EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR THE ELECTROCARDIOGRAM

Rutger R. van de Leur

Rutger R. van de Leur | 2024

Cover design: Rutger van de Leur
Layout: Sofia Casini (@fullofpride on Fiverr)
Printing: Gildeprint Enschede | www.gildeprint.nl
ISBN: 978-90-393-7737-6

Copyright © 2024 Rutger Roel van de Leur
All rights reserved. No part of this publication may be reproduced, stored in an archival system, of transmitted in any form or by any means, without prior permission of the copyright owner.

Explainable artificial intelligence for the electrocardiogram

Inzichtelijke kunstmatige intelligentie voor het electrocardiogram

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof. dr. H.R.B.M. Kummeling,
ingevolge het besluit van het College voor Promoties
in het openbaar te verdedigen op

donderdag 10 oktober 2024 des middags te 12.15 uur

door

Rutger Roel van de Leur

geboren op 9 augustus 1994
te Arnhem

Financial support by the Dutch Heart Foundation for the publication of this thesis is gratefully acknowledged. The research described in this thesis was supported by a grant from Dutch Heart Foundation (2019B011). Further financial support was provided by LANCAE, Chipsoft and Stichting Genetische Hartspierziekte PLN.

Promotor
Prof. dr. P.A.F.M. Doevendans

Copromotoren
Dr. R. van Es
Dr. R.J. Hassink

Beoordelingscommissie
Prof. dr. F.W. Asselbergs
Prof. dr. D.W. Donker
Prof. dr. M. Meine
Prof. dr. D.L. Oberski (voorzitter)
Dr. M. van Smeden

Table of contents

Preface and thesis outline 8

Big Data and Artificial Intelligence: Opportunities and Threats in Electrophysiology 15

Automatic Triage of 12-Lead ECGs Using Deep Convolutional Neural Networks 49

Automatic Triage of 12-lead ECGs using Deep Convolutional Neural Networks: A First Implementation Study 75

Improving explainability of deep neural network-based electrocardiogram interpretation using variational 141

Variational auto-encoders improve explainability over currently employed heatmap methods for deep learning-based interpretation of the electrocardiogram 181

Discovering and Visualizing Disease-Specific Electrocardiogram Features Using Deep Learning: Proof-of-Concept in Phospholamban Gene Mutation Carriers 189

ECG-only Explainable Deep Learning Algorithm Predicts Risk of Malignant Ventricular Arrhythmia in Phospholamban Cardiomyopathy 213

Life-threatening ventricular arrhythmia prediction in patients with dilated cardiomyopathy using explainable electrocardiogram-based deep neural networks 237

Electrocardiogram-based deep learning improves outcome prediction following cardiac resynchronization therapy 259

General discussion and future perspectives 291

Appendices 313

English summary 316

Nederlandse samenvatting 317

List of publications 320

Curriculum vitae 323

Dankwoord 325

Preface and thesis outline

The electrocardiogram (ECG) is one of the most used diagnostic tools in clinical practice, with approximately 300 million ECGs obtained worldwide each year.[1] The first recording of the heart's electrical activity was documented around 1880. However, it was Willem Einthoven who invented the first practical method for conducting an ECG in 1901. [2] In the subsequent years, Einthoven established naming conventions for the different waves observed in an ECG (P, Q, R, S, and T waves) as well as leads (Einthoven I, II, and III). Furthermore, Einthoven outlined criteria for determining a normal ECG and identified some initial abnormalities such as pathological Q-waves, ventricular extrasystole, AV-block, abnormal heart axis, and left and right ventricular hypertrophy as early as 1912.[3,4] By then, he also already realized there was much more to gain from this technology:

> *"The method of electrocardiography is still a young plant.*
> *We may reasonably expect that it will continue to bear good fruit."*
> Prof. dr. Willem Einthoven, Lancet (1912)

In the following decades, many novel electrophysiological abnormalities were discovered using the ECG, from Wolff-Parkinson-White syndrome in 1930 to Brugada syndrome in 1992.[5,6] In 1938, Frank Wilson develops the Wilson Central Terminal, which allows for measurement of the precordial leads (V1-V6). These leads were used to improve the diagnosis of myocardial infarction, and together with the Einthoven and augmented Einthoven leads we arrived at the 12-lead ECG that is used today.[7]

The 12-lead ECG became a fundamental tool in the everyday practice of clinical medicine, and its correct interpretation pivotal for a wide spectrum of cardiac abnormalities. Interpretation of the ECG is a complex task, that requires integration of knowledge on anatomy, (patho)physiology and electrophysiology with pattern recognition and the ability to quickly fixate the critical lead(s).[8] This requires extensive training, and although accuracy of interpretation increases from 42% for medical students to 75% for cardiologists, physicians at all training levels have deficiencies in ECG interpretation.[9,10] Therefore, in pre-hospital care and non-cardiology departments, expert knowledge to interpret ECGs might not always be readily available, and referral for such ECGs remains necessary. The life-threatening nature of a suspected acute coronary syndrome or ventricular arrhythmia requires not only accurate, but also timely ECG interpretation and this places a heavy logistic burden on clinical practice. More problematic is that even experts often do not agree on the interpretation of the ECG with each other, as well as their own previous interpretation of the same ECG.[9,11,12]

These deficiencies in the ECG interpretation, most notably the high inter- and intra-rater variability and logistical burden, instigated the introduction of algorithms for the computerized interpretation of the ECG (CIE) around 1960.[13] Significant progress was made in the development of ECG algorithms, but current versions have not been able to reach physician level accuracy in diagnosing cardiac abnormalities.[14] Consensus is therefore that all computer-based reports should be systematically overread.[14] While this leads to a decreased analysis time for experienced readers and improved diagnostic abilities when the interpretation is correct, it increases probability of errors when the interpretation is erroneous.[14,15] One of early pioneers in the field of CIE therefore concluded in 1966 that "computerized ECG interpretation will elude us until the next generation of cardiologists".[12]

Around the same time development in the field of CIE started, the term artificial intelligence (AI) emerged at a conference. AI refers to mimicking human intelligence in computers to perform tasks not explicitly programmed. The research field grew exponentially, and it was assumed that machines would soon become capable of doing anything a man can do. Around 1975 it turned out that early AI systems could not live up to the hype, and the first AI winter began. After a second unsuccessful AI 'boom', the field went into another AI winter around 2000. Due to access to large amounts of data, cheaper and faster computers, the third AI 'boom' started around 2010 and is currently probably reaching its peak. Research from the fields of image classification and speech recognition showed that one specific type of AI algorithm, called deep neural network (DNN), might be highly effective in the processing of raw data without the need for hand-crafted or rule-based feature engineering.[16,17] DNNs are computer algorithms based on the structure of the human brain and consist of layers of neurons that can be trained to discover complex patterns in images and signals using large datasets.[18]

Therefore, not one, but two generations later, developments in CIE and AI coincide and a substantial improvement of CIE is forthcoming. In 2018, at the start of this thesis project, Awni Hannun and colleagues were the first to robustly show that DNNs can learn a broad range of arrhythmias from the ECG with diagnostic performance similar to that of cardiologists.[19] The ECG turned out to be an ideal substrate for developing deep learning-based AI algorithms, especially because large, labeled ECG datasets were readily available. Since 2018, our research group at the University Medical Center Utrecht and other groups have shown that DNNs can not only be applied to enhance automated diagnosis, but also to detect (asymptomatic) cardiovascular disease that might not be readily apparent, even to expert eyes.[20,21]

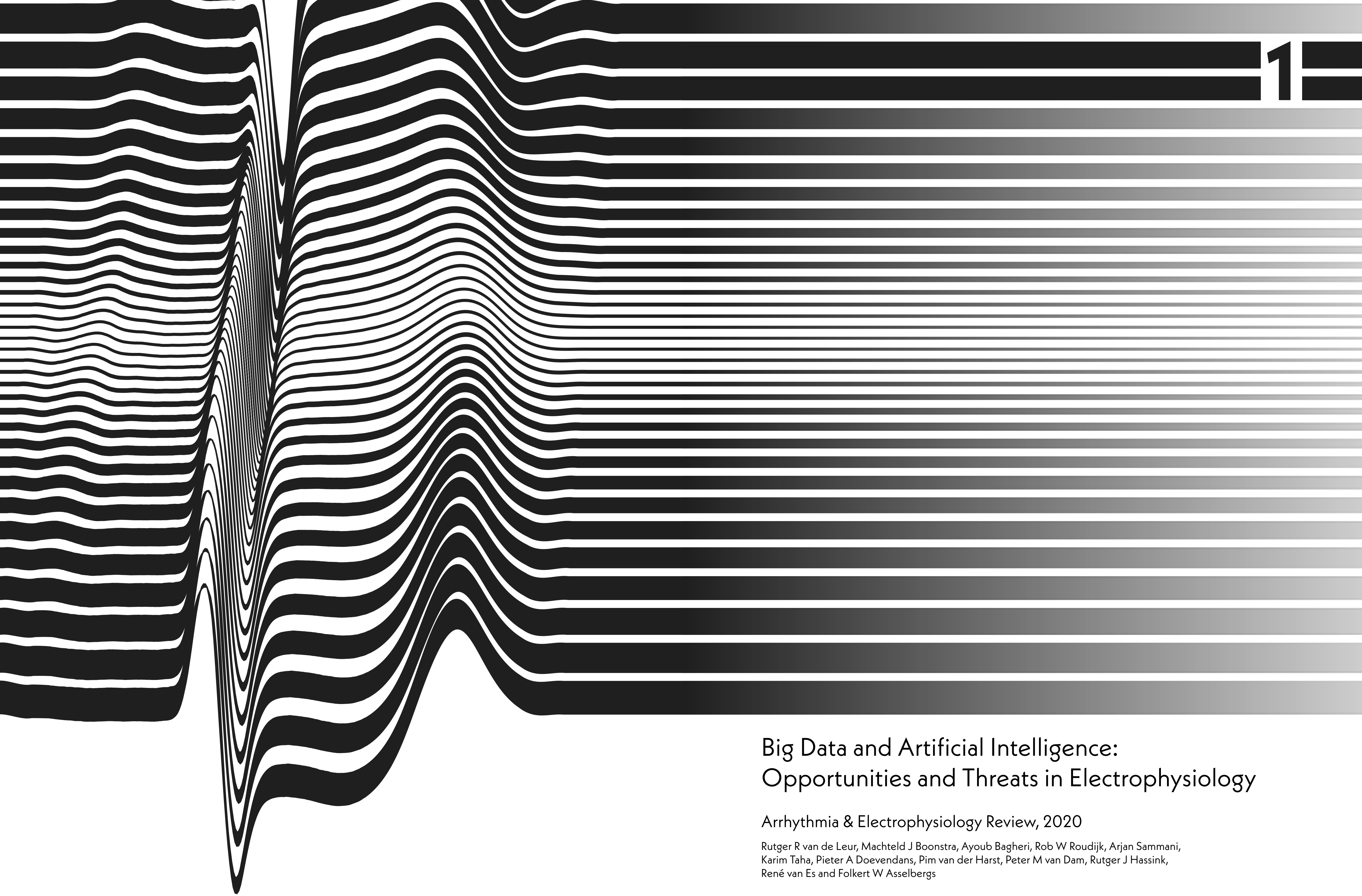
In the present thesis, we first investigated opportunities and treats for artificial intelligence in electrocardiography in a narrative review (**Chapter 1**) Next, we sought to develop a deep learning algorithm to optimize the diagnostic workflow on a large dataset with over 300.000 ECGs labelled by physicians. As we envisioned that the detailed ECG interpretation step would remain an expert task, we developed and validated an algorithm that could triage ECGs from normal to acute in **Chapter 2**. We showed that the algorithm performed excellently, but with this exciting progress challenges with implementing such algorithms in clinical practice became apparent. In the following chapters, we sought to address and find solutions to some of these challenges. Firstly, on the path towards clinical applicable AI for the ECG, we performed an implementation study with the triage algorithm. We investigated whether implementation of the triage algorithm in the hospital setting is safe and efficacious when considering clinical outcomes in **Chapter 3**. Secondly, we proposed a method to quantify how 'certain' an algorithm is in its prediction in **Chapter 4.** This measure of certainty was then evaluated as a safeguard to determine which ECGs will be automatically analyzed.

Next to the estimation of uncertainty, two other challenges with using deep neural networks arose: the lack of explainability and the need for very large datasets. Therefore, we designed and developed a novel method that leverages the power of DNN to interpret ECGs in an explainable manner. By training a variational auto-encoder on 1.1 million median beat ECGs, we were able to decompose the ECG morphology into 32 explainable factors (the FactorECG). In **Chapter 5**, we demonstrated that this explainable method performs on par with 'black box' DNNs for conventional ECG interpretation, but also for novel applications such a detection of reduced ejection fraction. Moreover, in **Chapter 6** we discussed why the FactorECG provides improved explainability over the heatmap-based methods used before.

Finally, as datasets with thousands of ECGs are not available for many clinically relevant questions, we evaluated the feasibility to transfer the knowledge that DNNs learned on *big data* to *small data*. In **Chapter 7**, we developed a deep learning algorithm to detect phospholamban (*PLN*) p.Arg14del variant carriers and used heatmaps to visualize which features were used by the algorithm. Afterwards, we applied our novel method, the FactorECG, to predict which *PLN* p.Arg14del variant carriers develop malignant ventricular arrhythmia (**Chapter 8**). We also applied the method in patients with dilated cardiomyopathy to predict ventricular arrhythmias (**Chapter 9**) and in patients that received cardiac resynchronization therapy to predict response to treatment and mortality (**Chapter 10**). The findings are discussed in **Chapter 11**.

REFERENCES

1. Holst H, Ohlsson M, Peterson C, Edenbrandt L. A confident decision support system for interpreting electrocardiograms. Clin Physiol 1999;19:410--418.
2. Waller AD. A Demonstration on Man of Electromotive Changes accompanying the Heart's Beat. J Physiology 1887;8:229–234.
3. Einthoven W. THE DIFFERENT FORMS OF THE HUMAN ELECTROCARDIOGRAM AND THEIR SIGNIFICATION. Lancet 1912;179:853–861.
4. Einthoven W. Het tele-cardiogram. Ned Tijdschr Geneeskd 1906:1517–1547.
5. Wolff L, Parkinson J, White PD. Bundle-branch block with short P-R interval in healthy young people prone to paroxysmal tachycardia. Am Heart J 1930;5:685–704.
6. Brugada P, Brugada J. Right bundle branch block, persistent ST segment elevation and sudden cardiac death: A distinct clinical and electrocardiographic syndrome A multicenter report. J Am Coll Cardiol 1992;20:1391–1396.
7. Wilson FN, Johnston FD, Rosenbaum FF, Erlanger H, Kossmann CE, Hecht H, Cotrim N, Oliveira RM de, Scarsi R, Barker PS. The precordial electrocardiogram. Am Heart J 1944;27:19–85.
8. Wood G, Batt J, Appelboam A, Harris A, Wilson MR. Exploring the Impact of Expertise, Clinical History, and Visual Search on Electrocardiogram Interpretation. Med Decis Making 2013;34:75–83.
9. Salerno SM, Alguire PC, Waxman HS. Competency in Interpretation of 12-Lead Electrocardiograms: A Summary and Appraisal of Published Evidence. Ann Intern Med 2003;138:751.
10. Cook DA, Oh S-Y, Pusic MV. Accuracy of Physicians' Electrocardiogram Interpretations. Jama Intern Med 2020;180:1461.
11. Holmvang L, Hasbak P, Clemmensen P, Wagner G, Grande P. Differences between local investigator and core laboratory interpretation of the admission electrocardiogram in patients with unstable angina pectoris or non-Q-wave myocardial infarction (a thrombin inhibition in myocardial ischemia [TRIM] substudy). Am J Cardiol 1998;82:54--60.
12. Simonson E, Tuna N, Okamoto N, Toshima H. Diagnostic accuracy of the vectorcardiogram and electrocardiogram A cooperative study. Am J Cardiol 1966;17:829–878.
13. PIPBERGER HV, FREIS ED, TACK L, ON HL. Preparation of Electrocardiographic Data for Analysis by Digital Electronic Computer. Circulation 1960;21:413–418.
14. Schläpfer J, Wellens HJ. Computer-Interpreted Electrocardiograms Benefits and Limitations. J Am Coll Cardiol 2017;70:1183–1192.
15. Southern WN, Arnsten JH. The effect of erroneous computer interpretation of ECGs on resident decision making. Med Decis Making 2009;29:372--376.
16. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster DR. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. Jama 2016;316:2402.
17. Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. 2013. p6645--6649.
18. Goodfellow I, Bengio Y, Courville A. Deep Learning. The MIT Press; 2016.
19. 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.
20. Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. Nature Reviews Cardiology 2021.
21. Leur RR van de, Boonstra MJ, Bagheri A, Roudijk RW, Sammani A, Taha K, Doevendans PA, Harst P van der, Dam P van, Hassink R, Es R van, Asselbergs FW. Big Data and Artificial Intelligence: Opportunities and Threats in Electrophysiology. Arrhythmia Electrophysiol Rev 2020;9:146–154.



1

Big Data and Artificial Intelligence: Opportunities and Threats in Electrophysiology

Arrhythmia & Electrophysiology Review, 2020

Rutger R van de Leur, Machteld J Boonstra, Ayoub Bagheri, Rob W Roudijk, Arjan Sammani,
Karim Taha, Pieter A Doevendans, Pim van der Harst, Peter M van Dam, Rutger J Hassink,
René van Es and Folkert W Asselbergs

Abstract

The combination of big data and artificial intelligence (AI) has an increasing impact in the field of electrophysiology. Algorithms are created to improve the automated diagnosis of clinical electrocardiograms (ECGs) or ambulatory rhythm devices. Furthermore, the use of AI during invasive electrophysiological studies or combining several diagnostic modalities into AI algorithms to aid diagnostics, are being studied. However, the clinical performance and applicability of created algorithms are yet unknown. In this narrative review, opportunities and threats of AI in the field of electrophysiology are described, mainly focusing on ECG-based AI. Current opportunities are discussed with their potential clinical benefits and concomitant challenges. Challenges in data acquisition, model performance, (external) validity, clinical implementation, algorithm interpretation as well as the ethical aspects of AI-minded research are discussed. This review aims to guide clinicians in the evaluation of the many exciting new and upcoming AI applications in electrophysiology, prior to clinical implementation.

Introduction

Artificial intelligence (AI) and big data-minded clinical research may aid the prediction and/or detection of (subclinical) cardiovascular diseases by providing additional knowledge about disease onset, progression or outcome. Clinical decision making, disease diagnostics, risk prediction or individualized therapy may be facilitated through new insights obtained from AI algorithms. As health records have become electronic, data of large populations are becoming increasingly accessible.[1] Within electrophysiology, the use of AI algorithms may be of particular interest, as large datasets of electrocardiograms (ECG) are often readily available. Moreover, data are continuously generated by implantable devices (e.g. pacemakers, implantable cardioverter defibrillators or loop recorders) or smartphone- and smartwatch applications.[2–6]

Interpretation of ECGs relies on expert opinion and requires training and clinical expertise which is subjected to considerable inter- and intra-clinician variability.[7–12] To facilitate clinical decision making, algorithms for the computerized interpretation of ECGs have been developed. However, these algorithms lack accuracy and may provide inaccurate diagnoses which may result in misdiagnosis when not reviewed carefully.[13–18]

Recently, substantial progress in the development of AI in electrophysiology has been made, which was mainly directed at ECG-based deep neural networks (DNN). For example, DNNs have been tested to identify arrhythmias, to classify supraventricular tachycardias, to predict left ventricular ejection fraction, to identify disease development in serial ECG measurements, to predict left ventricular hypertrophy and to perform comprehensive triage of ECGs.[6,19–23] DNNs are likely to aid non-cardiologists with improved ECG diagnostics and may provide the opportunity to expose yet undiscovered ECG characteristics indicating disease.

With this exciting progress, the challenges and threats of using AI techniques in clinical practice become apparent. In this narrative review, recent progress of AI in the field of electrophysiology is discussed together with opportunities and threats.

Artificial intelligence: A brief introduction

AI refers to mimicking human intelligence in computers to perform tasks not explicitly programmed. Machine learning (ML) is a branch of AI concerned with algorithms to train a model to perform a task. Two types of ML algorithms are supervised learning and unsupervised learning. Supervised learning refers to ML algorithms where input data are labelled with the outcome and the algorithm is trained to approximate the relation between input data and outcome. In unsupervised learning, input data are not labelled and the algorithm may discover data clusters in the input data.

In ML, an algorithm is trained to classify a dataset based on several statistical and probability analyses. In the training phase, model parameters are iteratively tuned by penalizing or rewarding the algorithm based on a true or false prediction. Deep learning is a sub-category of ML that uses neural networks as architecture to represent and learn data which are referred to as DNNs. The main difference between deep learning and other ML algorithms is that DNNs can learn from raw data (i.e. ECG waveforms) in an end-to-end manner; both feature extraction and classification are united in the algorithm (**Figure 1a**). For example, in ECG-based DNNs a matrix containing the time-stamped raw voltage values of each lead are used as input data. In other ML algorithms, features like heart rate or QRS duration are first (manually) extracted from the ECG and used as input data for the classification algorithm.

