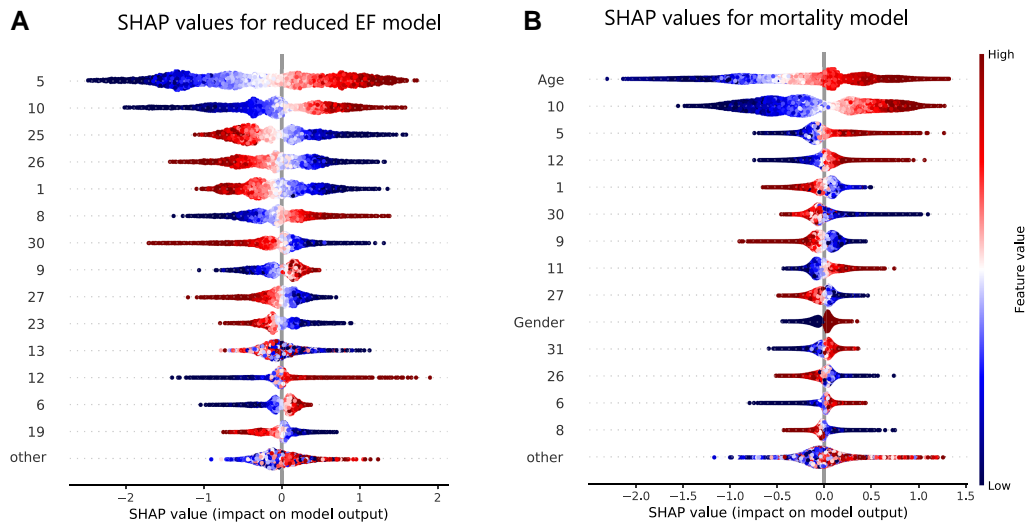


Figure 5 Explanations for the 1-year mortality and reduced ejection fraction models using Shapley Additive exPlanations values. (A) The most important model-level electrocardiogram factors for detecting reduced ejection fraction computed using Shapley Additive exPlanations values. Importance is ordered from top-to-bottom and colouring corresponds to the reconstructed electrocardiograms in [Figure 2](#). (B) The most important global electrocardiogram factors for predicting 1-year mortality. Importance is ordered from top-to-bottom and colouring corresponds to the reconstructed electrocardiograms in [Figure 2](#).

morphology of the ECG change, instead of only pointing at the location on the ECG's time axis ([Figure 6](#)).^{3,5,10,24}

Other studies investigated the use of (variational) auto-encoders on 12-lead ECGs in smaller datasets and showed that VAEs can be useful for compression of ECGs, data augmentation, clustering, and feature generation.^{25–28} Interestingly, Kuznetsov et al.²⁸ also determined that ~20–25 factors are needed to encode a single or median beat ECG. Our work makes the latent space of a VAE (i.e. the FactorECG) clinically useful and explainable to physicians, by (i) linking the ECG factors with known ECG measurements and diagnostic statements ([Figures 3 and 4](#) and [Tables 2 and 3](#)), (ii) providing extensive visualizations offline ([Figure 2](#)) and using an online tool (<https://decoder.ecgx.ai>), and (iii) showing that the ECG factors have adequate predictive power in various downstream tasks. [Figure 6](#) shows an example of how a FactorECG-based pipeline can be used in clinical practise. At model-level, the overall most important morphological ECG changes (i.e. ECG factors) for a specific task are shown and can be used to detect possible biases. At patient level, the user is provided with an individual explanation of which morphological ECG changes in this patient are causing the higher risk of reduced EF, for example. The online tool provides a possibility to upload ECGs to show the predictions and explanations, or to extract the FactorECGs to train new models using the code provided (<https://decoder.ecgx.ai> and <https://github.com/rutgervandeleur/ecgxai>).