To influence the speed and quality of the training phase, the setting of hyperparameters (e.g. the settings of the model architecture and training) is important. Furthermore, overfitting or underfitting the model to the available dataset must be prevented. Overfitting can occur when a complex model is trained using a small dataset. Then the model precisely describes the training dataset but fails to predict the outcome using other data (**Figure 1b**). On the other hand, when constraining the model too much, underfitting occurs (**Figure 1b**), also resulting in poor algorithm performance. To assess overfitting, a dataset is usually divided into a training dataset, a validation dataset and a test dataset or resampling methods are used, such as cross-validation or bootstrapping.[24]

To train and test ML algorithms, and in particular DNNs, preferably a large dataset (big data) is used. Performance of highly dimensional algorithms (e.g. algorithms with many model parameters), like DNNs, depends on the size of the dataset. For deep learning, more data is often required as DNNs have many non-linear parameters and non-linearity increases the flexibility of an algorithm. The size of a training dataset has to reasonably approximate the relation between input data and outcome and the amount of testing data has to reasonably approximate the performance measures of the DNN. Determining the exact size of a training and testing dataset is difficult.[25,26] It depends on the complexity of the algorithm (e.g. the number of variables), the type of the algorithm, the number of outcome classes and the difficulty to distinguish between outcome classes as inter-class differences might be subtle. Therefore, size of the dataset should be carefully reviewed per algorithm. A rule of thumb for the adequate size of a validation dataset is between 50-100 patients per outcome class to determine overfitting. Recent studies published in the field of ECG-based DNNs used between 50 thousand and 1.2 million patients; these numbers illustrate the amount of data used for this type of analyses.[6,19,21,27]

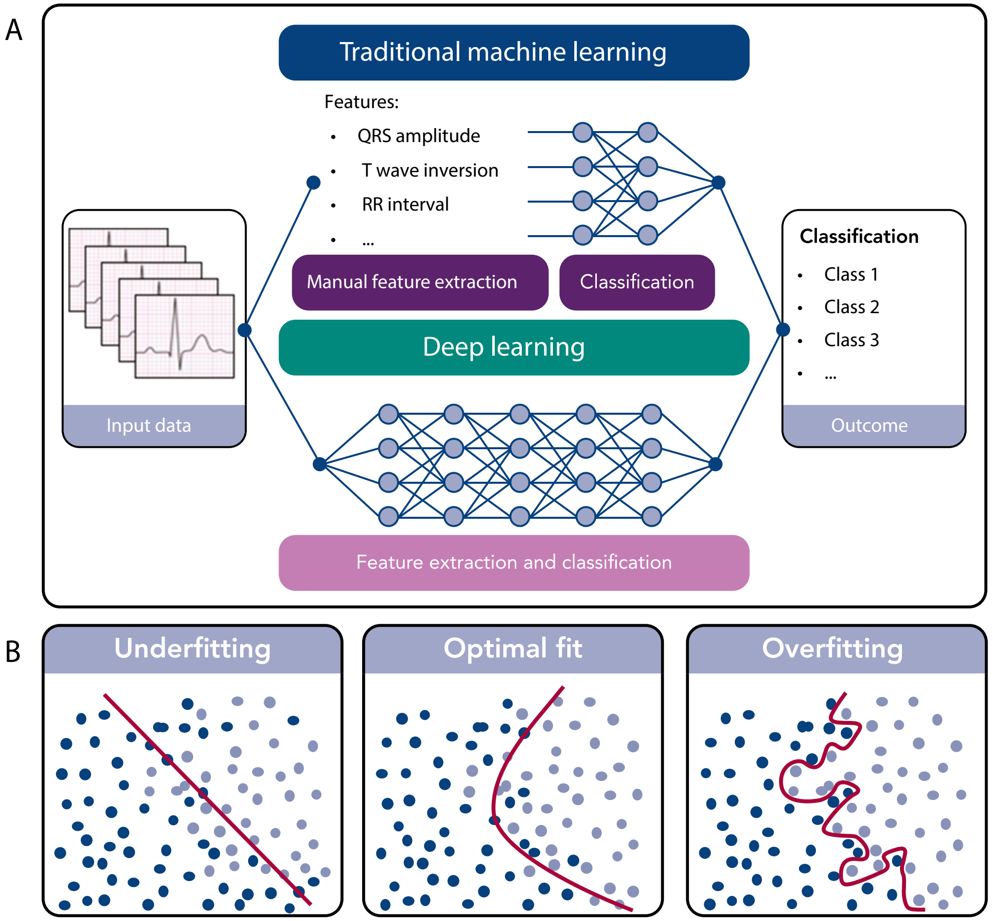


Figure 1. *A: traditional machine learning and deep learning and B: schematic representation fitting a model to a dataset.*

Prerequisites for AI in electrophysiology

Preferably, data used to create AI algorithms is objective, as subjectivity may introduce bias in the algorithm. To ensure clinical applicability of created algorithms, ease of access to input data and difference in data quality in different clinical settings and intended use of the algorithm should be considered. In this section, we mainly focus on the data quality of ECGs, as these data are easily acquired, and large datasets are readily available.

Technical specifications of ECGs

The ECG is obtained via electrodes on the body surface using an ECG device. The device samples the continuous body surface potentials and the recorded signals are filtered to obtain a clinically interpretable ECG.[28] As the diagnostic information of the ECG is contained below 100 Hz, a sampling rate of at least 200 Hz is required according to the Nyquist theorem.[29–33] Furthermore, an adequate resolution of at least 10µV is recommended to also obtain small amplitude fluctuations of the ECG signal. In the recorded signal, muscle activity, baseline wander, motion artefacts and powerline artefacts are also present, distorting the measured ECG. To remove noise and obtain an easily interpretable ECG, a combination of a high-pass filter of 0.67 Hz and a low-pass filter of 150-250 Hz is recommended, often combined with a notch filter of 50 Hz or 60 Hz. The inadequate setting of these filters might result in a loss of information such as QRS fragmentation or notching, slurring or distortion of the ST segment. Furthermore, a loss of QRS amplitude of the recorded signal might be the result of the inappropriate combination of a high-frequency cutoff and sampling frequency.[28,34] ECGs used as input for DNNs are often already filtered, thus potentially relevant information might already be lost. As DNNs process and interpret the input data differently; filtering might be unnecessary and potentially relevant information may be preserved. Furthermore, as filtering strategies differ between manufacturers and even different versions of ECG devices, the performance of DNNs might be affected when ECGs from different ECG devices are used as input data.

Apart from applied software settings like sampling frequency or filter settings, the hardware of ECG devices also differs between manufacturers. Differences in analogue to digital converters, type of electrodes used, or amplifiers also affect recorded ECGs. The effect of input data recorded using different ECG devices on the performance of AI algorithms is yet unknown. However, as acquisition methods may differ significantly between manufacturers, the performance of algorithms are likely to depend on the type or even version of the device.[35] Testing the performance of algorithms using ECGs recorded by different devices would illustrate the effect of these technical specifications on performance and generalizability.

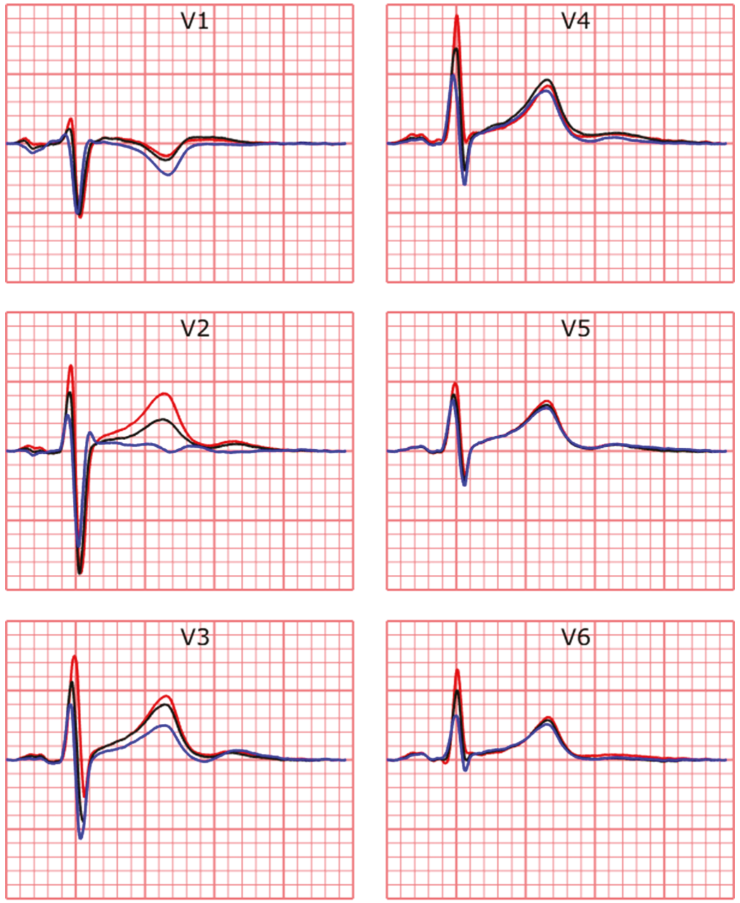


Figure 2. ***The effect of shifting precordial electrodes 4 cm upward (blue) or downward (red) from standard 12-lead electrode positioning (black).*** Displayed signals were simultaneously recorded using a 64-electrode measurement setup.

ECG electrodes

The recorded ECG is affected by electrode position with respect to the anatomical position of the heart and displacement of electrodes may result in misdiagnosis in a clinical setting.[36,37] For example, placement of limb electrodes on the trunk significantly affects the signal waveforms and lead reversal may mimic pathological conditions.[38–41] Furthermore, deviations in precordial electrode positions affect QRS and T-wave morphology (**Figure 2**). Besides the effect of cardiac electrophysiological characteristics like anisotropy, His-Purkinje anatomy, myocardial disease and cardiac anatomy on measured ECGs, cardiac position and cardiac movement also affect the ECG.[42–45]

Conventional clinical ECGs mostly consist of the measurement of eight independent signals; two limb leads and six precordial leads (**Figure 3b**). The remaining four limb leads are derived from the measured limb leads. However, body surface mapping studies identified the number of signals containing unique information up to twelve for ventricular depolarization and up to ten for ventricular repolarization.[46] Theoretically, to measure all information about cardiac activity from the body surface, the number of electrodes should be at least the number of all unique measurements. However, conventional 12-lead ECG is widely accepted for most clinical applications. Only when a posterior or right ventricle myocardial infarction or Brugada syndrome is suspected an adjustment of a lead position is considered.[27,47–50]

The interpretation of ECGs through computers and humans is fundamentally different and factors like electrode positioning or lead misplacement might influence algorithms. However, the effect of electrode misplacement or reversal, disease-specific electrode positions or knowledge of lead positioning on the performance on DNNs remains yet to be identified. A recent study was able to identify misplaced chest electrodes, implying that the effect of electrode misplacement might be identified and taken into account by algorithms.[51] Interestingly, studies suggested that DNNs can achieve similar performance when fewer leads are used.[50]

ECG input data format

ECGs can be obtained from the electronic database in three formats; as visualized signals (as used in standard clinical practice), as raw ECG signals or as median beats. To be used as input for DNNs, preferably, ECG signals consist of raw signals as visualized signals require digitization, consequently resulting in loss of signal resolution. Furthermore, raw ECG signals often consist of a continuous 10-second measurement of all recorded leads, whereas visualized signals may consist of 2.5 seconds per lead with only three simultaneously recorded signals per 2.5 seconds (**Figure 3**). A median beat per lead can also be used, computed from measured raw ECG signals or digitized visualized signals. Using the median beat might reduce noise, as noise is expected to cancel out by averaging all beats. Therefore, subtle changes in cardiac activation, invisible due to noise might become distinguishable for the algorithm. The use of the median beat may allow for precise analysis of waveform shapes or serial changes between individuals, but rhythm information is lost.

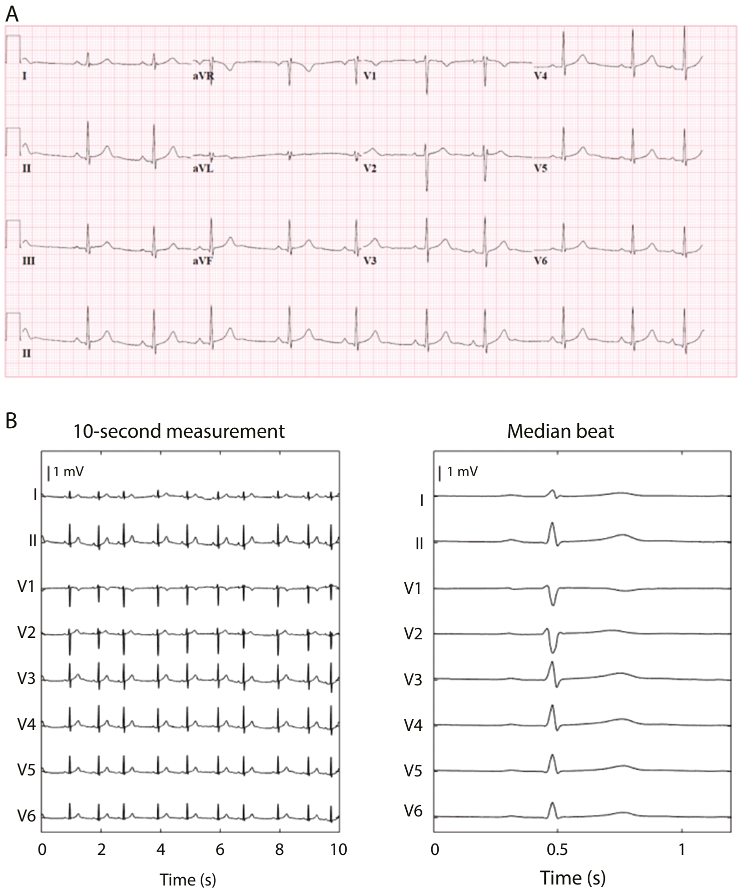


Figure 3. ***Standardized clinical visualized signals (A) with three simultaneously recorded 2.5-seconds measurements and raw signals (B) containing a10-second measurement and median beats of all recorded leads.*** Displayed signals are acquired using a General Electric Healthcare MAC5500 ECG device.

Opportunities for artificial intelligence in electrophysiology

Enhanced automated ECG diagnosis

An important opportunity of AI in electrophysiology is the enhanced automated diagnosis of clinical 12-lead ECGs.[8,11,12,20,52–54] Adequate computerized algorithms are especially important when expert knowledge is not readily available, like in prehospital care, non-cardiology departments, or low-resource facilities. If high-risk patients can be identified correctly, time-to-treatment can be reduced. However, currently available computerized ECG diagnosis algorithms lack accuracy.[11] Recently, progress has been made to use DNNs to automate diagnosis or triage of ECGs to improve time-to-treatment and decrease workload.[19,55] Using very large datasets, DNNs can achieve high diagnostic performance and outperform cardiology residents and non-cardiologists.[6,19] Moreover, progress has been made to use ECG data for predictive modeling such as atrial fibrillation in sinus rhythm ECGs or for the screening of hypertrophic cardiomyopathy.[56–58]

Combining other diagnostic modalities with ECG-based DNN

Interestingly, some studies suggest the possibility to use ECG-based DNNs combined with other diagnostic modalities to screen for disorders that are currently not associated with the ECG. In these applications, DNNs are thought to be able to detect subtle ECG changes. For example, when combined with large laboratory datasets patients with hyperkalemia could be identified, or when combined with echocardiographic results, reduced ejection fraction or aortic stenosis could be identified. In these three applications, the created DNNs identified these disorders from the ECG with high accuracy.[21,50,59] As a next step, supplementing ECG-based DNNs with body surface mapping data with a high spatial resolution (e.g. more than 12 measurement electrodes), inverse electrocardiography data or invasive electrophysiological mapping data, may result in the identification of subtle changes in the 12-lead ECG as a result of pathology.

Artificial intelligence for invasive electrophysiological studies

The application of AI before and during complex invasive electrophysiological procedures, like electro-anatomical mapping, is another major opportunity. By combining information from several diagnostic tools like magnetic resonance imaging (MRI), fluoroscopy or previous electro-anatomical mapping procedures, invasive catheter ablation procedure time might be reduced through the accelerated identification of arrhythmogenic substrates. Also, new techniques such as ripple mapping may be of benefit during electro-anatomical mapping studies.[60] Recent studies suggest that integration of fluoroscopy and electro-anatomical mapping with MRI is feasible using conventional statistical techniques or ML, whereas others suggest the use of novel anatomical mapping systems to circumvent fluoroscopy.[61–64] Furthermore, several ML algorithms were able to identify myocardial tissue properties using electrograms in vitro.[65]

Ambulatory device-based screening for cardiovascular diseases

One of the major current challenges in electrophysiology is the applicability of ambulatory rhythm devices into clinical practice. Several tools, such as implantable devices or smartwatch and smartphone-based devices, are becoming more widely used and continuously generate large amounts of data impossible for manual evaluation.[66] Arrhythmia detection algorithms based on DNNs trained on large cohorts of ambulatory patients with a single-lead plethysmography or ECG device have shown similar diagnostic performance as cardiologists or implantable loop recorders.[2,3,6] Another interesting application of DNNs algorithms are data from intracardiac electrograms before and during the activation of the defibrillator. Analysis of the signals before the adverse event might provide insight into the mechanism of the ventricular arrhythmia, providing the clinician with valuable insights. Continuous monitoring also provides the possibility to identify asymptomatic cardiac arrhythmias or detect complications post-surgery. Early detection of both might overcome serious adverse events and significantly improve fast and personalized healthcare.[6,19]

A promising benefit of smartphone-based applications for the early detection of cardiovascular disease is in early detection of atrial fibrillation (AF). As AF is a risk factor for stroke, early detection to prompt adequate (anticoagulant) treatment may be of importance.[67–69] Using smartphone or smartwatch acquired ECGs, an irregular rhythm can be accurately detected. Even the prediction whether a patient will develop AF in the future using smartphone acquired ECGs recorded during sinus rhythm has been recently reported.[69,70]Also, camera-based photoplethysmography recordings can be used to differentiate between irregular and regular cardiac rhythm.[71,72] However, under-detection of asymptomatic AF is expected as the use of applications requires active use and individuals are likely to only use applications in case of complaints. Therefore, a non-contact method with facial photoplethysmography recordings during regular smartphone use may be an interesting option to explore.[70,73,74]

Apart from the detection of (asymptomatic) AF, the prediction or early detection of ventricular arrhythmias using smartphone-based techniques are potentially clinically relevant. For example, smartphone-based monitoring of individuals with a known pathogenetic mutation might aid the early detection of disease onset. In some pathogenetic mutations, this may be especially relevant as sudden cardiac death can be the first manifestation of the disease. In these patients, close monitoring to prevent these adverse events by starting early treatment when subclinical signs are detected may provide clinical benefit.

Threats of artificial intelligence in electrophysiology

Data-driven versus hypothesis-driven research

Data from the electronic health records are almost always retrospectively collected, leading to data-driven research, instead of hypothesis-driven research. Research questions are often formulated based on readily available data, which increases the possibility of incidental findings and spurious correlations. While correlation might be sufficient for some predictive algorithms, causal relationships remain utmost important to define pathophysiological relationships and ultimately for the clinical implementation of AI algorithms. Therefore, big data research is argued to be in most cases solely used for hypothesis-generation and controlled clinical trials remain necessary to validate these hypotheses. When AI is used to identify novel pathophysiologic phenotypes (e.g. with specific ECG features), sequential prospective studies and clinical trials are crucial.[75]

Input data

For supervised learning, adequate labelling of input data is important.[18,76,77] Inadequate labelling of ECGs or the presence of for example pacemaker artefacts, comorbidities affecting the ECG or medication affecting the rhythm or conduction might influence the performance of DNNs.[13–18] Instead of true disease characteristics, ECG changes due to clinical interventions are used by the DNN to classify ECGs. For example, a DNN using chest X-rays provided insight into long-term mortality[78] but the presence of a thoracic drain and inadequate labelled input data resulted in an algorithm unsuitable for clinical decision making.[77–80] Therefore, the critical review of computerized labels and the identification of important features used by the DNN are essential.

Data extracted from ambulatory devices consist of real-time continuous monitoring data outside the hospital. As the signal acquisition is performed outside a standardized environment, signals are prone to errors. ECGs are more often exposed to noise due to motion artefacts, muscle activity artefacts, loosening or moved electrodes and alternating powerline artefacts. To accurately assess ambulatory data without the interference of artefacts, signals should be denoised or a quality control mechanism should be implemented. For both methods, noise should be accurately identified, after which adaptive filtering or noise qualification can be implemented.[81–83] However, as filtering might remove information, rapid real-time quality reporting of the presence of noise in the acquired signal is thought to be beneficial. With concise instructions, users can make adjustments to reduce artefacts and the quality of the recording will improve. Different analysis requires different data quality levels and through classification recorded data quality, the threshold for user notification can be adjusted per analysis.[84,85]

Generalizability and clinical implementation

With the increasing number of studies on ML algorithms, generalizability and implementation is one of the most important challenges to overcome. Diagnostic or prognostic prediction model research, from simple logistic regression to highly sophisticated DNNs, is characterized by three phases: 1) development and internal validation, 2) external validation and updating for other patients and 3) assessment of implementation of the model in clinical practice and its impact on patient outcomes.[86,87] During internal validation, the predictive performance of the model is assessed using the development dataset through train-test splitting, cross-validation or bootstrapping. Internal validation is however insufficient to test generalizability of the model in 'similar but different' individuals. Therefore, external validation of established models is important before clinical implementation. A model can be externally validated through temporal (same institution, later period), geographical (another institution with a similar patient group) or domain (different patient group) validation. Finally, implementation studies, such as cluster randomized trials, before-after studies or decision-analytic modelling studies, are required to assess the effect of implementing the model in clinical care.[86,87]

Most studies in automated ECG prediction and diagnosis performed some type of external validation. However, no study using external validation in a different patient group or implementation study has been published so far.[52] Recently, a study showed similar accuracy to predict low ejection fraction from the ECG using a DNN through temporal validation as in the development study.[88] A promising finding was a similar performance of the algorithm for different ethnic subgroups, even if the algorithm was trained on one subgroup.[89] As a final step to validate this algorithm, a cluster-randomized trial is currently being performed. This might provide valuable insight into the clinical usefulness ECG-based DNNs.[90]

Implementation studies for algorithms using ambulatory plethysmography and ECG data are ongoing. For example, the Apple Heart Study assessed the implementation of smartphone-based atrial fibrillation detection.[5] Over 400.000 patients were included using a mobile application, but only 450 patients were analyzed. Implementation was proven feasible as the number of false alarms was low, but the study lacks insight into the effect of smartphone-based atrial fibrillation detection on patient outcome. Currently, the HEARTLINE trial is performed and patients are randomized to use the smartwatch monitoring device. The need for treatment with anticoagulation of patients with device-detected subclinical AF is also being investigated.[4]

A final step for the successful clinical implementation of AI is to inform its users about adequate use of the algorithm. Standardized "Model Facts" leaflets have been proposed to instruct clinicians when, and more importantly when not, to use an algorithm.[91] This is particularly important if an algorithm is trained on a cohort using a specific subgroup of patients. Then, applying the model to a different population may potentially result in misdiagnosis. Therefore, describing the predictive performance in different (sub)groups (such as different age, sex, ethnicity and disease stage) is of utmost importance as AI algorithms are able to identify these by themselves.[89,92–94] However, as most ML algorithms are still considered to be 'black boxes', algorithm bias might remain difficult to detect.

Interpretability

Many sophisticated ML methods are considered 'black boxes' as they have many model parameters and abstractions. This is in contrast with the more conventional statistical methods used in medical research, like logistic regression and decision trees, where the influence of a predictor on the outcome is clear. The trade of complexity of models and interpretability for improved accuracy is important to acknowledge; with increased complexity of the network, interpretation becomes more complicated. But interpretability remains important to investigate false positives and negatives, to detect biased or overfitted models, to improve trust in new models or to use the algorithms as a feature detector.[95] Within electrophysiology, few studies investigated how the AI algorithms came to a certain result. For DNNs, three recent studies visualized individual examples using Guided Grad-CAM, a technique to show where the networks focus on. They showed that the DNN used the same segment of the ECG as a physician would (**Figure 4**).[19,27,96–98]

Visualization techniques may provide the ECG locations which the algorithms find important, but do not provide the specific feature. Therefore, the opportunity to identify additional ECG features remains dependent on expert opinion and analysis of the data by a clinician remains required. Visualization techniques and their results are promising and help to increase trust in DNNs for ECG analysis, but additional work is needed to further improve the interpretability of AI algorithms in clinical practice.[99,100]

Uncertainty estimation

In contrast to physicians or conventional statistical methods, DNNs struggle to inform their users *when they don't know*, i.e. to give uncertainty measures around their predictions. Current models always output a diagnosis or prediction, even if they have not seen the input before. In a real-world setting, clinicians acknowledge uncertainty and consult colleagues or literature, a DNN always makes a prediction. Therefore, methods that incorporate uncertainty are essential before implementation of such algorithms is possible.[101]

Ideally, the algorithm provides only results when it reaches a high threshold of certainty, while the uncertain cases will still be reviewed by a clinician.[101] For DNNs, several new techniques are available to obtain uncertainty measures, such as Bayesian deep learning, Monte Carlo dropout and ensemble learning, but these have never been applied in electrophysiologic research.[102] They have been applied to detect diabetic

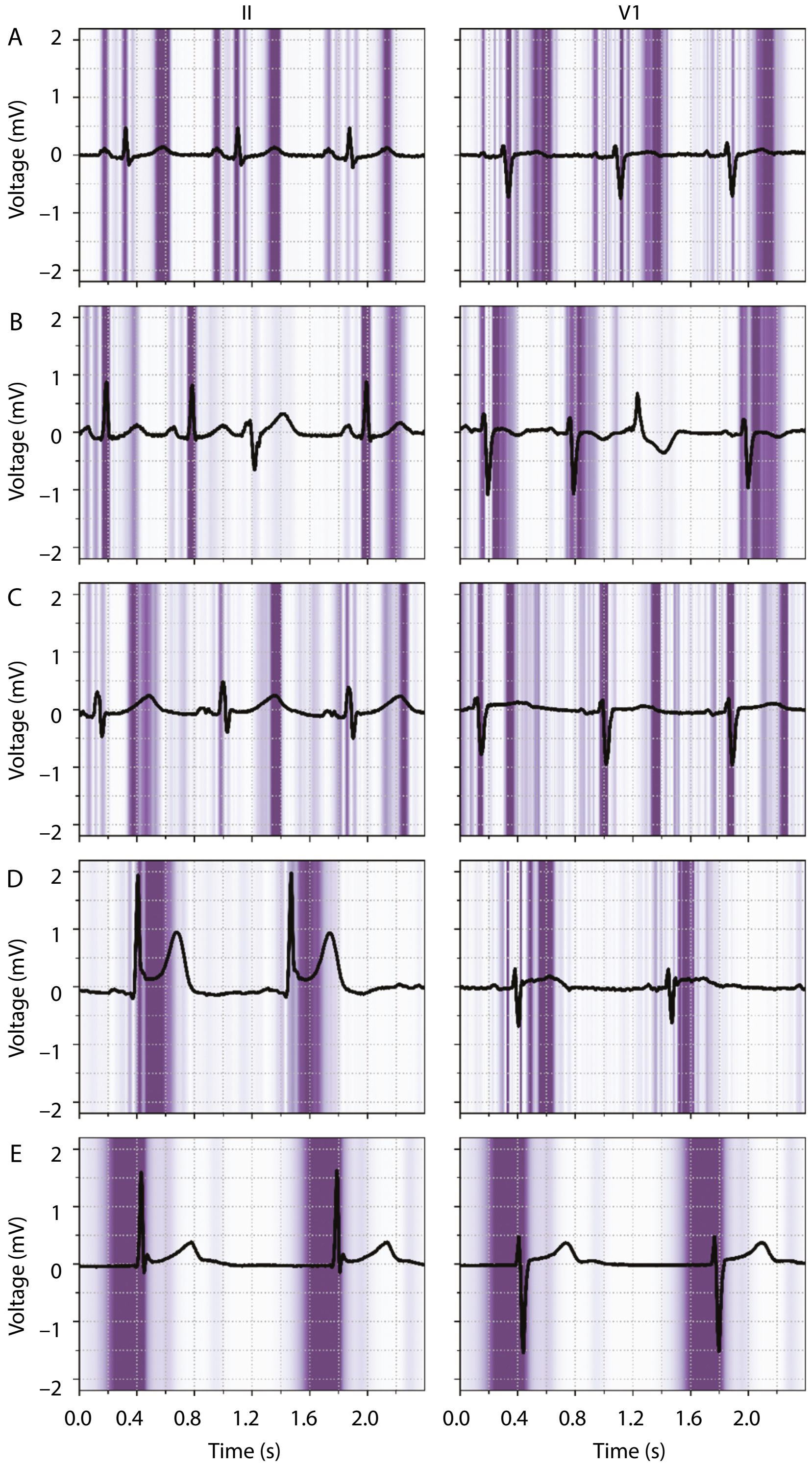


Figure 4.
ECG leads II and V1 with a superimposed guided Grad-CAM visualization showing regions important for the DNN to predict whether an ECG is normal, abnormal or acute.
A and B: Normal ECGs with focus on the P-wave, QRS- complex, and T- wave, while correctly ignoring a premature ventricular complex. C: Abnormal ECG with a long QT interval and a focus on the beginning and end of the QT- segment. D and E: Acute ECGs with an inferior ST-segment–elevation myocardial infarction (D) and a focus on the ST- segment with a junctional escape rhythm (E) and a focus on the pre-QRS- segment, where the P-wave is missing. Adopted from JAHA with permission [19].

retinopathy in fundus images using DNNs, where one study showed that overall accuracy could be improved when uncertain cases were referred to a physician[103] Another study suggested that uncertainty measures were able to detect when a different type of scanner was used that the algorithm had not seen before.[35] Moreover, combining uncertainty with active or online learning, allows the network to learn from previously uncertain cases, which are now reviewed by an expert.[104]

Ethical aspects

Several other ethical and legal challenges within the field of AI in healthcare are yet identified, like patient privacy, poor quality algorithms, algorithm transparency and liability concerns. Used data are subjected to privacy protections, confidentiality and data ownership, therefore requiring specific individual consent for use and reuse of data. However, with increasing the size of the dataset, anonymization techniques used nowadays might be inadequate and eventually result in the (de-)identification of patients.[105,106] Furthermore, as large datasets are required for DNNs, collaboration between institutions becomes inevitable. To facilitate data exchange, platforms have been established to allow for save and consistent data-sharing between institutions.[107] However, these databases may still contain sensitive personal data.[54,108] Therefore, federated learning architectures are proposed to provide data sharing while simultaneously obviating the need to share sensitive personal data (for example: andrea-consortium.org).

Another concerning privacy aspect is the continuous data acquisition through smartphone-based applications. In these commercial applications, data ownership and security are vulnerable aspects. Security between smartphones and applications is heterogeneous and data may be stored on commercial and poorly secured servers. Clear regulations and policies should be in place before these applications can enter the clinical arena.

Datasets contain information about medical history and treatment but may also encompass demographics, religious status or socioeconomic status. Apart from medical information, sensitive personal data might be taken into account by developed algorithms, possibly resulting in discrimination in for example ethnicity, gender or religion.[54,108–110]

As described, DNNs are 'black boxes' wherein input data is classified. Through the interpretation of DNNs and the incorporation of uncertainty measures, an estimate of the competency of an algorithm can be made. Traditionally, clinical practice mainly depends on the competency of a clinician. Decisions about diagnoses and treatments are based on widely-accepted clinical standards and the level of competency is protected by continuous intensive medical training. In the case of adverse events, clinicians are held responsible if they deviated from standard clinical care. However, the medical liability of the designed networks remains questionable. Incorrect computerized medical diagnoses or treatments result in adverse outcomes, thereby raising the question: who is accountable for mis diagnosis based on AI algorithms.

To guide the evaluation of ML algorithms, in particular DNNs, and accompanying literature in electrophysiology, a systematic overview of all relevant threats discussed in this review is presented in **Table 1**.

Table 1.* *Systematic overview of relevant threats of AI algorithms in electrophysiology. Adapted from [111]. Abbreviations: RCT = randomized controlled trial.

	KEY POINTS	**QUESTIONS**
INPUT	**Subjects**	Is an appropriate data source used with clean in- and exclusion criteria?
	Data	Is an appropriate data source used with clean in- and exclusion criteria?
PERFORMANCE	**Robustness**	How does the model perform? Was there a reasonable number of subjects? Were ECGs equally sampled per subject
	Overfitting and optimism	Was overfitting assessed using internal validation with train-test splitting, cross-validation or bootstrapping? Was the validation dataset of sufficient size (>100 participants with the outcome)?
	External validation	Are there external validation studies in different temporal, geographical or domain patient groups?
	Subgroups	Is subgroup analysis provided to minimize the risk of poor performance in subgroups? Is bias based on ethnicity, gender or other demographic factors present?
IMPLEMENTATION	**Subjects**	Is the population to use the algorithm similar to the (external) validation population? Is the disease prevalence similar?
	Data	Is the algorithm evaluated on the used diagnostic device of a specific manufacturer? Was data standardized according to general agreements?
	Implementation studies	Are there implementation studies (such as RCTs or before-after studies) performed? Does implementation of the model positively influence patient outcomes?
	Interpretation and uncertainty	Are there possibilities to check the predictions of the model in clinical practice (using visualizations)? Does the model provide uncertainty measures? How does the model deal with ECG noise or electrode misplacements? Is there a clear flowchart to refer specific uncertain cases to a physician?
	Etichal and legal	Are the ethical and legal aspects sufficiently addressed?

Conclusion

Many exciting opportunities arise when AI is applied to medical data, especially in cardiology and electrophysiology. Using AI technology, new ECG features, accurate automatic ECG diagnostics and new clinical insights can be rapidly obtained. In the near future, AI is likely to become one of the most valuable assets in clinical practice. However, as with every technique, AI has its limitations, also within the field of electrophysiology. To ensure the correct use of AI in a clinical setting, every clinician working with AI should be able to recognize the threats, limitations and challenges of the technique. Furthermore, during the creation of AI algorithms clinicians and data scientists should closely collaborate to ensure the creation of a clinically applicable and useful algorithm.

REFERENCES

1. Hemingway H, Asselbergs FW, Danesh J, et al. Big data from electronic health records for early and late translational cardiovascular research: challenges and potential. European heart journal 2018;39(16):1481–95. https://doi.org/10.1093/eurheartj/ehx487; PMID: 29370377.
2. Wasserlauf J, You C, Patel R, et al. Smartwatch Performance for the Detection and Quantification of Atrial Fibrillation. Circulation: Arrhythmia and Electrophysiology 2019;12(6):1–9. https://doi.org/10.1161/CIRCEP.118.006834; PMID: 31113234.
3. Bumgarner JM, Lambert CT, Hussein AA, et al. Smartwatch Algorithm for Automated Detection of Atrial Fibrillation. Journal of the American College of Cardiology 2018;71(21):2381–8. https://doi.org/10.1016/j.jacc.2018.03.003; PMID: 29535065.
4. Lopes RD, Alings M, Connolly SJ, et al. Rationale and design of the apixaban for the reduction of thrombo-embolism in patients with device-detected sub-clinical atrial fibrillation (ARTESiA) trial. American heart journal 2017;189:137–45. https://doi.org/10.1016/j.ahj.2017.04.008; PMID: 28625370.
5. Perez M v, Mahaffey KW, Hedlin H, et al. Large-scale assessment of a smartwatch to identify atrial fibrillation. New England Journal of Medicine 2019;381(20):1909–17. https://doi.org/10.1056/NEJMoa1901183; PMID: 31722151.
6. Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine 2019;25(January). https://doi.org/10.1038/s41591-018-0268-3; PMID: 30617320.
7. Kadish AH, Buxton AE, Kennedy HL, et al. ACC/AHA Clinical Competence Statement on Electrocardiography and Ambulatory Electrocardiography A Report of the ACC/AHA/ACP-ASIM Task Force on Clinical Competence (ACC/AHA Committee to Develop a Clinical Competence Statement on Electrocardiography and Amb. 2001; https://doi.org/10.1161/circ.104.25.3169; PMID: 11738321.
8. Salerno SM, Alguire PC, Waxman HS. Competency in interpretation of 12-lead electrocardiograms: a summary and appraisal of published evidence. Annals of Internal Medicine 2003;138(9):751–60. https://doi.org/138(9):751-60; PMID: 12729431.
9. Hill AC, Miyake CY, Grady S, et al. Accuracy of interpretation of preparticipation screening electrocardiograms. The Journal of pediatrics 2011;159(5):783–8. https://doi.org/10.1016/j.jpeds.2011.05.014; PMID: 21752393.
10. Dores H, Santos JF, Dinis P, et al. Variability in interpretation of the electrocardiogram in athletes: another limitation in pre-competitive screening. Revista Portuguesa de Cardiologia (English Edition) 2017;36(6):443–9. https://doi.org/10.1016/j.repc.2016.07.013; PMID: 28599797.
11. Schläpfer J, Wellens HJ. Computer-Interpreted Electrocardiograms Benefits and Limitations. J Am Coll Cardiol 2017;70(9):1183–92. https://doi.org/10.1016/j.jacc.2017.07.723; PMID: 28838369.
12. Viskin S, Rosovski U, Sands AJ, et al. Inaccurate electrocardiographic interpretation of long QT: the majority of physicians cannot recognize a long QT when they see one. Heart Rhythm 2005;2(6):569–74. https://doi.org/10.1016/j.hrthm.2005.02.011; PMID: 15922261.
13. Willems JL, Abreu-Lima C, Arnaud P, et al. The diagnostic performance of computer programs for the interpretation of electrocardiograms. New England Journal of Medicine 1991;325(25):1767–73. https://doi.org/10.1056/NEJM199112193252503; PMID: 1834940.
14. Guglin ME, Thatai D. Common errors in computer electrocardiogram interpretation. International journal of cardiology 2006;106(2):232–7. https://doi.org/10.1016/j.ijcard.2005.02.007; PMID: 16321696.
15. Shah AP, Rubin SA. Errors in the computerized electrocardiogram interpretation of cardiac rhythm. Journal of electrocardiology 2007;40(5):385–90. https://doi.org/10.1016/j.jelectrocard.2007.03.008; PMID: 17531257.
16. Bae MH, Lee JH, Yang DH, et al. Erroneous computer electrocardiogram interpretation of atrial fibrillation and its clinical consequences. Clinical cardiology 2012;35(6):348–53. https://doi.org/10.1002/clc.22000; PMID: 22644921.
17. Anh D, Krishnan S, Bogun F. Accuracy of electrocardiogram interpretation by cardiologists in the setting of incorrect computer analysis. Journal of electrocardiology 2006;39(3):343–5. https://doi.org/10.1016/j.jelectrocard.2006.02.002; PMID: 16777525.
18. Zhang K, Aleexenko V, Jeevaratnam K. Computational approaches for detection of cardiac rhythm abnormalities: Are we there yet? Journal of Electrocardiology 2020;59:28–34. https://doi.org/10.1016/j.jelectrocard.2019.12.009; PMID: 31954954.
19. Van de Leur RR, Blom LJ, Gavves E, et al. Automatic Triage of 12-Lead Electrocardiograms Using Deep Convolutional Neural Networks. Journal of the American Heart Association 2020;9(e015138). https://doi.org/10.1161/JAHA.119.015138; PMID: 32406296.
20. Perlman O, Katz A, Amit G, et al. Supraventricular tachycardia classification in the 12-lead ECG using atrial waves detection and a clinically based tree scheme. IEEE journal of biomedical and health informatics 2015;20(6):1513–20. https://doi.org/10.1109/JBHI.2015.2478076; PMID: 26415192.
21. Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nature Medicine 2019;25(1):70–4. https://doi.org/10.1038/s41591-018-0240-2; PMID: 30617318.
22. Sbrollini A, de Jongh MC, ter Haar CC, et al. Serial electrocardiography to detect newly emerging or aggravating cardiac pathology: a deep-learning approach. Biomedical engineering online 2019;18(1):15. https://doi.org/10.1186/s12938-019-0630-9; PMID: 30755195.
23. Wu JM-T, Tsai M-H, Xiao S-H, et al. A deep neural network electrocardiogram analysis framework for left ventricular hypertrophy prediction. Journal of Ambient Intelligence and Humanized Computing 2020;1–17. https://doi.org/10.1007/s12652-020-01826-1.
24. Hastie T, Tibshirani R, Friedman J: The elements of statistical learning: data mining, inference, and prediction. 2009, Springer Science & Business Media, New York City.
25. Van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC medical research methodology 2014;14(1):137. https://doi.org/10.1186/1471-2288-14-137; PMID: 25532820.
26. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Statistics in medicine 2016;35(2):214–26. https://doi.org/10.1002/sim.6787; PMID: 26553135.
27. Raghunath S, Cerna AEU, Jing L, et al. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. Nature Medicine 2020;1–6. https://doi.org/10.1038/s41591-020-0870-z; PMID: 32393799.
28. Kligfield P, Gettes LS, Bailey JJ, et al. Recommendations for the standardization and interpretation of the electrocardiogram: part I: the electrocardiogram and its technology a scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clin. Journal of the American College of Cardiology 2007;49(10):1109–27. https://doi.org/10.1016/j.jacc.2007.01.024; PMID: 17349896.