We hypothesized that an ECG can be explained by a few underlying explanatory factors of variation and showed that it is possible to encode the median beat ECG morphology in 21 continuous factors, from which the ECG can be reconstructed with high precision (Pearson correlation between original and reconstructed ECG 0.90 in internal validation and 0.88 in external validation). An online tool for clinicians to interactively visualize the factors can be found via <https://decoder.ecgx.ai>. When relating the ECG factor traversals ([Figure 2](#) and [Supplementary material online](#),

[Figure S2](#)) to diagnostic ECG statements and conventional ECG measurements ([Figures 3 and 4](#)), we were able to relate many of them to the underlying anatomical and (patho)physiological factors ([Table 3](#)). For example, Factor 10 has a clear linear relationship with ventricular frequency and therefore shows high values for sinus tachycardia and low values for sinus bradycardia. Moreover, the factor traversals ([Figure 2](#)) show the changes in the ECG associated to the ventricular frequency, such as the length of the QT interval and appearance of the T-wave of the previous beat. Factors 6, 23, and 27 account for the P-wave size and are related to diagnoses that involve the P-wave, such as junctional bradycardia and atrial fibrillation, while PR interval (or location of the P-wave) is encoded in factor 8. Factors 25, 26, and 30 encode ventricular conduction delays, such as right and LBBB, while ventricular repolarization is mainly encoded in Factors 1, 5, 9, 13, and 30. ST elevation is most prominent in Factors 1 and 5, which are subsequently important for predicting diagnoses such as acute pericarditis and early repolarization. Next to these more common ECG variations, rare abnormalities are also represented, as for example Wolff-Parkinson-White pattern (with pre-excitation and short PR interval) is encoded using a combination of Factors 8 and 12. An overview of the ECG morphology and most important associations for each ECG factor can be found in [Table 3](#).

For the reduced EF task, we found that the performance of the explainable pipeline is equivalent to both the black box DNN in our dataset and in the original publication by Attia et al.⁴ This finding was externally validated in the UK Biobank, a population-based cohort that is very different from the academic hospital-derived training population, and shown to be robust with a similar AUROC as in the internal validation dataset. Most important ECG indicators for reduced EF were consistent with previous findings that indicated similar features to be predictive of heart failure: inferolateral negative T-waves, increased ventricular rate, P-wave area, prolonged PR interval, RBBB, LBBB, but also inferolateral ST elevation.²⁹ The