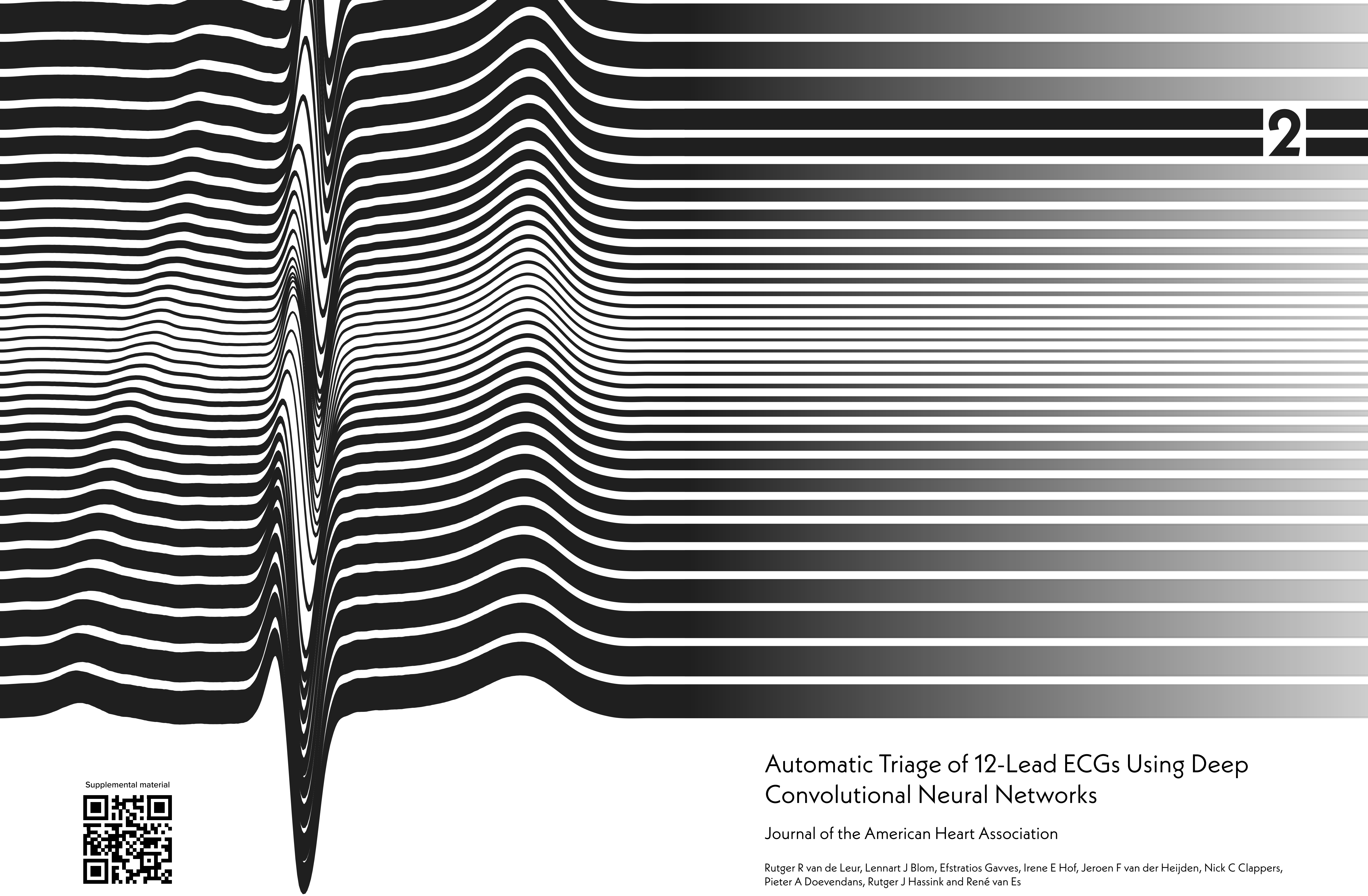
29. Jain R, Singh R, Yamini S, et al. Fragmented ECG as a Risk Marker in Cardiovascular Diseases. Current Cardiology Reviews 2014;10(3):277–86. https://doi.org/10.2174/1573403x10666140514103451; PMID: 24827794.
30. Korkmaz A, Yildiz A, Demir M, et al. The relationship between fragmented QRS and functional significance of coronary lesions. Journal of Electrocardiology 2017;50(3):282–6. https://doi.org/10.1016/j.jelectrocard.2017.01.005; PMID: 28117101.
31. Das MK, Zipes DP. Fragmented QRS: A predictor of mortality and sudden cardiac death. Heart Rhythm 2009;6(3 SUPPL.). https://doi.org/10.1016/j.hrthm.2008.10.019; PMID: 19251229.
32. Thakor N v, Webster JG, Tompkins WJ. Estimation of QRS Complex Power Spectra for Design of a QRS Filter. IEEE Transactions on Biomedical Engineering 1984;BME-31(11):702–6. https://doi.org/10.1109/tbme.1984.325393; PMID: 6500590.
33. Thakor N v, Webster JG, Tompkins WJ. Optimal QRS detector. Medical & Biological Engineering & Computing 1983;21(3):343–50. https://doi.org/10.1007/bf02478504; PMID: 6876910.
34. García-Niebla J, Serra-Autonell G, de Luna AB. Brugada syndrome electrocardiographic pattern as a result of improper application of a high pass filter. American Journal of Cardiology 2012;110(2):318–20. https://doi.org/10.1016/j.amjcard.2012.04.038; PMID: 22732021.
35. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature medicine 2018;24(9):1342–50. https://doi.org/10.1038/s41591-018-0107-6; PMID: 30104768.
36. Herman M v, Ingram DA, Levy JA, et al. Variability of electrocardiographic precordial lead placement: a method to improve accuracy and reliability. Clinical cardiology 1991;14(6):469–76. PMID: 1810683.
37. Hill NE, Goodman JS. Importance of accurate placement of precordial leads in the 12-lead electrocardiogram. Heart & lung: the journal of critical care 1987;16(5):561. https://doi.org/10.1007/s11517-013-1115-9; PMID: 3308780.
38. Chanarin N, Caplin J, Peacock A. "Pseudo reinfarction": a consequence of electrocardiogram lead transposition following myocardial infarction. Clinical cardiology 1990;13(9):668–9. https://doi.org/10.1002/clc.4960130916; PMID: 2208827.
39. Peberdy MA, Ornato JP. Recognition of electrocardiographic lead misplacements. The American Journal of Emergency Medicine 1993;11(4):403–5. https://doi.org/10.1016/0735-6757(93)90177-d; PMID: 8216526.
40. Rautaharju PM, Prineas RJ, Crow RS, et al. The effect of modified limb electrode positions on electrocardiographic wave amplitudes. Journal of Electrocardiology 1980;13(2):109–13. https://doi.org/10.1016/s0022-0736(80)80040-9; PMID: 7365351.
41. Rajaganeshan R, Ludlam CL, Francis DP, et al. Accuracy in ECG lead placement among technicians, nurses, general physicians and cardiologists. International Journal of Clinical Practice 2007;62(1):65–70. https://doi.org/10.1111/j.1742-1241.2007.01390..x; PMID: 17764456.
42. Van Oosterom A, Hoekema R, Uijen GJH. Geometrical factors affecting the interindividual variability of the ECG and the VCG. Journal of electrocardiology 2000;33:219–28. https://doi.org/10.1054/jelc.2000.20356; PMID: 11265725.
43. Hoekema R, Uijen GJH, van Erning L, et al. Interindividual variability of multilead electrocardiographic recordings: influence of heart position. Journal of electrocardiology 1999;32(2):137–48. https://doi.org/10.1016/S1053-0770(99)90050-2; PMID: 10338032.
44. Mincholé A, Zacur E, Ariga R, et al. MRI-based computational torso/biventricular multiscale models to investigate the impact of anatomical variability on the ECG QRS complex. Frontiers in physiology 2019;10:1103. https://doi.org/10.3389/fphys.2019.01103; PMID: 31507458.
45. Nguyên UC, Potse M, Regoli F, et al. An in-silico analysis of the effect of heart position and orientation on the ECG morphology and vectorcardiogram parameters in patients with heart failure and intraventricular conduction defects. Journal of Electrocardiology 2015;48(4):617–25. https://doi.org/10.1016/j.jelectrocard.2015.05.004; PMID: 26025201.
46. Hoekema R, Uijen G, van Oosterom A. The number of independent signals in body surface maps. Methods of information in medicine 1999;38(02):119–24. https://doi.org/10.1055/s-0038-1634176; PMID: 10431516.
47. Shimizu W, Matsuo K, Takagi M, et al. Body Surface Distribution and Response to Drugs of ST Segment Elevation in Brugada Syndrome: Clinical Implication of Eighty-Seven-Lead Body Surface Potential Mapping and Its Application to Twelve-Lead Electrocardiograms. Journal of Cardiovascular Electrophysiology 2000;11(4):396–404. https://doi.org/10.1111/j.1540-8167.2000.tb00334.x; PMID: 10809492.
48. Priori SG, Blomström-Lundqvist C, Mazzanti A, et al. 2015 ESC Guidelines for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death. European Heart Journal 2015;36(41):2793–867. https://doi.org/10.1093/eurheartj/ehv316; PMID: 26837728.
49. Ibánez B, James S, Agewall S, et al. 2017 ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation. Revista espanola de cardiologia (English ed) 2017;70(12):1082. https://doi.org/10.1016/j.rec.2017.11.010; PMID: 29198432.
50. Galloway CD, Valys A v., Shreibati JB, et al. Development and Validation of a Deep-Learning Model to Screen for Hyperkalemia From the Electrocardiogram. JAMA Cardiology 2019;55905:1–9. https://doi.org/10.1001/jamacardio.2019.0640; PMID: 30942845.
51. Rjoob K, Bond R, Finlay D, et al. Data driven feature selection and machine learning to detect misplaced V1 and V2 chest electrodes when recording the 12lead electrocardiogram. Journal of electrocardiology 2019;57:39–43. https://doi.org/10.1016/j.jelectrocard.2019.08.017; PMID: 31476727.
52. Hong S, Zhou Y, Shang J, et al. Opportunities and Challenges in Deep Learning Methods on Electrocardiogram Data: A Systematic Review. ArXiv. Available at: https://arxiv.org/abs/2001.01550. Published: December 2019. Accessed: May 12, 2020.
53. Helbing D. Societal, Economic, Ethical and Legal Challenges of the Digital Revolution: From Big Data to Deep Learning, Artificial Intelligence, and Manipulative Technologies. SSRN Electronic Journal 2015; https://doi.org/10.2139/ssrn.2594352.
54. Balthazar P, Harri P, Prater A, et al. Protecting Your Patients' Interests in the Era of Big Data, Artificial Intelligence, and Predictive Analytics. Journal of the American College of Radiology 2018;15(3):580–6. https://doi.org/10.1016/j.jacr.2017.11.035; PMID: 29402532.
55. Ribeiro AH, Ribeiro MH, Paixão GMM, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nature communications 2020;11(1):1–9. https://doi.org/10.1038/s41467-020-15432-4; PMID: 32273514.
56. Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. The Lancet 2019;6736(19):1–7. https://doi.org/10.1016/S0140-6736(19)31721-0; PMID: 31378392.
57. Ko W-Y, Siontis KC, Attia ZI, et al. Detection of Hypertrophic Cardiomyopathy Using a

Convolutional Neural Network-Enabled Electrocardiogram. Journal of the American College of Cardiology 2020;75(7):722–33. https://doi.org/10.1016/j.jacc.2019.12.030; PMID: 32081280.

58. Attia ZI, Sugrue A, Asirvatham SJ, et al. Noninvasive assessment of dofetilide plasma concentration using a deep learning (neural network) analysis of the surface electrocardiogram: A proof of concept study. PLoS ONE 2018;13(8):1–12. https://doi.org/10.1371/journal.pone.0201059; PMID: 30133452.
59. Kwon JM, Lee SY, Jeon KH, et al. Deep Learning-Based Algorithm for Detecting Aortic Stenosis Using Electrocardiography. Journal of the American Heart Association 2020;9(7):e014717. https://doi.org/10.1161/JAHA.119.014717; PMID: 32200712.
60. Katritsis G, Luther V, Kanagaratnam P, et al. Arrhythmia mechanisms revealed by ripple mapping. Arrhythmia & Electrophysiology Review 2018;7(4):261. https://doi.org/10.15420/aer.2018.44.3; PMID: 30588314.
61. Van den Broek HT, Wenker S, van de Leur R, et al. 3D Myocardial Scar Prediction Model Derived from Multimodality Analysis of Electromechanical Mapping and Magnetic Resonance Imaging. Journal of cardiovascular translational research 2019;12(6):517–27. https://doi.org/10.1007/s12265-019-09899-w; PMID: 31338795.
62. Van Es R, van den Broek HT, van der Naald M, et al. Validation of a novel stand-alone software tool for image guided cardiac catheter therapy. The international journal of cardiovascular imaging 2019;35(2):225–35. https://doi.org/10.1007/s10554-019-01541-9; PMID: 30689193.
63. Zollei L, Grimson E, Norbash A, et al. 2D-3D rigid registration of X-ray fluoroscopy and CT images using mutual information and sparsely sampled histogram estimators. In: IEEE Computer Society. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001 IEEE. Kauai, HI, USA: IEEE, 2001(2);II–II. https://doi.org/10.1109/CVPR.2001.991032.
64. Walsh KA, Galvin J, Keaney J, et al. First experience with zero-fluoroscopic ablation for supraventricular tachycardias using a novel impedance and magnetic-field-based mapping system. Clinical Research in Cardiology 2018;107(7):578–85. https://doi.org/10.1007/s00392-018-1220-8; PMID: 29476203.
65. Cantwell CD, Mohamied Y, Tzortzis KN, et al. Rethinking multiscale cardiac electrophysiology with machine learning and predictive modelling. Computers in biology and medicine 2019;104:339–51. https://doi.org/10.1016/j.compbiomed.2018.10.015; PMID: 30442428.
66. Bansal A, Joshi R. Portable out-of-hospital electrocardiography: A review of current technologies. Journal of Arrhythmia 2018;(December 2017):129–38. https://doi.org/10.1002/joa3.12035; PMID: 29657588.
67. Mairesse GH, Moran P, van Gelder IC, et al. Screening for atrial fibrillation: a European heart rhythm Association (EHRA) consensus document endorsed by the heart rhythm Society (HRS), Asia Pacific heart rhythm Society (APHRS), and Sociedad Latinoamericana de Estimulación Cardíaca Y Electrofisiología (SOLAECE). Ep Europace 2017;19(10):1589–623. https://doi.org/10.1093/europace/eux177; PMID: 29048522.
68. Freedman B, Camm J, Calkins H, et al. Screening for atrial fibrillation: a report of the AF-SCREEN international collaboration. Circulation 2017;135(19):1851–67. https://doi.org/10.1161/CIRCULATIONAHA.116.026693; PMID: 28483832.
69. Wegner FK, Kochhäuser S, Ellermann C, et al. Prospective blinded Evaluation of the smartphone-based AliveCor Kardia ECG monitor for Atrial Fibrillation detection: The PEAK-AF study. European Journal of Internal Medicine 2020;73:72–5. https://doi.org/10.1016/j.ejim.2019.11.018; PMID: 31806411.
70. Galloway C, Treiman D, Shreibati J, et al. 5105 A deep neural network predicts atrial fibrillation from normal ECGs recorded on a smartphone-enabled device. European Heart Journal 2019;40(Supplement_1):ehz746-0041. https://doi.org/10.1093/eurheartj/ehz746.0041.
71. Brasier N, Raichle CJ, Dörr M, et al. Detection of atrial fibrillation with a smartphone camera: first prospective, international, two-centre, clinical validation study (DETECT AF PRO). Ep Europace 2019;21(1):41–7. https://doi.org/10.1093/europace/euy176; PMID: 30085018.
72. McManus DD, Chong JW, Soni A, et al. PULSE- SMART: pulse- based arrhythmia discrimination using a novel smartphone application. Journal of cardiovascular electrophysiology 2016;27(1):51–7. https://doi.org/10.1111/jce.12842; PMID: 26391728.
73. Couderc J-P, Kyal S, Mestha LK, et al. Detection of atrial fibrillation using contactless facial video monitoring. Heart Rhythm 2015;12(1):195–201. https://doi.org/10.1016/j.hrthm.2014.08.035; PMID: 25179488.
74. Yan BP, Lai WHS, Chan CKY, et al. Contact- free screening of atrial fibrillation by a smartphone using facial pulsatile photoplethysmographic signals. Journal of the American Heart Association 2018;7(8):e008585. https://doi.org/10.1161/JAHA.118.008585; PMID: 29622592.
75. Caliebe A, Leverkus F, Antes G, et al. Does big data require a methodological change in medical research? BMC medical research methodology 2019;19(1):125. https://doi.org/10.1186/s12874-019-0774-0; PMID: 31208367.
76. Hashimoto DA, Rosman G, Rus D, et al. Artificial Intelligence in Surgery. Annals of Surgery 2018;268(1):70–6. https://doi.org/10.1097/sla.0000000000002693; PMID: 29389679.
77. Wang X, Peng Y, Lu L, et al. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: IEEE. Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, HI, USA, 2017;2097–106. https://doi.org/10.1109/CVPR.2017.369.
78. Lu MT, Ivanov A, Mayrhofer T, et al. Deep Learning to Assess Long-term Mortality From Chest Radiographs. JAMA Network Open 2019;2(7):e197416. https://doi.org/10.1001/jamanetworkopen.2019.7416; PMID: 31322692.
79. Baltruschat IM, Nickisch H, Grass M, et al. Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification. Scientific Reports 2019;9(1). https://doi.org/10.1038/s41598-019-42294-8; PMID: 31011155.
80. Oakden-Rayner L. Exploring the ChestXray14 dataset: problems. Wordpress: Luke Oakden Rayner 2017. https://doi.org/10.1016/j.acra.2019.10.006; PMID: 31706792.
81. Moeyersons J, Smets E, Morales J, et al. Artefact detection and quality assessment of ambulatory ECG signals. Computer methods and programs in biomedicine 2019;182:105050. https://doi.org/10.1016/j.cmpb.2019.105050; PMID: 31473442.
82. Clifford GD, Behar J, Li Q, et al. Signal quality indices and data fusion for determining clinical acceptability of electrocardiograms. Physiological measurement 2012;33(9):1419. https://doi.org/10.1088/0967-3334/33/9/1419; PMID: 22902749.
83. Xia H, Garcia GA, McBride JC, et al. Computer algorithms for evaluating the quality of ECGs in real time. In: IEEE. 2011 Computing in Cardiology. Hangzhou, China, 2011;369–72.
84. Li Q, Rajagopalan C, Clifford GD. A machine learning approach to multi-level ECG signal quality classification. Computer methods and programs in biomedicine 2014;117(3):435–47. https://doi.org/10.1016/j.cmpb.2014.09.002; PMID: 25306242.
85. Redmond SJ, Xie Y, Chang D, et al. Electrocardiogram signal quality measures for un-

supervised telehealth environments. Physiological Measurement 2012;33(9):1517. https://doi.org/10.1088/0967-3334/33/9/1517; PMID: 22903004.

86. Moons KGM, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart 2012;98(9):691–8. https://doi.org/10.1136/heartjnl-2011-301247; PMID: 22397946.
87. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology 2018;286(3):800–9. https://doi.org/10.1148/radiol.2017171920; PMID: 29309734.
88. Attia ZI, Kapa S, Yao X, et al. Prospective validation of a deep learning electrocardiogram algorithm for the detection of left ventricular systolic dysfunction. Journal of Cardiovascular Electrophysiology 2019;30(5):668–74. https://doi.org/10.1111/jce.13889; PMID: 30821035.
89. Noseworthy PA, Attia ZI, Brewer LC, et al. Assessing and Mitigating Bias in Medical Artificial Intelligence: The Effects of Race and Ethnicity on a Deep Learning Model for ECG Analysis. Circulation Arrhythmia and electrophysiology 2020;(March):208–14. https://doi.org/10.1161/CIRCEP.119.007988; PMID: 32064914.
90. Yao X, McCoy RG, Friedman PA, et al. ECG AI-Guided Screening for Low Ejection Fraction (EAGLE): Rationale and design of a pragmatic cluster randomized trial. American Heart Journal 2020;219:31–6. https://doi.org/10.1016/j.ahj.2019.10.007; PMID: 31710842.
91. Sendak MP, Gao M, Brajer N, et al. Presenting machine learning model information to clinical end users with model facts labels. NPJ Digital Medicine 2020;3(1):1–4. https://doi.org/10.1038/s41746-020-0253-3; PMID: 32219182.
92. Macfarlane PW, Katibi IA, Hamde ST, et al. Racial differences in the ECG—selected aspects. Journal of electrocardiology 2014;47(6):809–14. https://doi.org/10.1016/j.jelectrocard.2014.08.003; PMID: 25193321.
93. Rijnbeek PR, van Herpen G, Bots ML, et al. Normal values of the electrocardiogram for ages 16–90 years. Journal of electrocardiology 2014;47(6):914–21. https://doi.org/10.1016/j.jelectrocard.2014.07.022; PMID: 25194872.
94. Attia ZI, Friedman PA, Noseworthy PA, et al. Age and Sex Estimation Using Artificial Intelligence From Standard 12-Lead ECGs. Circulation: Arrhythmia and Electrophysiology 2019;12(9):1–11. https://doi.org/10.1161/CIRCEP.119.007284; PMID: 31450977.
95. Carvalho D v, Pereira EM, Cardoso JS. Machine Learning Interpretability: A Survey on Methods and Metrics. Electronics 2019;8(8):832. https://doi.org/10.3390/electronics8080832.
96. Selvaraju RR, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision 2017. Venice, Italy; 2017;618–26.
97. Springenberg JT, Dosovitskiy A, Brox T, et al. Striving for simplicity: The all convolutional net. San Diego, United States: International Conference on Learning Representations. 2015;1-14.
98. Strodthoff N, Strodthoff C. Detecting and interpreting myocardial infarction using fully convolutional neural networks. Physiological measurement 2019;40(1):015001. https://doi.org/10.1088/1361-6579/aaf34d; PMID: 30523982.
99. The LRM. Opening the black box of machine learning. The Lancet Respiratory medicine 2018;6(11):801. https://doi.org/10.1016/S2213-2600(18)30425-9; PMID: 30343029.
100. Sturmfels P, Lundberg S, Lee S-I. Visualizing the Impact of Feature Attribution Baselines. Distill 2020;5(1):e22. https://doi.org/10.23915/distill.00022.
101. Filos A, Farquhar S, Gomez AN, et al. A Systematic Comparison of Bayesian Deep Learning Robustness in Diabetic Retinopathy Tasks. ArXiv. Available at: https://arxiv.org/abs/1912.10481. Published: December 2019. Accessed: June 2, 2020.
102. Tagasovska N, Lopez-Paz D. Single-Model Uncertainties for Deep Learning. In: Advances in Neural Information Processing Systems. Vancouver, Canada, 2019;6414–25.
103. Leibig C, Allken V, Ayhan MS, et al. Leveraging uncertainty information from deep neural networks for disease detection. Scientific reports 2017;7(1):1–14. https://doi.org/10.1038/s41598-017-17876-z; PMID: 29259224.
104. Gal Y, Islam R, Ghahramani Z. Deep bayesian active learning with image data. ArXiv. Available at: https://arxiv.org/abs/1703.02910. Published: March 2017. Accessed: June 2, 2020.
105. Carter RE, Attia ZI, Lopez-Jimenez F, et al. Pragmatic considerations for fostering reproducible research in artificial intelligence. NPJ digital medicine 2019;2(1):1–3. https://doi.org/10.1038/s41746-019-0120-2; PMID: 31304388.
106. Rocher L, Hendrickx JM, de Montjoye Y-A. Estimating the success of re-identifications in incomplete datasets using generative models. Nature communications 2019;10(1):1–9. https://doi.org/10.1038/s41467-019-10933-3; PMID: 31337762.
107. Mandel JC, Kreda DA, Mandl KD, et al. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. Journal of the American Medical Informatics Association 2016;23(5):899–908. https://doi.org/10.1093/jamia/ocv189; PMID: 26911829.
108. Vayena E, Blasimme A. Health Research with Big Data: Time for Systemic Oversight. The Journal of Law, Medicine & Ethics 2018;46(1):119–29. https://doi.org/10.1177/1073110518766026; PMID: 30034208.
109. Vayena E, Blasimme A. Biomedical Big Data: New Models of Control Over Access, Use and Governance. Journal of Bioethical Inquiry 2017;14(4):501–13. https://doi.org/10.1007/s11673-017-9809-6; PMID: 28983835.
110. McCall B. What does the GDPR mean for the medical community? The Lancet 2018;391(10127):1249–50. https://doi.org/10.1016/s0140-6736(18)30739-6; PMID: 29619949.
111. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Annals of internal medicine 2019;170(1):51–8. https://doi.org/10.7326/M18-1376; PMID: 30596875.



2

Supplemental material

Automatic Triage of 12-Lead ECGs Using Deep Convolutional Neural Networks

Journal of the American Heart Association

Rutger R van de Leur, Lennart J Blom, Efstratios Gavves, Irene E Hof, Jeroen F van der Heijden, Nick C Clappers, Pieter A Doevendans, Rutger J Hassink and René van Es

Abstract

Background

The correct interpretation of the electrocardiogram (ECG) is pivotal for accurate diagnosis of many cardiac abnormalities and conventional computerized interpretation has not been able to reach physician level accuracy in detecting (acute) cardiac abnormalities yet. This study aims to develop and validate a deep neural network (DNN) for comprehensive automated ECG triage in daily practice.

Methods and results

We developed a 37-layer convolutional residual DNN on a dataset of free text physician-annotated 12-lead electrocardiograms. The DNN was trained on a dataset with 336.835 recordings from 142.040 patients and validated on an independent validation dataset (n = 984), annotated by a panel of five cardiologists-electrophysiologists. The 12-lead ECGs were acquired on all non-cardiology departments of the University Medical Center Utrecht. The algorithm learned to classify these ECGs into four triage categories: normal, abnormal not acute, subacute and acute. Discriminative performance is presented with overall and category-specific c-statistics, polytomous discrimination indexes (PDI), sensitivities, specificities, positive and negative predictive values. The patients in the validation dataset had a mean age of 60.4 years and 54.3% were male. The DNN showed excellent overall discrimination with an overall c-statistic of 0.93 (95% CI 0.92 – 0.95) and PDI of 0.83 (95% CI 0.79 – 0.87).

Conclusions

This study demonstrates that an end-to-end DNN can accurately be trained on unstructured free text physician annotations and used to consistently triage 12-lead electrocardiograms. When further finetuned with other clinical outcomes and externally validated in clinical practice, the demonstrated deep learning-based ECG interpretation can potentially improve time-to-treatment and decrease healthcare burden.

Introduction

With more than 300 million ECGs obtained annually worldwide, the electrocardiogram (ECG) is a fundamental tool in the everyday practice of clinical medicine.[1] The correct interpretation of the ECG is pivotal for accurate diagnosis of a wide spectrum of cardiac abnormalities and requires the expertise of an experienced cardiologist. The life-threatening nature of a suspected acute coronary syndrome and ventricular arrhythmias requires not only accurate, but also timely ECG interpretation and places a heavy logistic burden on clinical practice.

Automated triage of ECGs in categories that need acute, non-acute or no attention may therefore be of great support in daily practice. Accurately prioritizing different ECGs could lead to improvements in time-to-treatment and possibly decrease healthcare costs.[2] Especially in pre-hospital care and non-cardiology departments, expert knowledge to interpret ECGs might not always be readily available.[3–5] However, a consistent and fast automated algorithm that supports the physician in comprehensive triage of the ECG remains lacking.

Computerized interpretation of the ECG (CIE) was introduced over 50 years ago and became increasingly important in aiding the physician interpretation in many clinical settings. However, current CIE algorithms have not been able to reach physician level accuracy in diagnosing cardiac abnormalities.[5] Accurate interpretation of arrythmias and ST-segment abnormalities remains the most problematic and many algorithms suffer from high amounts of false positives for these disorders.[4–9] Overdiagnosis and failure to correct the erroneous interpretation by overreading physician has shown to lead to unnecessary interventions and medication use.[10,11]

With the development of algorithms that can benefit from large-scale processing of raw data without the need for hand-crafted feature extraction, a substantial improvement of CIE is forthcoming. Several of these techniques, deep neural networks (DNN) in particular, have shown to be highly effective in similar applications as speech recognition and image classification.[12–14] DNNs are computer algorithms based on the structure and function of the human brain. Their hidden layers of neurons can be trained to discover complex patterns in signals such as the ECG.[15] In comparison to conventional CIE algorithms, DNNs have the advantage that they jointly optimize both pattern discovery and classification in an end-to-end approach that only needs the raw waveforms as input. In medicine, deep learning showed promising results when applied to arrhythmia detection in single lead ECG recordings and to early detection of atrial fibrillation in normal sinus rhythm ECGs.[16,17] When combined with ultrasound or laboratory findings, deep learning algorithms were able to detect reduced ejection fraction and hyperkalemia in 12-lead ECGs.[18,19]

This study aims to develop and validate a DNN for comprehensive automated ECG triage that could support daily clinical practice.

Methods

Study participants

The dataset contained all 12-lead ECGs from patients between 18 and 85 years old recorded in the University Medical Center Utrecht from January 2000 to August 2019, obtained at non-cardiology departments. All extracted data were de-identified in accordance with the EU General Data Protection Regulation and written informed consent was therefore not required by the UMC Utrecht ethical committee.

Training data acquisition and annotation

All ECGs were recorded on a General Electric MAC 5500 (GE Healthcare, Chicago, IL, United States). We extracted raw 10 second 12-lead ECG data waveforms from the MUSE ECG system (MUSE version 8, GE Healthcare, Chicago, IL, United States). All recordings in the UMC Utrecht acquired in non-cardiology departments were systematically annotated by a physician as part of the regular clinical workflow. These physicians were all trained to interpret and annotate an ECG as part of their cardiology residency. During the annotation, the physicians had access to the name, sex and age of the patient, the computer-calculated conduction intervals, the previous ECG recordings and the full patients records. The ECGs were divided into 4 triage categories, based on how quickly a cardiologist has to be consulted: [1] normal, [2] not acute abnormal (consultation with low priority), [3] subacute abnormal (consultation with moderate priority) and [4] acute abnormal (consultation with high priority).

The free-text physician ECG annotations were labelled into one of the four triage categories using a text mining-based approach. Firstly, the annotations were tokenized and all frequent (i.e. occurring more than 20 times) terms and multi-word collocations were extracted. These terms, such as “STEMI”, and collocations, such as “first degree AV-block” and “1st degree AV block”, contained multiple variations of diagnostic ECG statements. Therefore, they were mapped to the standardized statements of the “American Heart Association’s Electrocardiography Diagnostic Statement List”.[20] Secondly, a panel of three electrophysiologists defined the triage category for every standardized diagnostic statement. The used diagnostic statements and their corresponding triage category are appreciated in **Figure 1**. Thirdly, a final triage category was assigned to every ECG. When multiple statements were given, the final triage category was the maximum category. All text mining step were performed with the quanteda package for R (version 3.5, R Foundation for Statistical Computing, Vienna, Austria). [21] An overview of the text mining steps can be found in **Figure 2**.

Figure 1.
ECG diagnoses with their corresponding triage category. Triage categories as defined by the panel of electrophysiologists, with *[1]* normal, *[2]* not acute abnormal (consultation without priority), *[3]* subacute abnormal (consultation with some priority) and *[4]* acute abnormal (consult immediately). The ECG diagnoses derived from the text mining algorithm were used to categorize the training data using these rules. When multiple diagnoses were given, the final triage category was the maximum category. AV: atrioventricular. AVNRT: atrioventricular nodal reentrant tachycardia, AVRT: atrioventricular reentrant tachycardia.

	Normal	Abnormal, not acute	Abnormal, subacute	Abnormal, acute
RHYTHM	Sinus rhythm Sinus arrhythmia Atrial ectopic beat Ventricular ectopic beat	Sinus tachycardia *> (220 - age)/min* Sinus bradycardia *< 50/min* Ectopic atrial rhythm Atrial fibrillation *< 100/min* Paced rhythm	Atrial fibrillation *> 100/min* Atrial flutter	Extreme bradycardia *< 30/min* AVNRT/AVRT Ventricular tachycardia Junctional escape Ventricular escape Undefined rhythm
CONDUCTION	First degree AV block	Intraventricular conduction delay Right bundle branch block Left bundle branch block Other blocks	Second degree AV block QT interval *> 500ms*	Third degree AV block
ISCHEMIA		History of transmural infarction pathological Q-waves		ST-elevation and/or -depression T-wave inversion
OTHER		Nonspecific ST/T abnormalities Left ventricular hypertrophy	Pericarditis	

Consultation of cardiologist	
Normal	No
Abnormal, not acute	< 24 hours
Abnormal, subacute	< 3 hours
Abnormal, acute	Immediately

Validation data annotation

For the validation of the DNN, a dataset with higher annotation reliability was required. Therefore, an independent dataset was annotated and triaged by the reference standard, a panel of five practicing senior electrophysiologists or cardiologists. All records were annotated by two independent annotators, who were blinded to the other annotation. In case of disagreement in the triage category, a third annotator was consulted, and the majority vote was used as the final label. The recordings with three discordant votes were discussed in a joint panel meeting and recordings of insufficient quality were excluded. Annotation was performed using an online tool, where the expert had access to the 12-lead ECG, computer-calculated conduction intervals and the age and sex of the patient. The experts were instructed to classify the ECGs into one of the four triage categories based on the rules in **Figure 1**. The input and annotation steps in the validation dataset are schematically shown in **Figure 2**.

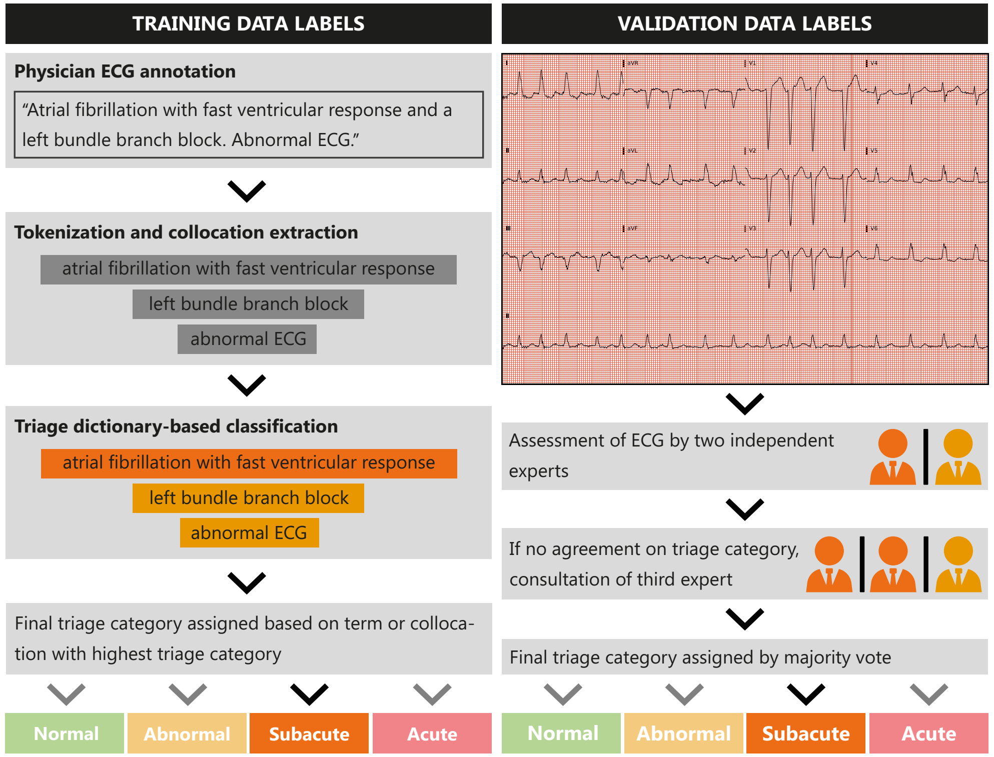

TRAINING DATA LABELS
Physician ECG annotation
"Atrial fibrillation with fast ventricular response and a left bundle branch block. Abnormal ECG."
Tokenization and collocation extraction
atrial fibrillation with fast ventricular response
left bundle branch block
abnormal ECG
Triage dictionary-based classification
atrial fibrillation with fast ventricular response
left bundle branch block
abnormal ECG
Final triage category assigned based on term or collocation with highest triage category
Normal
Abnormal
Subacute
Acute
VALIDATION DATA LABELS
Assessment of ECG by two independent experts
If no agreement on triage category, consultation of third expert
Final triage category assigned by majority vote
Normal
Abnormal
Subacute
Acute

Figure 2. ***Overview of the labelling into triage categories in the training and validation datasets.*** The training labels (left), used for training, are derived from the free-text annotation given to the ECG by a single physician in daily practice. The ECG diagnoses are mapped to triage categories using rules defined by a panel of electrophysiologists (***Figure 1***). The validation labels (right), used for validation of the DNN, are given by the expert panel based on visual inspection of a 12-lead ECG.

As manual annotation by a panel is time-intensive, a sample size calculation was performed to achieve adequate precision of the validation performance measures. For this, a minimum of 50 cases per category was needed.[22] As the smallest triage category in the training dataset has a prevalence of approximately 5%, the validation dataset consisted of 1000 recordings from unique patients. All ECGs of these patients were excluded from the training dataset.

Algorithm development

Considering that lead III, aVR, aVL and aVF are derived from lead I and II, we used the raw 10 second 8-channel waveforms (I, II and V1-V6), sampled at 500Hz, as the input for our DNN. We applied an architecture similar to the Inception ResNet network, by combining blocks of convolutional layers in parallel with residual connections.[23,24] This network is built with layers of identical blocks with a pre-activation design, consisting of two 1-dimensional convolutional layers, preceded by batch normalization, rectified linear unit (ReLU) activation and dropout.[25–27] Every residual block consists of three parallel branches; one with a normal convolutional layer, one with a dilated convolutional layer and one with a shortcut connection, where the input to the block is added to the output unadjusted.[28] This enables the network to determine features in two different time-dimensions, where the dilated convolution covers a complete heartbeat. The output of the last block was flattened and used as input to a fully connected layer with a ReLU nonlinearity, followed by dropout. The output layer consisted of four nodes, one for every triage category, and a softmax function was used to produce a probability distribution over all triage categories. A similar auxiliary output was added in the middle of the network and its loss was added to the total loss during training.

After hyperparameter and architecture optimization, the final selected network consisted of 16 residual blocks with two 1D convolutional layers with filter size 5 and dilation of 100 (**Supplemental Figure 1**). Every other block downsampled the input using a strided convolution and the number of filters was doubled every fourth block. Dropout was performed with a probability of 30%. The fully connected layer consisted of 256 nodes. This resulted in a final network with 37 layers.

This network was trained using the Adam optimizer with a learning rate of 0.0005 and a mini batch size of 128.[29] Weighted focal loss was used to counteract the category imbalance in the dataset and to minimize the number of false negatives.[30] Training was terminated when the loss stopped decreasing in the 5% subset of the training dataset. Network training was performed using the PyTorch package (version 1.3) on a Titan Xp GPU (NVIDIA Corporation, Santa Clara, CA, United States).[31]

The different network architectures and hyperparameters were chosen using a combination of manual tuning and random grid search. The network with the lowest loss in a 5% randomly sampled subset of the training dataset was chosen. When multiple architectures showed similar performance, the simplest architecture was selected. The following hyperparameters were assessed: the use of dilated convolutions, residual connections, max pooling, an auxiliary loss and/or fully connected layers, the number of layers, the size and number of convolutional filters, the dropout rate, the learning rate and the weights of the loss. We also experimented with an ordinal loss method instead of a multinomial loss method and with adding age and sex to the flattened layer, but this did not result in increased performance.[32,33]

Visualization of the DNN

To improve understanding of the decisions of the DNN, Guided Grad-CAM, a technique for visual explanations in convolutional neural networks, was adjusted for use in 1-dimensional data.[34] Guided Grad-CAM is a combination between the fine-grained Guided Backpropagation and Grad-CAM, which produces a coarse class-discriminative heatmap based on the final convolutional layer.[34,35] The heatmap is superimposed over the ECG recording and shows the regions in the ECG important to the DNN for predicting a specific triage category.

Statistical analysis

Inter-observer agreement was quantified using squared weighted Cohen's kappa for two reviewers or tests and ordinal Krippendorff's alpha for more than two reviewers.[36,37] Considering the imbalance in category frequencies, overall algorithm discriminatory performance was assessed with the unweighted mean of all pairwise concordance (or c) statistics (also known as area under the receiver operating curve (AUROC)) and the polytomous discriminatory index (PDI).[38–40] The first metric estimates the probability to correctly distinguish between all pairs of two patients from different categories, where a value of 0.5 denotes random performance and 1 perfect performance. The second assesses the discrimination between all categories simultaneously in a set approach. It estimates the probability to correctly identify a specific patient in a set of patients from every category, where 0.25 denotes random and 1 perfect performance with four categories.[38,39,41] As a second step, category-specific performance is assessed with the c-statistic, PDI, sensitivity, specificity and positive and negative predictive values. All category-specific measures, except the PDI, were applied in a one-versus-other approach.

All statistical analyses were performed using R version 3.5 (R Foundation for Statistical Computing, Vienna, Austria). The TRIPOD Statement for reporting of diagnostic models was followed, where appropriate.[42] All data are presented as mean ± standard deviation or median with interquartile range (IQR). The 95% confidence intervals around the performance measures were obtained using 2000 bootstrap samples.

Results

The total training dataset consisted of 336.835 recordings of 142.040 patients. The distribution of triage categories was unbalanced with the most recordings in category 2 (45.5%) and the least in category 4 (4.8%). In the validation dataset, there was consensus between the two experts in 736 cases (73.6%). After consultation of a third tie-breaker expert (248 cases, 24.8%), the panel meeting (29 cases, 2.9%) and exclusion of recordings of insufficient quality, 984 validation cases were used for the analysis. There was good inter-observer agreement, with a Krippendorff's alpha of 0.72. Conflicts in the first expert annotation round occurred the most between category 1 and 2 (162/255, 64%), between category 2 and 3 (30/255, 12%) and between category 2 and 4 (24/255, 9.5%). Disagreement between category 1 and 2 was mostly due to different assessments on the presence of nonspecific ST-segment or T-wave abnormalities. For category 2 and 3 and category 3 and 4, the most common difference was the interpretation of ST-segment elevation or depression. **Table 1** summarizes the patient demographics and triage category distributions of the recordings in the training and validation datasets.

The overall discrimination, as measured by the unweighted mean of pairwise c-statistics and the PDI, of the DNN demonstrated in this paper were 0.93 (95% CI 0.92 – 0.95) and 0.83 (95% CI 0.79 – 0.87), respectively. C-statistics, PDIs, sensitivities, specificities, positive predictive values (PPV) and negative predictive values (NPV) per triage category in a one-versus-other approach are shown in **Table 2**, while the confusion matrix is appreciated in **Figure 3**. Visualizations of the regions in the ECG important for the DNN to predict a specific category are shown in **Figure 4.** The full 12-lead ECGs can be found in **Supplemental Figures 2-6**.

The DNN predicted a lower triage category than the true category (undertriage) in 88 (8.9%) and a higher category (overtriage) in 107 (11%) of the recordings in the validation dataset. Most undertriage (59/88, 67%)

Table 1.* *Patient demographics and distribution of triage categories in the training and validation datasets. A 5% randomly sampled subset of the training dataset was used for model tuning and internal validation. The validation dataset is independent from the training dataset. [†] Distribution based on text mining categorization of annotations by physician in daily practice. [‡] Distribution based on the expert consensus panel annotations.

		TRAINING (n = 336.835)	**VALIDATION** (n = 984)
MALE SEX (n (%))		188858 (56.1)	402 (54.3)
AGE (mean (sd))		60.8 (15.5)	60.4 (15.3)
LOCATION (n (%))	**Emergency Department**	92532 (27.5)	310 (31.5)
	Intensive Care Unit	20045 (6.0)	63 (6.4)
	Non-cardiology outpatient clinic	73170 (21.7)	161 (16.4)
	Non-cardiology ward	86630 (25.7)	263 (26.7)
	Pre-operative screening	6300 (1.9)	8 (0.8)
	Recovery Ward	53994 (16.0)	163 (16.6)
	Other	4164 (1.2)	80 (8.1)[‡]
TRIAGE CATEGORY (n (%))	**Normal**	142456 (42.3) [†]	418 (42.5) [‡]
	Abnormal, not acute	153360 (45.5) [†]	410 (41.7) [‡]
	Abnormal, subacute	24731 (7.3) [†]	76 (7.7) [‡]
	Abnormal, acute	16288 (4.8) [†]	80 (8.1) [‡]

Table 2.
Diagnostic performance measures per triage category for the deep neural network in the panel annotated validation dataset.
The c-statistics, sensitivities, specificities, positive and negative predictive values are calculated in a one-vs-other approach and compare the category with the highest probability to the reference standard. The PDI estimates the probability that a patient from that category is correctly identified from a set of cases from every category.
CI: confidence interval, C-statistic: concordance statistic, equivalent to area under the receiver operating characteristic curve, PDI: polytomous discriminatory index.