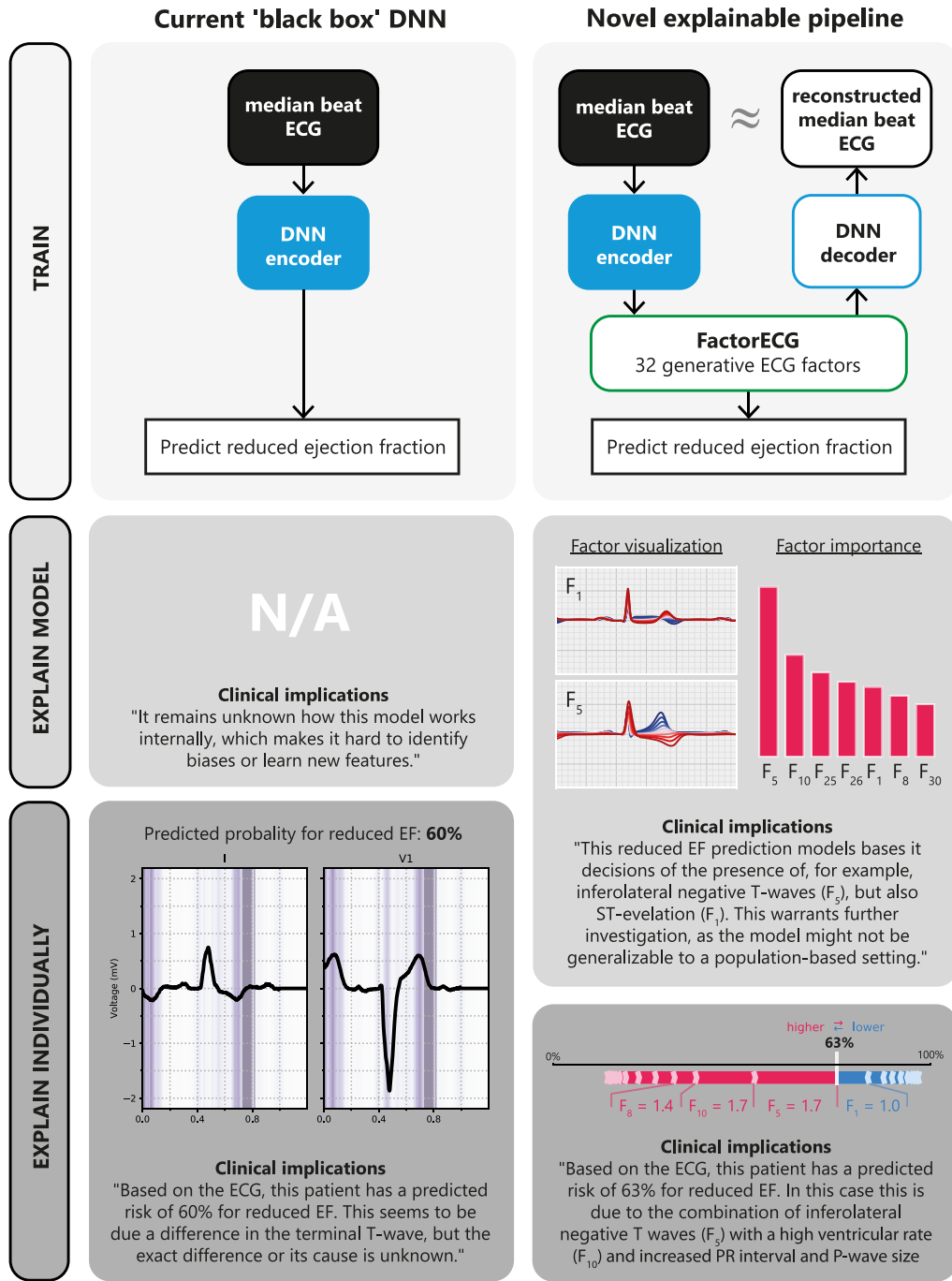


Figure 6 Comparison of architecture and model- and individual patient-level explainability using the novel inherently explainable approach as compared with *post hoc* heatmap-based explainability for detection of reduced ejection fraction. The conventional 'black box' deep neural network contains only a single encoder to interpret the electrocardiogram. Afterwards, Guided Grad-CAM is applied to show what segments of the electrocardiogram were important for prediction on the patient level. Model-level explainability is not possible. The novel explainable pipeline adds a generative part to the architecture, which allows for precise visualizations of the morphological electrocardiogram features. By combining factor Shapley Additive exPlanations importance scores and factor traversals, we obtain model-level explainability. Individual patient-level explainability is achieved using individual Shapley Additive exPlanations importance score.

importance of this latter feature illustrates that the DNN also picks up reduced EF due to acute ischaemia. This could hamper the generalizability of such models for screening purposes in the general population as these patients are only present in large hospitals and is one

of the reasons why explainable models are imperative.^{8,30} Although the model for 1-year mortality performs worse than in the original paper by Raghunath *et al.*,⁵ it does perform similarly to the 'black box' DNN on our dataset. The difference in performance is likely

Table 3 Summarizing description of electrocardiogram morphology and associations of the 21 significant electrocardiogram factors