	NORMAL	ABNORMAL, NOT ACUTE	ABNORMAL, SUBACUTE	ABNORMAL, ACUTE
C-STATISTIC (95% CI)	0.95 (0.94 – 0.96)	0.91 (0.89 – 0.93)	0.94 (0.90 – 0.97)	0.94 (0.90 – 0.97)
PDI (95% CI)	0.91 (0.87 – 0.93)	0.80 (0.75 – 0.84)	0.82 (0.75 – 0.88)	0.80 (0.73 – 0.87)
SENSITIVITY	0.87	0.76	0.64	0.79
SPECIFICITY	0.88	0.89	0.98	0.94
POSITIVE PREDICTIVE VALUE	0.85	0.83	0.78	0.55
NEGATIVE PREDICTIVE VALUE	0.90	0.84	0.97	0.98

occurred between categories 1 and 2 and these undertriaged recordings were categorized as 2 by the panel based on nonspecific ST-segment abnormalities (26/59, 44%), old ischemia (12/59, 20%), left ventricular hypertrophy (7/59, 12%) or other reasons (14/59, 24%). All 9 acute category 4 recordings triaged as category 2 contained ST-depression or T-wave inversion and no ST-elevation. In the category 2 recordings overtriaged as 4, the panel did mention nonspecific ST-segment abnormalities in 20/34 (59%) recordings and old ischemia in 8/34 (24%).

As the labelling procedures for the training and validation datasets differ (**Figure 2**), the performance of the DNN could be dependent on errors in two steps in the training labelling procedure. Firstly, the inter-observer agreement between manual categorization of the free-text physician ECG annotations into triage categories and the text mining-based categorization was excellent in the validation dataset, with a weighted Cohen's kappa of 0.96. Secondly, agreement between the text mining-based triage categories and the reference standard was good (Cohen's kappa 0.74). The overall c-statistic and PDI for predicting the reference standard triage category with the text mining-based categories were 0.86 (95% CI 0.85 – 0.88) and 0.48 (95% CI 0.43 – 0.53), respectively.

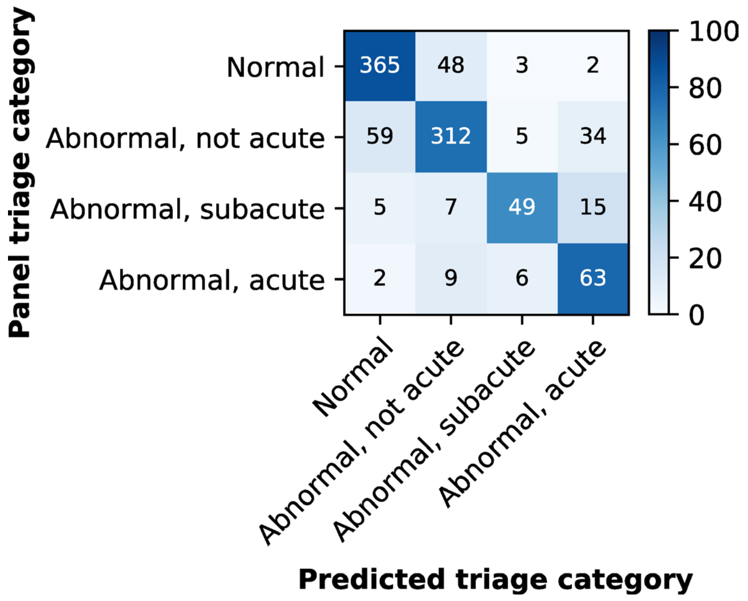


Figure 3.
Confusion matrix for the deep neural network.
Rows represent the categories given by the reference standard (expert panel) and columns the categories predicted by the deep neural network (DNN). The colormap is normalized per row and represents the percentage in the true triage category.

Discussion

This study is among the first to apply DNNs to a large dataset of 12-lead ECGs for automatic interpretation. We demonstrated that a deep learning approach performs well in detecting abnormalities for triage of 12-lead ECGs. Our DNN has an excellent c-statistic of 0.93 (95% CI 0.92 – 0.95) and good PDI of 0.83 (95% CI 0.79 – 0.87), with high positive- and negative predictive values across all triage categories. These findings indicate that a deep learning approach may be used to support the physician in ECG triage and reduce clinical workload with an improved prioritization of ECGs for interpretation by the cardiologist.

Interpretation of the ECG requires extensive knowledge of the wide variety of electrical manifestations of heart disease and a good understanding of normal variety. This has been a challenge for both manual and computerized interpretation, and has led to a collection of definitions, measurements and criteria to aid clinical decision making.[5,20] This challenge is extensively described in earlier studies but comparisons are difficult, as the studies demonstrate wide variations in diagnostic measures and an international accepted standard for validation of ECG diagnosis is still missing.[4,5] For comprehensive ECG interpretation non-cardiologist physicians correctly identified 36% to 96% of the diagnoses, with significant differences between physicians and increasing performance for more experienced physicians.[4,43–45] Most studies focused on particular aspects of ECG interpretation, such as normal-abnormal differentiation, arrythmia classification and detection of ST-elevation myocardial infarction (STEMI). Overall, for these aspects physicians have higher false negative rates, while computerized algorithms have higher false positive rates, when compared to expert panels.[4,5,7–9,43–45] The DNN could improve both the high false positive and negative rates, while producing consistent results not dependent on external factors, such as physician experience.

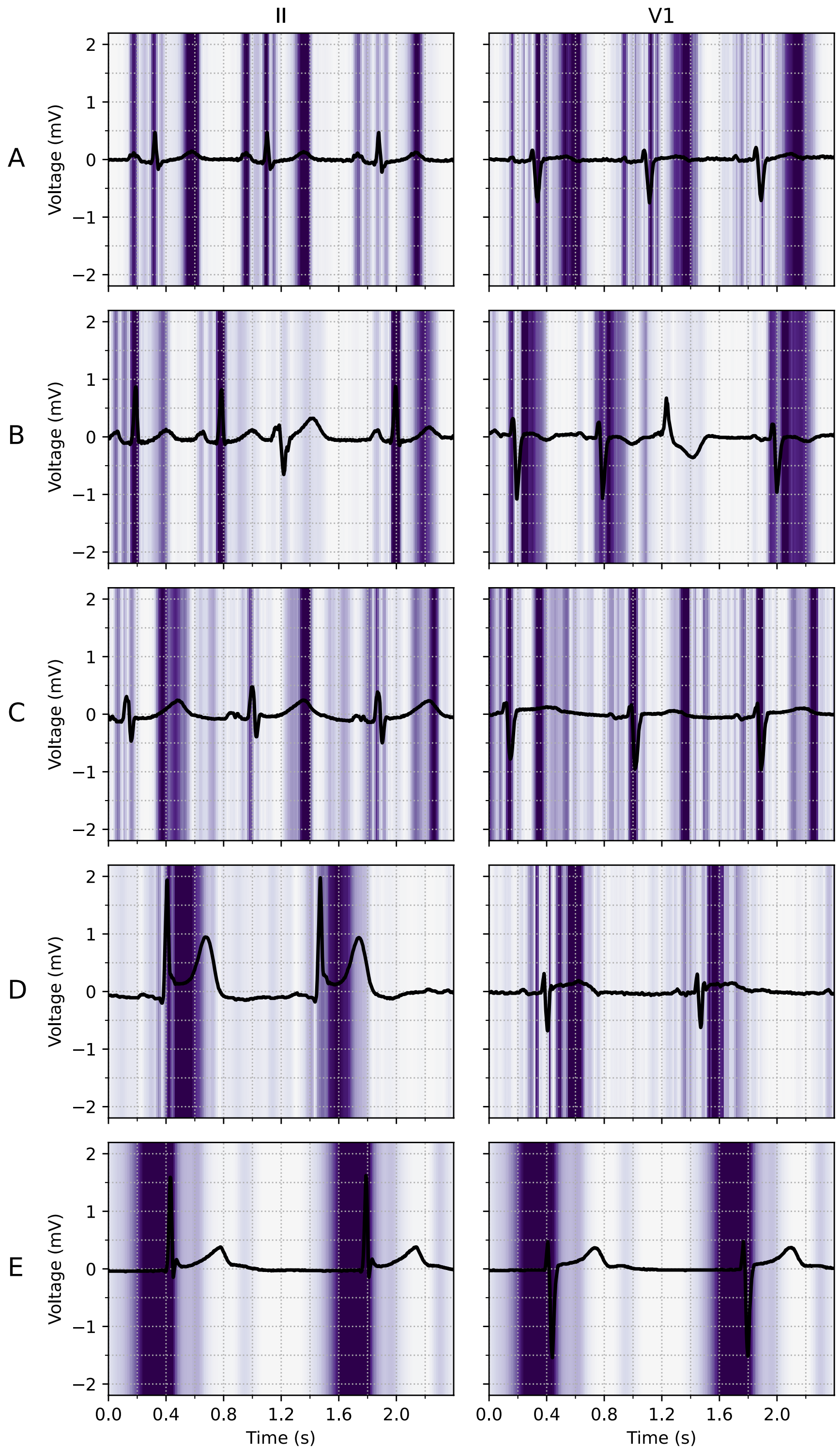


Figure 4. *Examples of electrocardiogram (ECG) leads II and V1 with superimposed a Guided Grad-CAM visualization showing regions important for the deep neural network (DNN) to predict a certain triage category.* A Normal ECG with focus on the P-wave, QRS-complex and T-wave. B Normal ECG with a single ignored premature ventricular complex (PVC). C Subacute ECG with a long QT interval and a focus on the beginning and end of the QT-segment. D Acute ECG with an inferior ST-elevation myocardial infarction and a focus on the ST-segment and J-point. E Acute ECG with a junctional escape rhythm and a focus on the pre-QRS-segment, where the P-wave is missing. The full 12-lead ECGs are available in Supplementary Figures 2-6.

Conventional CIE uses manually derived features, which only capture a fraction of the available information for any manifestation of heart disease in the obtained raw signal. This is one of the reasons that could explain the excellent performance of our algorithm and DNNs in general, as their integrated feature discovery and classification incorporates the whole raw input signal. In addition, conventional CIE algorithms are tuned to produce complete interpretations of the ECG and are less focused on one of their most important uses, quick triage. By training on a large physician-annotated 12-lead ECG dataset, where the labels are mapped to predefined triage categories, we focus on a single task and are able to achieve high accuracy. The large size of the dataset makes that the network has seen a wide variety of ECGs and should therefore be well generalizable.

Although the DNN does not use any manually selected features of the signal, visualizations show that the network bases its decisions on same regions in the ECG as experts would. As shown in **Figure 4** the network correctly identifies a normal ECG, a long QT-segment, ST-elevation myocardial infarction and a junctional escape rhythm in sensible regions and correctly ignores a premature ventricular complex in a normal ECG. Furthermore, inspection of the misclassifications of the DNN shows a similar pattern to the disagreement between the experts in the panel. The correct interpretation of ST-segment and T-wave abnormalities is apparently challenging for both the cardiologist and the DNN.

The DNN is trained on triage category labels that were automatically derived using text mining on free text annotations by a single physician in daily practice. Disagreement measures show that the text mining categories do not completely agree with the labels given by the expert panel. Most of this disagreement is caused by disagreement between the expert panel labels and the automatically categorized single physician labels (Cohen's kappa 0.74). Considering this substantial disagreement between training and validation labels, we might expect that the DNN cannot outperform the performance measures for prediction with only the text mining-based triage categories. However, the DNN exceeds both the overall c-statistic and PDI of the text mining-based triage categories and shows to be robust against considerable training label noise. This is in line with previous research that showed that deep neural network can handle label noise quite well.[46]

Other research demonstrated the value of DNN for ECG interpretation for similar problems, where a single-lead ECG was used for arrhythmia classification and a 12-lead ECG for early detection of atrial fibrillation, contractile dysfunction and hyperkalemia.[16–19] Our study shows that, in comparison to combining ECG recordings with other imaging modalities or laboratory findings, it is also feasible to use the less structured and noisy physician labels to successfully train a DNN for comprehensive ECG triaging. Moreover, this is one of the first studies to visualize regions in the ECG important for the decisions of the DNN.[34]

Triage is the process of classifying according to the severity of the case to determine how quickly action is needed. Careful triage is needed to prioritize those cases where timely action reduces morbidity and mortality among patients. For a triage algorithm to be effective, it is important that undertriage, e.g. failure to detect patients with acute disease, and overtriage, e.g. false alarms, are minimized. Our DNN shows very high negative predictive values for the highest categories, subacute and acute (**Table 2**). This can potentially reduce time-to-treatment for patients with acute cardiac disorders, as the algorithm is able to provide triage advice immediately after the ECG is acquired and before the ECG is assessed by a physician with sufficient expertise. However, the sensitivities for the subacute and acute categories of 64% and 79% are partly due to undertriage (**Figure 3**) and therefore need further improvement before clinical implementation is possible. The algorithm shows relatively high positive predictive values, which will decrease the amount of false alarms in otherwise normal ECGs. Since most hospital-acquired ECGs fall into this category, a modest improvement can already significantly decrease the workload for physicians.

This study has several limitations to address. Though a reasonable large training dataset was used, the acute categories remained relatively small. This is customary to an unselected real-world dataset but entails a chance of underprediction. We made use of the focal loss method, used in computer-vision DNN algorithms, to counteract this problem.[26] In the validation dataset the triage category distribution was similar, but confidence intervals showed adequate precision in the smaller categories too. The representative sampling of the validation is also a strength, making it possible to derive positive and negative predictive values, which are most important to the patient. We believe that the panel-annotated validation dataset in this study provides a good measure of generalizability of the physician annotation-based deep learning to hospital populations comparable to ours. It has been shown that ethnicity influences the ECG and could be taken into account to improve automated interpretation.[47] External validation is therefore needed when used with different recording machines and in very different populations, such as patients in the general practice or populations with a different ethnical composition. This is beyond the scope of this study and will most likely require (re)training on a such a dataset.

Both manual and computerized ECG interpretation are hard to standardize, as can be seen by high disagreement rates between the experts (Krippendorff's alpha 0.72). This number is comparable to earlier studies on the inter-observer agreement between experts on ECG interpretation.[4,6,48] The panel-annotated validation dataset used in this study is the current best reference standard available, but in clinical practice, many other diagnostic tests are used to interpret the ECG findings. Therefore, we suspect the diagnostic accuracy of our algorithm could be further optimized with hard clinical outcome data, such as a diagnosis and localization of myocardial infarction with coronary artery angiography, cardiac enzymes and electrolyte disorders from laboratory data and even mortality. Both optimization with clinical outcome data and external validation is necessary before clinical implementation is possible.

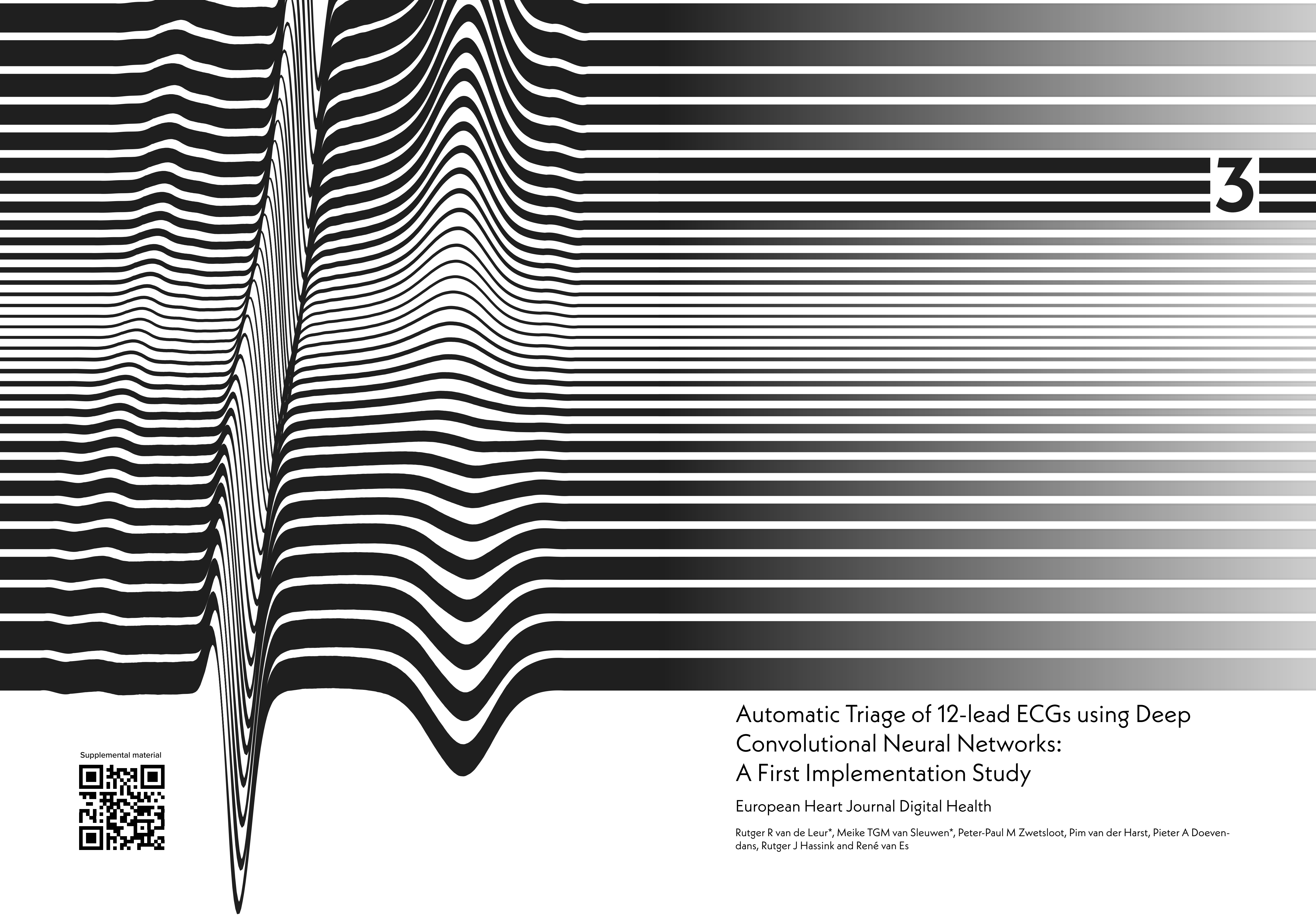
Another future perspective of the DNN is their capability to continuously improve and learn by adding new cases. Traditionally, neural networks did not provide uncertainly around their predictions, but new insights with several different Bayesian methods changed this.[49] When combining uncertainty around the predictions with active learning, it becomes possible to let uncertain cases be annotated by a cardiologist and improve the algorithm, while easier cases can be classified automatically.[50] Moreover, to determine the most important ECG leads, the algorithm could be trained and evaluated with fewer input channels. This could make the use of a similar algorithm with home-monitoring devices with less leads possible.

In conclusion, our end-to-end DNN can triage 12-lead electrocardiograms into normal, abnormal and acute with high discrimination across all categories. In clinical practice, this could lead to improved time-to-treatment for acute cardiac disorders and decreased and better-balanced workload for clinicians. Further improvement with other clinical outcomes, prospective validation in other populations and implementation studies are needed before implementation in clinical practice is possible.

REFERENCES

1. Holst H, Ohlsson M, Peterson C, Edenbrandt L. A confident decision support system for interpreting electrocardiograms. Clin Physiol. 1999;19:410–418.
2. Diercks DB, Kontos MC, Chen AY, et al. Utilization and Impact of Pre-Hospital Electrocardiograms for Patients With Acute ST-Segment Elevation Myocardial Infarction. J Am Coll Cardiol. 2009;53:161–166.
3. Eslava D, Dhillon S, Berger J, et al. Interpretation of electrocardiograms by first-year residents: the need for change. J Electrocardiol. 2009;42:693–697.
4. Salerno SM, Alguire PC, Waxman HS. Competency in Interpretation of 12-Lead Electrocardiograms: A Summary and Appraisal of Published Evidence. Ann Intern Med. 2003;138:751.
5. Schläpfer J, Wellens HJ. Computer-Interpreted Electrocardiograms. J Am Coll Cardiol. 2017;70:1183–1192.
6. Holmvang L, Hasbak P, Clemmensen P, et al. Differences between local investigator and core laboratory interpretation of the admission electrocardiogram in patients with unstable angina pectoris or non-Q-wave myocardial infarction (a thrombin inhibition in myocardial ischemia [TRIM] substudy). Am J Cardiol. 1998;82:54–60.
7. Berge HM, Steine K, Andersen TE, et al. Visual or computer-based measurements: Important for interpretation of athletes' ECG. Br J Sports Med. 2014;48:761–767.
8. Garvey JL, Zegre-Hemsey J, Gregg R, Studnek JR. Electrocardiographic diagnosis of ST segment elevation myocardial infarction: An evaluation of three automated interpretation algorithms. J Electrocardiol. 2016;49:728–732.
9. Shah AP, Rubin SA. Errors in the computerized electrocardiogram interpretation of cardiac rhythm. J Electrocardiol. 2007;40:385–390.
10. Bogun F, Anh D, Kalahasty G, et al. Misdiagnosis of atrial fibrillation and its clinical consequences. Am J Med. 2004;117:636–642.
11. Southern WN, Arnsten JH. The effect of erroneous computer interpretation of ECGs on resident decision making. Med Decis Mak. 2009;29:372–376.
12. Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. ; 2013: 6645–6649.
13. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA - J Am Med Assoc. 2016;316:2402–2410.
14. Tschandl P, Rosendahl C, Akay BN, et al. Expert-Level Diagnosis of Nonpigmented Skin Cancer by Combined Convolutional Neural Networks. JAMA Dermatology. 2018;155:58–65.
15. Goodfellow I, Bengio Y, Courville A. Deep Learning. The MIT Press; 2016.
16. Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med. 2019;25. doi:10.1038/s41591-018-0268-3.
17. Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. 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;6736:1–7.
18. Galloway CD, Valys A V., Shreibati JB, et al. Development and Validation of a Deep-Learning Model to Screen for Hyperkalemia From the Electrocardiogram. JAMA Cardiol. 2019;55905:1–9.
19. Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nat Med. 2019;25:70–74.
20. Mason JW, Hancock EW, Gettes LS. Recommendations for the standardization and interpretation of the electrocardiogram: Part II: Electrocardiography diagnostic statement list: A scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council . Circulation. 2007;115:1325–1332.
21. Benoit K, Watanabe K, Wang H, et al. quanteda: An R package for the quantitative analysis of textual data. J Open Source Softw. 2018;3:774.
22. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: A resampling study. Stat Med. 2016;35:214–226.
23. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. CoRR. 2015;abs/1512.0. Available at http://arxiv.org/abs/1512.03385.
24. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. 2016. Available at http://arxiv.org/abs/1602.07261.
25. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: Bach F, Blei D, eds. Proceedings of the 32nd International Conference on Machine Learning. Lille, France: PMLR; 2015: 448–456.
26. Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J Mach Learn Res. 2014;15:1929–1958.
27. Glorot X, Bordes A, Bengio Y. Deep Sparse Rectifier Neural Networks Xavier. Proc 14th Int Con- ference Artif Intell Stat. 2011;15.
28. Yu F, Koltun V. Multi-Scale Context Aggregation by Dilated Convolutions. 2015. Available at http://arxiv.org/abs/1511.07122.
29. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. AIP Conf Proc. 2014;1631:58–62.
30. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal Loss for Dense Object Detection. Proc IEEE Int Conf Comput Vis. 2017;2017-Octob:2999–3007.
31. Paszke A, Gross S, Chintala S, et al. Automatic differentiation in PyTorch. 2017.
32. Cheng J, Wang Z, Pollastri G. A neural network approach to ordinal regression. Proc Int Jt Conf Neural Networks. 2008;:1279–1284.
33. Van Leeuwen KG, Sun H, Tabaeizadeh M, et al. Detecting Abnormal Electroencephalograms Using Deep Convolutional Networks. Clin Neurophysiol. 2018;130:77–84.
34. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Proc IEEE Int Conf Comput Vis. 2017;2017-Octob:618–626.
35. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for Simplicity: The All Convolutional Net. 2014;:1–14.
36. Krippendorff K, Hayes AF. Answering the call for a standard reliability measure for coding data. Commun Methods Meas. 2007;1:77–89.
37. Fleiss JL, Cohen J. The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. Educ Psychol Meas. 1973;33:613–619.
38. Van Calster B, Vergouwe Y, Looman CWN, et al. Assessing the discriminative ability of risk models for more than two outcome categories. Eur J Epidemiol. 2012;27:761–770.
39. Van Calster B, Van Belle V, Vergouwe Y, et al. Extending the c-statistic to nominal polytomous outcomes: The Polytomous Discrimination Index. Stat Med. 2012;31:2610–2626.

40. Hand DJ, Till RJ. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Mach Learn. 2001;45:171–186.
41. Li J, Gao M, D'Agostino R. Evaluating classification accuracy for modern learning approaches. Stat Med. 2019;38:2477–2503.
42. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Ann Intern Med. 2015;162:55.
43. Goy JJ, Schlaepfer J, Stauffer JC. Competency in interpretation of the 12-lead electrocardiogram among swiss doctors. Swiss Med Wkly. 2013;143:8–10.
44. Veronese G, Germini F, Ingrassia S, et al. Emergency physician accuracy in interpreting electrocardiograms with potential ST-segment elevation myocardial infarction: Is it enough? Acute Card Care. 2016;18:7–10.
45. McCabe JM, Armstrong EJ, Ku I, Kulkarni A, Hoffmayer KS, Bhave PD, Waldo SW, Hsue P, Stein JC, Marcus GM, Kinlay S, Ganz P. Physician accuracy in interpreting potential ST-segment elevation myocardial infarction electrocardiograms. J Am Heart Assoc. 2013;2:1–9.
46. Rolnick D, Veit A, Belongie S, Shavit N. Deep Learning is Robust to Massive Label Noise. ArXiv. 2017. Available at http://arxiv.org/abs/1705.10694.
47. Macfarlane PW, Katibi IA, Hamde ST, et al. Racial differences in the ECG - Selected aspects. J Electrocardiol. 2014;47:809–814.
48. Brosnan M, La Gerche A, Kumar S, Lo W, Kalman J, Prior D. Modest agreement in ECG interpretation limits the application of ECG screening in young athletes. Hear Rhythm. 2015;12:130–136.
49. Blundell C, Cornebise J, Kavukcuoglu K, Wierstra D. Weight Uncertainty in Neural Networks. 2015;37. Available at http://arxiv.org/abs/1505.05424.
50. Gal Y, Islam R, Ghahramani Z. Deep Bayesian Active Learning with Image Data. 2017. Available at http://arxiv.org/abs/1703.02910.



3

Supplemental material

Automatic Triage of 12-lead ECGs using Deep Convolutional Neural Networks: A First Implementation Study

European Heart Journal Digital Health

Rutger R van de Leur*, Meike TGM van Sleuwen*, Peter-Paul M Zwetsloot, Pim van der Harst, Pieter A Doevendans, Rutger J Hassink and René van Es

Abstract

Background and aims

Expert knowledge to correctly interpret electrocardiograms (ECGs) is not always readily available. An artificial intelligence (AI) based triage algorithm (DELTAnet), able to support physicians in ECG prioritization, could help reduce current logistic burden of overreading ECGs and improve time-to treatment for acute and life-threatening disorders. However, the effect of clinical implementation of such AI algorithms is rarely investigated.

Methods

Adult patients at non-cardiology departments who underwent ECG testing as a part of routine clinical care were included in this prospective cohort study. DELTAnet was used to classify 12-lead ECGs into one of the following triage classes: normal, abnormal not acute, subacute, and acute. Performance was compared to triage classes based on the final clinical diagnosis. Moreover, the associations between predicted classes and clinical outcomes were investigated.

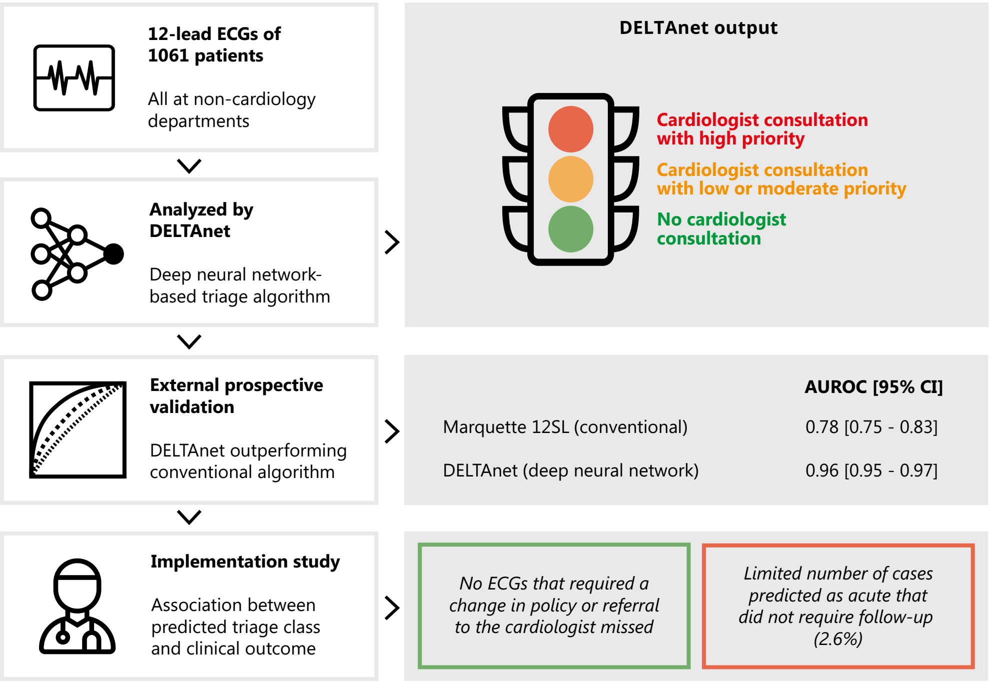
Results

A total of 1061 patients and ECGs were included. Performance was good with a mean concordance statistic of 0.96 [95% CI 0.95-0.97] when comparing DELTAnet to the clinical triage classes. Moreover, zero ECGs that required a change in policy or referral to the cardiologist were missed and there was a limited number of cases predicted as acute that did not require follow-up (2.6%).

Conclusions

This study is the first to prospectively investigate the impact of clinical implementation of an ECG-based AI triage algorithm. It shows that DELTAnet is efficacious and safe to be used in clinical practice for triage of 12-lead ECGs in non-cardiology hospital departments.

Graphical abstract


12-lead ECGs of 1061 patients
All at non-cardiology departments
Analyzed by DELTAnet
Deep neural network-based triage algorithm
External prospective validation
DELTAnet outperforming conventional algorithm
Implementation study
Association between predicted triage class and clinical outcome
DELTAnet output
Cardiologist consultation with high priority
Cardiologist consultation with low or moderate priority
No cardiologist consultation
AUROC [95% CI]
Marquette 12SL (conventional)
0.78 [0.75 - 0.83]
DELTAnet (deep neural network)
0.96 [0.95 - 0.97]
No ECGs that required a change in policy or referral to the cardiologist missed
Limited number of cases predicted as acute that did not require follow-up (2.6%)

Overview of the study and its outcomes. AUROC: area under the receiver operating curve, CI: confidence interval, ECG: electrocardiogram.

Introduction

Correct and timely interpretation of the electrocardiogram (ECG) is important for accurate diagnosis of a variety of cardiac abnormalities, as early treatment results in lower mortality and decreases disease burden for life-threatening cardiac disorders.[1–3] Expert knowledge to interpret ECGs is often not readily available, especially in prehospital care and non-cardiology departments.[4–6] Accurately prioritizing which ECGs need expert attention could lead to improvements in time to treatment and enhance the cost-effectiveness of current healthcare.[7,8]

Recent advancements in the field of artificial intelligence have shown that deep neural networks (DNN) can learn to interpret ECGs with high accuracy.[9] Previous studies have shown that DNNs can be used to detect many separate ECG abnormalities, such as rhythm and conduction disorders.[10–13] Our group developed a comprehensive DNN based triage algorithm (DELTAnet) that is able to consistently triage all 12-lead hospital ECGs.[14] DELTAnet was trained to classify each ECG into one of the following 4 classes based on how quickly a cardiologist should be consulted: (1) normal; no action needed, (2) abnormal not acute; consultation with low priority, (3) sub-acute; consultation with moderate priority, or (4) acute; consultation with high priority. This algorithm was validated in an expert-annotated test set and shows potential to support physicians in comprehensive triage and decision making regarding the prioritization of a newly acquired ECG.

Despite the rise in AI-optimized ECG interpretation approaches, clinical implementation of these algorithms is limited. Essential steps should be completed before clinical implementation is possible: (1) development and internal validation, (2) external validation in other populations, and (3) assessment of the implementation of the model in clinical practice with its impact on patient outcomes.[15] Most studies regarding automated ECG applications address the first two phases, but their implications for implementation remain unclear. In this study, we aim to prospectively validate the performance of DELTAnet and investigate the impact of possible implementation of DELTAnet in clinical practice when applied to 12-lead ECGs from multiple non-cardiology hospital departments.

Methods

Study Setting and Participants

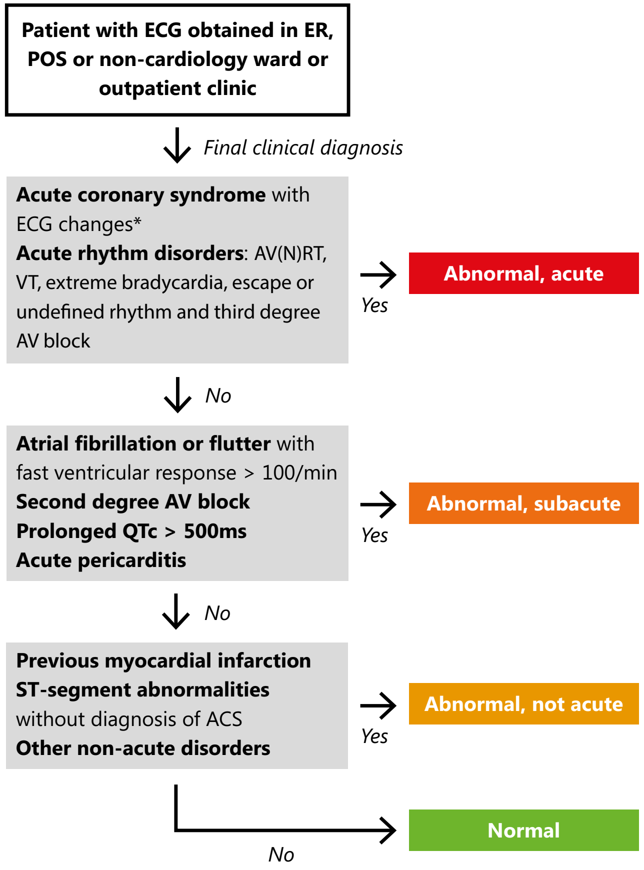
We conducted a prospective, single-center, consecutive and observational cohort study with adult inpatients who underwent ECG testing as a part of routine clinical care at University Medical Center Utrecht (UMCU, Utrecht, the Netherlands). Patients were included when their ECG was acquired in one of the following departments: the emergency room (ER), pre-operative screening department (POS), a non-cardiology ward or a non-cardiology outpatient clinic between October 1st and 31st of 2019. All ECGs were interpreted by a cardiologist or cardiology resident as part of the regular clinical workflow. Patients were excluded if their ECG was of insufficient quality, as annotated by the overreading physician. For patients with multiple ECGs acquired during their hospital stay, only the first ECG was selected for analysis. All ECGs were acquired using a General Electric MAC 5500 (GE Healthcare, Chicago, IL, United States) and electrodes could have been placed both in the standard or Mason-Likar configuration. The study was conducted under a protocol approved by the UMCU Institutional Review Board using a waiver of written informed consent.

Triage classification of the ECG

We used a previously described deep learning-based triage algorithm (DELTAnet) that was developed and validated for comprehensive triage of 12-lead ECGs.[14] In short, DELTAnet is a 37-layer convolutional neural network trained to triage ECG using a dataset of 336.835 ECGs from 142.040 patients. For training, the physician annotations of each ECG were translated into one of the triage classes based on predefined criteria, and these triage classes were used to train the algorithm. Hyperparameters were tuned using a combination of manual tuning and random grid search on a subset of 5% of the training dataset. DELTAnet was validated on an export-annotated test set of 984 ECGs from 984 patients. The algorithms outputs one of 4 triage categories, based on how quickly a cardiologist has to be consulted: [1] normal, [2] not acute abnormal (consultation with low priority), [3] subacute abnormal (consultation with moderate priority) and [4] acute abnormal (consultation with high priority). For this study, custom software automatically extracts the raw ECG data from the MUSE system (GE Healthcare, Chicago, IL, United States) and this data was then triaged by the DELTAnet algorithm on a standard desktop computer. The DELTAnet prediction was not shown to the physician in this study.

We evaluated its performance by comparing the predicted triage classes to the triage classes as based on the final clinical diagnosis. In the development study, the model was trained and validated using only the physician annotation of the ECG categorized into one of the triage categories. Detailed ECG interpretation also needs additional

Figure 1. ***Labeling into triage classes as based on final clinical diagnosis.*** For few cases where the cardiologist annotated diagnosis was not clear (e.g. whether specific or non-specific ST-abnormalities), the ECG was assessed to determine the appropriate category. When multiple diagnoses were visible on the ECG, the highest triage class was chosen. All other non-acute disorders can be found in Supplemental Figure 1. * ECG changes were defined as ST-segment deviations or T-wave changes associated with ischemia. Abbreviations: ACS = acute coronary syndrome, AV(N)RT = Atrioventricular (nodal) reentry tachycardia, VT = ventricular tachycardia, AV block = atrioventricular block, AF = atrial fibrillation.


Patient with ECG obtained in ER, POS or non-cardiology ward or outpatient clinic
Final clinical diagnosis
Acute coronary syndrome with ECG changes*
Acute rhythm disorders: AV(N)RT, VT, extreme bradycardia, escape or undefined rhythm and third degree AV block
Yes
Abnormal, acute
No
Atrial fibrillation or flutter with fast ventricular response > 100/min
Second degree AV block
Prolonged QTc > 500ms
Acute pericarditis
Yes
Abnormal, subacute
No
Previous myocardial infarction
ST-segment abnormalities without diagnosis of ACS
Other non-acute disorders
Yes
Abnormal, not acute
No
Normal

clinical information, such as patient history, previous ECGs and results of other tests. In the current analysis, the final clinical diagnosis was therefore extracted from medical record data and used to determine the clinical triage classes using the flowchart in **Figure 1** and the diagnostic statement to triage class matrix in **Supplemental Figure 1**. The major difference between the current class definition and the one used for training is that ST-segment abnormalities are classified as either acute or not-acute based on the outcomes of laboratory tests and coronary angiography.

For comparison purposes we sought a commercially available and widely used conventional rule-based (ie. not deep learning-based) algorithm for interpretation of 12-lead ECGs. The Marquette 12SL algorithm (GE Healthcare, Chicago, IL, United States) was selected, as it is used in all GE ECG systems and currently provides the computerized interpretation of the ECG in our hospital.[16] Marquette 12SL diagnostic statements were manually mapped to triage classes based on **Supplemental Figure 1**.

Association with clinical care and outcomes

Currently, in our hospital all ECGs acquired at non-cardiology departments are systematically overread by a cardiologist or cardiology resident within 24 hours, or faster when another physician asks for a consult. This a time-consuming tasks, which places a heavy logistic burden on clinical practice. To optimize this process, DELTAnet recommends the physician whether cardiologist consultation or overreading of the ECG is necessary and within which timeframe. Normal ECGs are no longer overread, while for acute ECGs consultation is immediate. To assess the effect of implementing the DELTAnet recommendation in clinical practice, the association between the predicted triage class and the currently chosen management for this patient was evaluated. For all patients, the following events were logged: cardiologist consultation, whether there was a change in patient management (diagnostics, a medication change or a cardiac procedure, such as electrical cardioversion or coronary angiography), follow-up appointment at a cardiology clinic, the final clinical diagnosis (whether cardiac/non-cardiac) and clinical outcomes (length of hospital stay and in-hospital mortality). We evaluated whether using DELTAnet to guide physicians resulted in similar management for patients as in current clinical practice in our hospital.

Outcome Measures

The outcome of this study is the expected impact of future implementation of DELTAnet into clinical practice in non-cardiology wards. We evaluated the classification performance as compared to the final clinical diagnosis in both the overall cohort as well as in several subgroups (age, gender, hospital location). Moreover, we compared the management of the patients between the predicted triage classes and focus on two outcomes: 1) no important undertriage; defined as ECGs as normal that should have required cardiac follow-up, and 2) a limited proportion of overtriage; defined as ECGs predicted as acute that did not require any cardiac follow-up or had no final diagnosis of cardiac disease.

Statistical Analyses

For descriptive analysis, proportions and percentages and means with standard deviations (SD), or medians with interquartile ranges (IQR) when data was not normally distributed, were calculated. Overall classification performance was evaluated in terms of the unweighted mean of all pairwise concordance-statistics (c-statistics, equivalent to area under the receiver operating curve).[17,18] This method is robust to class imbalance and calculates the area under the receiver operating curve for all pairs of classes. Given the number of classes , any pair of classes and and the measure of separability between two classes, this metric is defined as follows:[18]

$$M = \frac{2}{c(c-1)} \sum_{i<j} \hat{A}(i,j)$$

For category-specific performance, we assessed c-statistics, sensitivity, specificity, positive predictive values (PPV) and negative predictive values (NPV). All category-specific measures were applied in a one-versus-other approach. To estimate the 95% confidence interval (CI) of the performance metrics, we used 2000 rounds of bootstrapping. C-statistics were compared using permutation tests. The TRIPOD Guidelines were followed where applicable.[19]

Results

Patient characteristics

A total of 1061 patients were found eligible, and 48 were excluded due to technically insufficient recording quality of the ECG. The distribution of predicted triage categories was unbalanced with the most recordings being normal (52%) and the least belonging to the subacute group (4%). Most ECGs were acquired at the ER (42%) and the smallest subset contained ECGs obtained at non-cardiology wards (14%). **Table 1** summarizes the patient characteristics, hospital locations and predicted triage class distributions of the data set.

Classification performance

The overall classification performance of DELTAnet, as measured by the unweighted mean of pairwise c-statistics, was 0.96 [95% CI 0.95 - 0.97] when comparing the predicted triage classes to the final clinical diagnosis. DELTAnet outperformed the Marquette 12SL algorithm, which had an unweighted mean of pairwise c-statistics of 0.78 [95% CI 0.75-0.83, $p < 0.001$]. The c-statistics, sensitivities, specificities, positive predictive values, and negative predictive values per triage category of DELTAnet are shown in **Table 2** and the corresponding confusion matrix in **Figure 2**. Classification performance was good for all subgroups (Supplemental Figure 2). None of the pairwise combinations showed significant differences between subgroups.