Factor	High values		Low values	
	ECG morphology	Associations	ECG morphology	Associations
1	Inferolateral horizontal ST depression	Left ventricular hypertrophy	Inferolateral horizontal ST elevation	Pericarditis, reduced EF, and 1-year mortality
5	Inferolateral T-wave inversion	T-wave axis, LBBB, inferior and lateral ischaemia, low QRS voltage, reduced EF, and 1-year mortality	Inferolateral concave ST elevation	T-wave axis, pericarditis, early repolarization
6	Increased P-wave amplitude		Reduced P-wave amplitude	Atrial fibrillation, atrial flutter
8	Shorter PR-interval and P-wave duration	First-degree AV block and reduced EF	Longer PR-interval and P-wave duration	WPW pattern
9	Anterior concave ST-elevation	LBBB and reduced EF	Anterior T-wave inversion	RBBB, RVH, posterior ischaemia, T-wave inversion, and 1-year mortality
10	Shorter QT-interval and TP-interval	Increased ventricular frequency, sinus tachycardia, atrial fibrillation, atrial flutter, SVT, low QRS voltage, reduced EF, and 1-year mortality	Longer QT-interval and TP-interval	Decreased ventricular frequency, sinus rhythm, and sinus bradycardia
11	Subtle QRS- and T-wave changes	One-year mortality	Subtle QRS- and T-wave changes	
12	Earlier onset of depolarization	Reduced PR-interval, WPW pattern, LAFB, and 1-year mortality	Later onset of depolarization	Increased PR-interval and first-degree AV block
13	Anterior horizontal ST-elevation	Anterior and septal ischaemia	Anterior horizontal ST-depression	Reduced EF
15	P/T overlap	Sinus tachycardia	Reduced P-wave amplitude	Third-degree AV-block and junctional bradycardia
16	Subtle T-wave changes	LAFB	Subtle T-wave changes	Posterior and lateral ischaemia
17	Lateral horizontal ST-elevation	Lateral ischaemia and right ventricular hypertrophy	Lateral horizontal ST-depression	Inferior ischaemia
19	Slower R-wave progression		Faster R-wave progression	
22	Baseline shift		Baseline shift	
23	Reduced P-wave amplitude	Atrial fibrillation, junctional bradycardia, third-degree AV block	Increased P-wave amplitude	
25	Shorter QRS duration		Longer QRS duration with slurred S-wave	RBBB, LBBB, ventricular tachycardia, NICD, WPW pattern, and reduced EF
26	—		Deep and broad S-wave in V1 with monophasic broad lateral R-waves and negative T-waves	LBBB and reduced EF
27	P- and R-axis deviation to the left with increasing P- and R-wave amplitudes		P- and R-axis deviation to the right with decreasing P- and R-wave amplitudes	Low QRS voltage, left axis deviation, third-degree AV-block, atrial fibrillation, atrial flutter, SVT, junctional bradycardia, reduced EF, and 1-year mortality
30	Shorter QT-interval		Longer QT-interval	Prolonged QT interval, reduced EF, and 1-year mortality
31	R-axis deviation to the right	Right axis deviation	R-axis deviation to the left	Left axis deviation, LAFB, and LVH
32	Decreased precordial QRS-amplitude		Increased precordial QRS-amplitude	LVH

The influence of an ECG factor on median beat ECG morphology is determined using visual inspection of the factor traversals (Figure 2). A summary of the most important associations of every ECG factor with conventional ECG measurements, ECG diagnostic statements, reduced EF, and 1-year mortality is obtained by combining results from Figures 3, 4, and 5. EF, ejection fraction; LAFB, left anterior fascicular block; LBBB, left bundle branch block; LVH, left ventricular hypertrophy; NICD, non-specific intraventricular conduction delay; RBBB, right bundle branch block; RVH, right ventricular hypertrophy; SVT, supraventricular tachycardia, WPW: Wolff-Parkinson-White.

caused by differences in the population, as the predictive value of just age and sex is also lower than in the original paper. We observed that the predictors for 1-year mortality are increasing age, higher ventricular frequency, negative T-waves, and ST-depression and elevation, and prolonged QT interval, which are all known risk factors for mortality.^{31,32}

There are several limitations to acknowledge. Firstly, the algorithm is trained on a very large dataset with over 1 million ECGs, but we could not account for the imbalance in ECG abnormalities due to the unsupervised nature of training. Therefore, less common ECG abnormalities might not be accurately encoded, as also demonstrated by the lower performance on for example ischaemia classes and lower correlation coefficients of the reconstructed ECGs (see [Supplementary material online, Table S1](#)). Future studies could experiment with balancing the dataset based on labelled abnormalities and the effect it may have on encoding rare ECG abnormalities. Secondly, the reduced performance of the explainable pipeline in diagnosing low QRS voltage and left ventricular hypertrophy is most likely due to the inability of the VAE to always reconstruct the amplitude of the R-wave correctly (see [Supplementary material online, Table S1](#)). Further research in the field of generative models for ECGs is needed to address this limitation and to improve the reconstruction quality. Finally, only one DNN architecture was investigated for comparison to a 'black box' DNN, which was similar to the encoder of the VAE for accurate comparison. As the performance of the current architecture is on par with other state-of-the-art models for similar tasks in this and other research of our group, we do not expect much gain from other DNN architectures.^{4,10,22,33,34}

Future studies should focus on evaluating the use of inherently explainable pipelines on other ECG tasks, as the dimensionality reduction of our algorithm to 21 factors broadens the usability of DNNs greatly to much smaller labelled datasets than before. Another important perspective is using the approach on full 10-s rhythm ECGs, to take additional ECG information into account. Rhythm disorders that are not visible in the median ECG beat, such as second-degree AV block and premature ventricular and atrial complexes, could add interesting information to the model. Finally, explainability of the current approach is hampered by the fact that some of the factors in the current FactorECG are still ambiguous and represent multiple ECG changes at the same time. Further developments in the field of DNN-based feature generation are needed to better disentangle the ECG factors.

In conclusion, we leveraged a large dataset of over 1 million ECGs to train a generative DNN that learned 21 valid underlying anatomical and (patho)physiological explanatory factors of variation in median beat 12-lead ECG data. We showed that our pipeline is not only able to interpret ECGs with highly accurate performance on par with 'black box' DNNs but also provide improved explainability on which ECG morphologies were important. These findings demonstrate that inherently explainable pipelines should be the future of ECG interpretation, as they allow reliable clinical interpretation of these models without performance reduction, while also broadening their applicability to many other (rare) diseases.

Code availability

The decoder for the FactorECG is publicly available at <https://decoder.ecgx.ai>. Researchers can request the ECG factors for their

own ECGs using a tool at <https://encoder.ecgx.ai>. The code for training and evaluating the β -VAE and the black box DNN is available at <https://github.com/rutgervandeleur/ecgxai>.

Ethics committee approval

This study was approved by the University Medical Center Utrecht ethical committee with number 18–827.

Supplementary material

[Supplementary material](#) is available at European Heart Journal – Digital Health online.

Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

Funding

This study was financed by the Netherlands Organisation for Health Research and Development (ZonMw) with grant number 104021004 and the Dutch Heart Foundation with grant number 2019B011.

Conflict of interest: None declared.

Data availability

The training datasets used in this study are not openly available due to privacy concerns. The expert-annotated test set is available upon request to the corresponding author.

References

- van de Leur RR, Boonstra MJ, Bagheri A, Roudijk RW, Sammani A, Taha K, Doevendans PA, van der Harst P, van Dam P, Hassink R, van Es R, Asselbergs FW. Big data and artificial intelligence: opportunities and threats in electrophysiology. *Arrhythmia Electrophysiol Rev* 2020;**9**:146–154.
- Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, Ng AY. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med* 2019;**25**:65–69.
- van de Leur RR, Blom LJ, Gavves E, Hof IE, van der Heijden JF, Clappers NC, Doevendans PA, Hassink RJ, Es RV. Automatic triage of 12-lead electrocardiograms using deep convolutional neural networks. *J Am Heart Assoc* 2020;**9**:e015138.
- Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, Pelliikka 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.
- Raghunath S, Cerna AEU, Jing L, VanMaanen DP, Stough J, Hartzel DN, Leader JB, Kirchner HL, Stumpe MC, Hafez A, Nemani A, Carbonati T, Johnson KW, Young K, Good CW, Pfeifer JM, Patel AA, Delisle BP, Alsaid A, Beer D, Haggerty CM, Fornwalt BK. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nat Med* 2020;**26**:886–891.
- 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.
- Kundu S. AI In medicine must be explainable. *Nat Med* 2021;**27**:1328.
- 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.
- Goodman B, Flaxman S. European Union regulations on algorithmic decision-making and a "right to explanation.". *Ai Mag* 2017;**38**:50–57.
- 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**:e009056.
- Kwon J, 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, multicentre study. *Lancet Digital Heal* 2020;**2**:e358–e367.

12. 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.
13. Hooker S, Erhan D, Kindermans P-J, Kim B. A benchmark for interpretability methods in deep neural networks. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2019. 9737–9748.
14. Kingma DP, Welling M. Auto-Encoding variational Bayes. In: Bengio Y and LeCun Y (eds.), *2nd International Conference on Learning Representations*. Banff, AB, Canada: Conference Track Proceedings; 2014.
15. Higgins I, Matthey L, Pal A, Burgess CP, Glorot X, Botvinick M, Mohamed S, Lerchner A. Beta-VAE: learning basic visual concepts with a constrained variational framework. In: 5th International Conference on Learning Representations. Toulon, France: Conference Track Proceedings; 2017.
16. Petersen SE, Sanghvi MM, Aung N, Cooper JA, Paiva JM, Zemrak F, Fung K, Lukaschuk E, Lee AM, Carapella V, Kim YJ, Piechnik SK, Neubauer S. The impact of cardiovascular risk factors on cardiac structure and function: insights from the UK biobank imaging enhancement study. *PLoS One* 2017;**12**:e0185114.
17. Petersen SE, Aung N, Sanghvi MM, Zemrak F, Fung K, Paiva JM, Francis JM, Khanji MY, Lukaschuk E, Lee AM, Carapella V, Kim YJ, Leeson P, Piechnik SK, Neubauer S. Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in caucasians from the UK biobank population cohort. *J Cardiovasc Magn Reson* 2017;**19**:18.
18. Petersen SE, Matthews PM, Francis JM, Robson MD, Zemrak F, Boubertakh R, Young AA, Hudson S, Weale P, Garratt S, Collins R, Piechnik S, Neubauer S. UK Biobank's cardiovascular magnetic resonance protocol. *J Cardiovasc Magn Reson* 2016;**18**:8.
19. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015;**12**:e1001779.
20. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. East Lansing, MI, USA: ACM; 2016. p785–794.
21. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S and Garnett R (eds.), *Advances in neural information processing systems 30*: Curran Associates, Inc.; 2017. p4765–4774.
22. Bos MN, van de Leur RR, Vranken JF, Gupta DK, van der Harst P, Doevendans PA, van Es R. Automated comprehensive interpretation of 12-lead electrocardiograms using Pre-trained exponentially dilated causal convolutional neural networks. In: 2020 Comput Cardiol 2020: IEEE; 2020:1–4.
23. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multi-variable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;**162**: 55–63.
24. Kwon J, Lee SY, Jeon K, Lee Y, Kim K, Park J, Oh B, Lee M. Deep learning-based algorithm for detecting aortic stenosis using electrocardiography. *J Am Heart Assoc* 2020;**9**:e014717.
25. Jang JH, Kim TY, Lim HS, Yoon D. Unsupervised feature learning for electrocardiogram data using the convolutional variational autoencoder. *PLoS One* 2021;**16**: e0260612.
26. Yildirim O, Tan RS, Acharya UR. An efficient compression of ECG signals using deep convolutional autoencoders. *Cogn Syst Res* 2018;**52**:198–211.
27. Liu H, Zhao Z, Chen X, Yu R, She Q. Using the VQ-VAE to improve the recognition of abnormalities in short-duration 12-lead electrocardiogram records. *Comput Meth Prog Bio* 2020;**196**:105639.
28. Kuznetsov VV, Moskalenko VA, Gribov DV, Zolotykh NY. Interpretable feature generation in ECG using a variational autoencoder. *Front Genet* 2021;**12**:638191.
29. O'Neal WT, Mazur M, Bertoni AG, Bluemke DA, Al-Mallah MH, Lima JAC, Kitzman D, Soliman EZ. Electrocardiographic predictors of heart failure with reduced versus preserved ejection fraction: the multi-ethnic study of atherosclerosis. *J Am Heart Assoc* 2017;**6**:e006023.
30. Yao X, McCoy RG, Friedman PA, Shah ND, Barry BA, Behnken EM, Inselman JW, Attia ZI, Noseworthy PA. ECG AI-guided screening for low ejection fraction (EAGLE): rationale and design of a pragmatic cluster randomized trial. *Am Heart J* 2020;**219**:31–36.
31. Kannel WB, Kannel C, Paffenbarger RS, Cupples LA. Heart rate and cardiovascular mortality: the Framingham study. *Am Heart J* 1987;**113**:1489–1494.
32. Porthan K, Viitasalo M, Jula A, Reunanen A, Rapola J, Väänänen H, Nieminen MS, Toivonen L, Salomaa V, Oikarinen L. Predictive value of electrocardiographic QT interval and T-wave morphology parameters for all-cause and cardiovascular mortality in a general population sample. *Heart Rhythm* 2009;**6**:1202–1208.e1.
33. Kashou AH, Ko W-Y, Attia ZI, Cohen MS, Friedman PA, Noseworthy PA. A comprehensive artificial intelligence-enabled electrocardiogram interpretation program. *Cardiovasc Digital Heal J* 2020;**1**:62–70.
34. Siegersma KR, van de Leur RR, Onland-Moret NC, Leon DA, Diez-Benavente E, Rozendaal L, Bots ML, Coronel R, Appelman Y, Hofstra L, van der Harst P, Doevendans PA, Hassink RJ, den Ruijter HM, van Es R. Deep neural networks reveal novel sex-specific electrocardiographic features relevant for mortality risk. *Eur Heart J* 2022;**3**:245–254.