Under- and overtriage

For 59 patients (5.6%), DELTAnet predicted a lower triage class than was determined by the final clinical diagnosis (undertriage). Most undertriage consisted of patients classified as not acute but predicted to be normal by DELTAnet (57/59, 97% of all undertriage). These patients were classified as not acute based on nonspecific ST-abnormalities (21/57, 37%), incomplete right bundle branch block (9/57, 16%), aspecific intraventricular conduction delay (7/57, 12%), previous ischemia (6/57, 11%), bradycardia <50 bpm (6/57, 11%), left ventricular hypertrophy (4/57, 7%), low QRS voltage (2/57, 4%), or a combination of mentioned abnormalities (2/57, 4%). One undertriage case (1/59, 2%) concerned a patient classified as acute but predicted to be non-acute by DELTAnet. This represented a patient that presented at the ER with chest pain, atypical ST-elevation in lead V2 and V3 without reciprocal depression and low troponin. The patient was clinically triaged as acute because the final diagnosis was unstable angina with coronary stenosis on coronary angiography (**Supplemental Figure 3**). The last undertriage case (1/59, 2%) reflected a patient classified as sub-acute because of QTc >500ms, but was predicted as non-acute by DELTAnet.

		OVERALL **n=1013**
DEMOGRAPHICS	**Age, median (IQR)**	64 (52-73)
	Female sex, n (%)	439 (43%)
	BMI, median (IQR)	29 (25-30)
	History of cardiovascular disease, n (%)	506 (49%)
	Cardiac procedure in history, n (%)	275 (27%)
RISK FACTORS	**Hypertension**	467 (45%)
	Diabetes	298 (29%)
	High Cholesterol	232 (23%)
	Smoking	464 (46%)

		OVERALL **n=1013**
LOCATION, N (%)	**Emergency room**	430 (42%)
	Non-cardiology outpatient clinic	253 (25%)
	Non-cardiology ward	143 (14%)
	Pre-operative screening	187 (18%)
PREDICTED TRIAGE CLASS, N (%)	**Normal**	529 (52%)
	Abnormal, not acute	373 (37%)
	Abnormal, subacute	29 (3%)
	Abnormal, acute	82 (8%)

Table 1 *Patient characteristics and distributions in hospital location and triage classes.*

For 102 patients (9.6%), DELTAnet predicted a higher triage class than was determined by the final clinical diagnosis (overtriage). Most overtriage occurred for patients classified as non-acute but predicted to be acute by DELTAnet (47/102, 46%). Overpredictions mainly represented patients with ST-abnormalities on the ECG but no final clinical diagnosis of acute ischemia (39/47, 72%) (**Supplemental Figure 4**). Other non-acute overpredictions concerned patients with atrial fibrillation on the ECG (7/47, 15%), of which three patients showed atrial fibrillation in combination with other non-acute abnormalities (nonspecific ST-abnormalities or left fascicular block) (**Supplemental Figure 5**), and the last case (1/47, 2%) represented a paced rhythm (**Supplemental Figure 6**).

With the Marquette 12SL algorithm, undertriage was observed in 73 patients (6.8%) and overtriage in 155 patients (15%). The four patients with an acute final diagnosis, but misclassified as not acute, presented with pan-ischemia (2/2, 50%) and high-degree AV block (2/2, 50%). The 13 patients with a subacute final diagnosis, but misclassified as not acute or normal, presented with a prolonged QT interval (6/13, 46%), atrial fibrillation with fast ventricular response (5/13, 38%) or pericarditis (2/13, 16%). In 62 of the 66 ECG misclassified as acute (94%), but with a final not acute diagnosis, signs of ischemia were mentioned in the Marquette 12SL diagnosis.

Associations with clinical care and outcomes

Overall, patients with a higher predicted triage class were more often referred for cardiac follow-up, more often diagnosed with cardiac disease, and had worse clinical outcomes (i.e. longer hospital admission and higher mortality rate, **Table 3**). Moreover, patients with a higher predicted class in general also represented patients with more severe clinical diagnoses (**Supplemental Table 1**).

Of the 529 (52% of the cohort) patients with an ECG classified as normal by DELTAnet, a cardiologist was consulted in 79/529 (15%) of the cases, mostly in the ER. For most patients this did not results in a change of management (45/529, 8.5%). For the other 34/529 (6.4%) patients, follow-up was recommended (additional diagnostics or admission to a cardiology ward) in 15/34 (44%) patients, a change in medication was made in 14/34

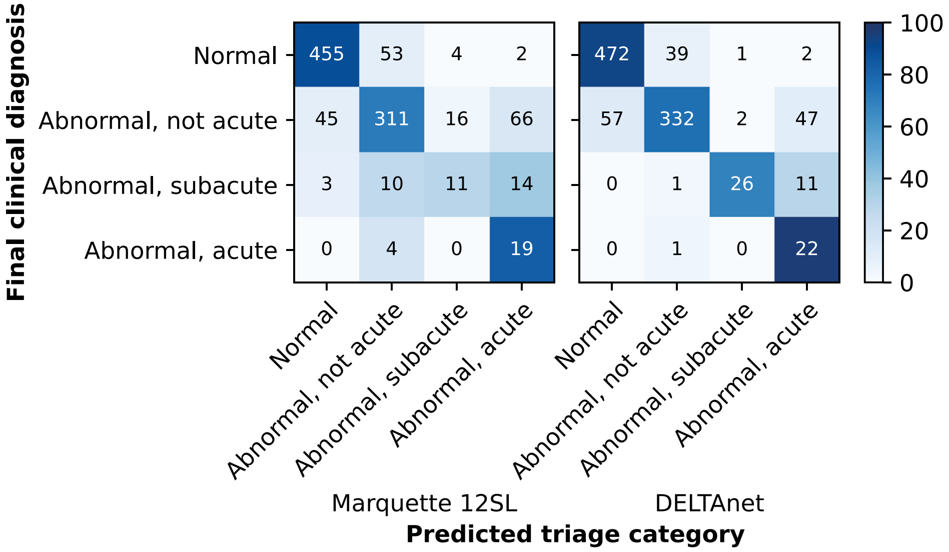


Figure 2. *Confusion matrix comparing Marquette 12SL and DELTAnet predictions to the clinical triage classes (based on final clinical diagnosis).* The color map represents the percentage in the clinical triage class, normalized per row.

(41%) patients and a cardiac procedure (percutaneous coronary intervention, coronary artery bypass grafting surgery, pericardiocentesis or pacemaker implantation) was performed in 5/34 (15%) patients. Of these 34 patients with a change in management, 1 patient represented acute pathology (acute coronary syndrome without clear ECG abnormalities), the others all represented normal or abnormal non-acute patients. In these cases, a cardiologist was consulted based on cardiac complaints, or for other questions (e.g. to evaluate possible cardiac spread of infections, or help in determining the appropriate treatment plan because of a history of cardiac disease: 17/34 patients were already known with cardiac disease). All 34 cases are described in detail in **Supplemental Figures 7 to 40**.

Of the 82 patients with an ECG predicted as acute by DELTAnet, a cardiologist was consulted in 55/82 (67%) patients. In the 27 (34%) patients where no cardiologist was consulted, most patients (15/27, 56%) had ECG abnormalities consistent with previous ECGs and therefore did not require follow-up. These ECGs were mostly acquired for other reasons than clinical complaints: only 2/27 (7%) patients had symptoms of chest pain, while 8/27 (30%) were routine control ECGs at the POS or outpatient clinic, 3/27 (11%) were acquired to evaluate whether a medication change would be allowed (risk for long QT abnormalities) and for the other 14/27 (52%) the reason for ECG was not documented.

	NORMAL	**ABNORMAL, NOT ACUTE**	**ABNORMAL, SUBACUTE**	**ABNORMAL, ACUTE**
C-STATISTIC (95% CI)	0.95(0.94-0.96)	0.91 (0.89 – 0.93)	0.99 (0.98-1.00)	0.98 (0.97-0.99)
SENSITIVITY	0.92	0.76	0.68	0.96
SPECIFICITY	0.89	0.93	1.0	0.94
PPV	0.89	0.89	0.90	0.27
PPV	0.91	0.83	0.99	1.0

Table 2. ***Performance measures per triage class comparing predicted triage classes by DELTAnet with clinical triage classes.*** C-statistics (concordance statistic), sensitivities, specificities, positive and negative predictive values are all calculated in a 1-vs-other approach. CI = confidence interval, PPV = positive predictive value, NPV = negative predictive value.

		NORMAL n=529	**ABNORMAL, NOT ACUTE n=373**	**ABNORMAL, SUBACUTE n=29**	**ABNORMAL, ACUTE n=82**	**P-VALUE**
FOLLOW-UP, N (%)	**Cardiologist consulted?**	79 (15%)	117 (31%)	19 (66%)	54 (66%)	<0.0001**
	Change in management after ECG*	34 (6%)	72 (19%)	17 (59%)	47 (57%)	<0.0001**
	Follow-up appointment cardio clinic?	36 (7%)	80 (21%)	11 (38%)	31 (38%)	<0.0001**
	Final diagnosis of cardiac disease, n (%)	27 (5%)	194 (52%)	24 (83%)	53 (65%)	<0.0001**
CLINICAL OUTCOMES	**Length of Stay, mean (SD), days**	3 ± 13	4 ± 10	6 ± 13	6 ± 21	<0.0001***
	Hospital Mortality	3 (0.6%)	8 (2%)	0 (0%)	10 (12%)	<0.0001***

Table 3. ***Differences in provided clinical follow-up, diagnosis and outcomes per predicted triage class. A p-value of <0.05 denotes significant difference.*** *change in management defined as either a cardiac medication change, a performed cardiac procedure, when the patient was admitted to a cardiology department, or when cardiac follow-up diagnostics (e.g. ECG/lab) were proposed.
**pairwise comparisons showed a significant difference for all pairwise comparisons (p < 0.001), except for the difference between the sub-acute and acute group (all p > 0.05).
***pairwise comparisons showed a significant difference for the normal group versus either the abnormal not acute, sub-acute or acute group (p < 0.001) for the in-hospital length-of stay. Pairwise comparisons showed a significant difference between the normal and acute group and the abnormal not-acute and acute group (p < 0.001) for both in-hospital and 1-year mortality.

Discussion

This study is the first to prospectively assess the impact of implementing an ECG-based AI algorithm for triage of 12-lead ECGs in non-cardiology departments. We demonstrated DELTAnet to be safe when implemented in clinical practice: no important ECGs were missed and the number of ECGs predicted as acute that did not require follow-up was very limited (2.6%). Moreover, we showed excellent classification performance for both the overall population and when stratified in subgroups, similar to the original test dataset, and outperforming the currently employed Marquette 12SL algorithm.[14] Therewith, this indicates that DELTAnet can be safely used to prioritize non-cardiology ECGs by automatically assessing normal ECGs and by potentially warning the physician for acute ECGs requiring immediate follow-up by a cardiologist.

Classification performance of DELTAnet was excellent in this prospective validation dataset with an overall c-statistic of 0.96 [95% CI 0.95 - 0.97], comparable to the performance during internal validation (c-statistic 0.93 [95% CI 0.92 – 0.95]).[14] The internal validation dataset was annotated by a panel of electrophysiologists that only had access to the ECG. For some ECG abnormalities, such as wide complex tachycardia or ST-segment deviations, additional information from previous ECGs, follow-up or additional diagnostics is needed for accurate triage. In the current validation dataset we therefore took all clinical data into account to determine the final clinical diagnosis associated to this ECG. This led to many previously acute ECGs being classified as not acute. It turns out that in the previous study the panel labeled many ECGs with ST-segment abnormalities or wide complex tachycardia as acute when having no knowledge of the other clinical information. The reclassification in the current analysis led to an increase of the sensitivity of the acute class from 79% to 96%, without reducing specificity.

One other study investigated the use of deep neural networks for triage of ECGs in the Emergency Department and showed improved performance over a conventional rule-based ECG algorithm.[20] Although their algorithm shows similar sensitivity and specificity for differentiating normal and abnormal ECGs, DELTAnet greatly outperforms their sensitivity in detecting acute ECGs (53% vs 96%), making it much safer for use in clinical practice. Comparison to other studies remains challenging, as wide varieties of ECG abnormalities are assessed in different studies using different metrics. One important observation from a recent meta-analysis is, however, that non-cardiologist physicians perform poorly in interpreting ECGs with a pooled accuracy of 69%.[5] This is exactly the area where the current algorithm can be used to prevent important ECGs from being missed while saving time by prioritizing other ECGs.

Over- and undertriage

For DELTAnet to be safe and efficacious for implementation into clinical practice, undertriage (failure to identify patients that need to be referred) and overtriage (false alarms, unnecessary consultations of the cardiologist) should be minimized. DELTAnet showed very high negative predictive values compared to clinical triage classes for the acute classes (NPV = 0.99, **Table 2**). This is among the most important findings of the current study, as it allows for safe implementation of the algorithm in clinical practice. It must be noted that there were some cases of non-ST-elevation acute coronary syndrome and unstable angina classified as normal or not acute, but these patients did not have ECG abnormalities at the time (**Supplemental Table 1**). Therefore, one should realize that the main goal of DELTAnet is to support physicians in decision-making regarding the acuteness and prioritization of new acquired ECGs; DELTAnet does not aim to (and will not be able to) substitute clinical decision-making. Only patients with ECG abnormalities at time of ECG can be detected using such an algorithm.

Most undertriage was seen for the abormal, not acute class, where 7.7% of ECGs with that final diagnosis were classified as normal (**Figure 2**). Detailed inspection of the cases showed that this was mostly due to disagreement between the treating physician and algorithm on the meaning of non-specific ST-segment abnormalities. In practice, in 15% of the patients with an ECG classified as normal by DELTAnet a cardiologist was consulted, which resulted in a change of management in 6% of patients. These cases concerned patients where the clinical presentation of the patient was leading in clinical decision-making and no or minimal abnormalities were seen on the ECG. None of these patients would therefore have been wrongfully overlooked by a cardiologist if DELTAnet would have been implemented.

The main challenge of the algorithm resides in the overtriage of acute disease, showing a lower positive predictive value (0.27) for this class. This lower PPV results from weighting in the training phase, where the algorithm was penalized for undertriaging acute ECGs, as this might cause undesirable false negatives in clinical practice. The PPV in the current validation set is lower than the original study. The panel that labeled the validation dataset in the original study marked many ECGs as acute based on ST-segment abnormalities. These are now classified as abnormal not acute when taking previous ECGs and other tests (such as troponin testing and the outcomes of coronary angiography) into account. This distinction between non-acute and acute ST-abnormalities remains a challenge, especially as DELTAnet cannot take into account symptoms or previous ECG without a current diagnosis of ACS. Overtriage is not expected to have much negative consequences: ST-abnormalities can be dangerous when undetected, so consultation with an expert to justify or rule-out possible ischemia seems appropriate in these cases. The high rate of false positives for the acute class could lead to alarm fatigue, as most of the ECGs predicted as acute do not need acute follow-up.[21] Overall, however, these false positives only account for 5.9% of the ECGs, lowering the risks of alarm fatigue (**Figure 2**).

Limitations

There are several limitations to address. First, the number of times a cardiologist was consulted might be underestimated, as this may not always have been logged. However, important cases are always documented so it can be assumed that when not documented, no further follow-up was required. Second, our study is a background implementation study and therefore we were not able to perform extra diagnostic tests to justify the results or perform further investigation. This could have lead to an underestimation of undertriage of the acute class, as patients with acute myocardial ischemia could have been missed completely. Also, we are not able to assess the effect that implementation of DELTAnet would have on clinical decision-making, despite the prospective nature of this study.

Implications for future work

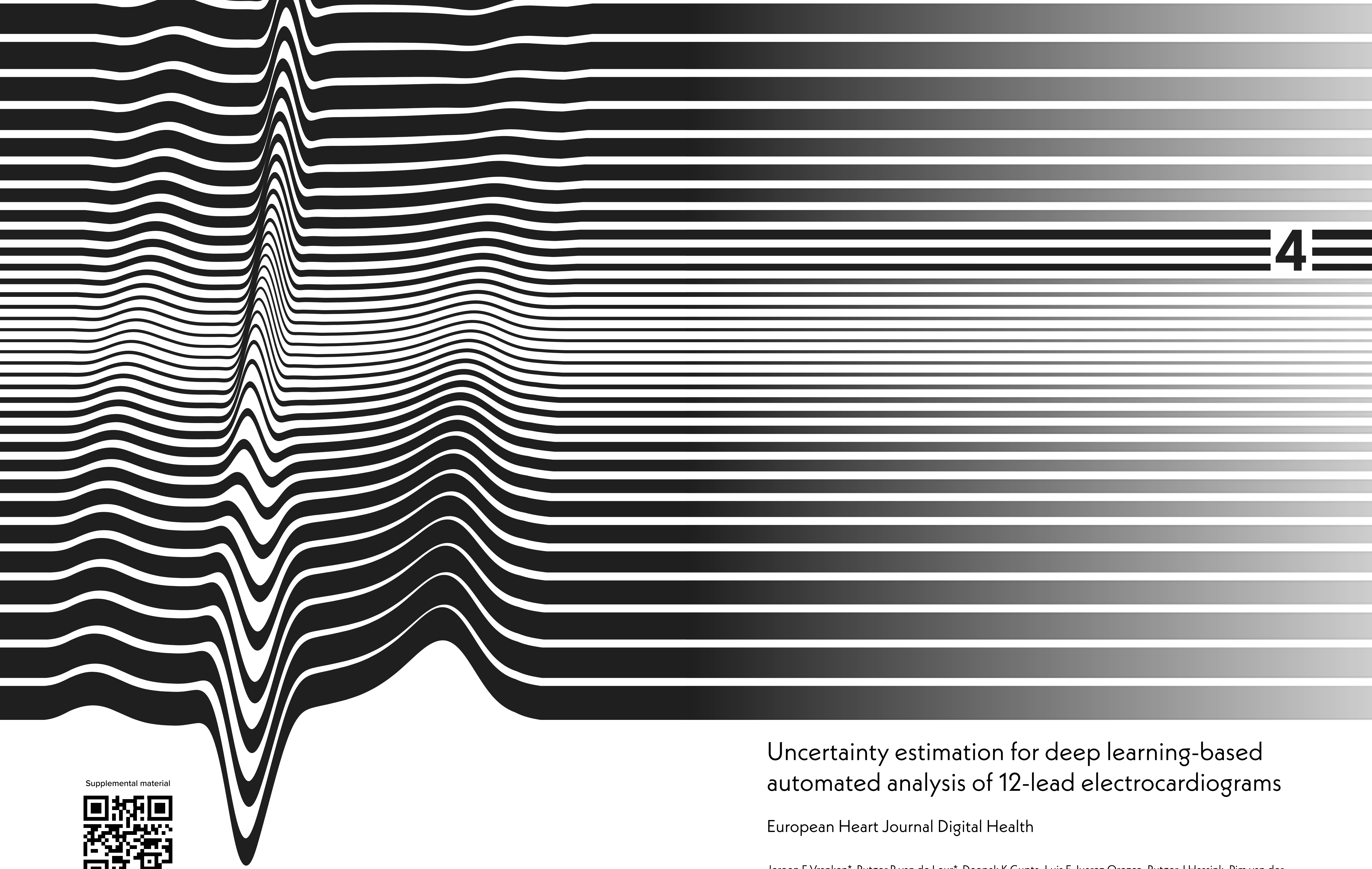
An important next step will be to perform a randomized controlled trial to evaluate implementation in real-world clinical practice with its true impact on clinical care and patient outcomes. Other future perspectives to improve its clinical applicability include adding visualization methods and uncertainty models that can identify the cases the algorithm is prone to misdiagnose.[22,23] In addition, another future goal is to investigate whether automated comparison of a new acquired ECG to previous ECGs would be possible. Eventually, an 'AI-ECG dashboard' needs to be developed that is able to clearly present the ECG with predicted triage categories along with ECG features important for prediction. At last, a goal is to implement DELTAnet in mobile ECG devices, making it applicable for use in pre-hospital settings or places where standard 12-lead ECG is not readily available.

Conclusion

This study is the first to prospectively validate an ECG-based AI triage algorithm and provide insight into its clinical implementation. We demonstrated that DELTAnet is safe to be used in clinical practice for triage of 12-lead ECGs, acquired at non-cardiology departments, and outperformed the currently employed algorithm for computerized interpretation of the ECG (Marquette 12SL). Implementation of DELTAnet could possibly lead to decreased workload for physicians and quicker recognition of acute life-threatening cardiac disorders. As a next step, a randomized study will be performed to evaluate its added value on clinical care and patient outcomes compared to current care.

REFERENCES

1. Ibanez B, James S, Agewall S, et al. 2017 ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation. *Eur Heart J* 2018; 39: 119--177.
2. Foo CY, Bonsu KO, Nallamothu BK, et al. Coronary intervention door-to-balloon time and outcomes in ST-elevation myocardial infarction: A meta-analysis. *Heart* 2018; 104: 1362--1369.
3. Terkelsen CJ, Sørensen JT, Maeng M, et al. System Delay and Mortality Among Patients With STEMI Treated With Primary Percutaneous Coronary Intervention. *Jama* 2010; 304: 763–771.
4. Eslava D, Dhillon S, Berger J, et al. Interpretation of electrocardiograms by first-year residents: the need for change. *J Electrocardiol* 2009; 42: 693--697.
5. Cook DA, Oh S-Y, Pusic MV. Accuracy of Physicians' Electrocardiogram Interpretations. *Jama Intern Med* 2020; 180: 1461.
6. Salerno SM, Alguire PC, Waxman HS. Competency in Interpretation of 12-Lead Electrocardiograms: A Summary and Appraisal of Published Evidence. *Ann Intern Med* 2003; 138: 751.
7. Schläpfer J, Wellens HJ. Computer-Interpreted Electrocardiograms Benefits and Limitations. *J Am Coll Cardiol* 2017; 70: 1183–1192.
8. Diercks DB, Kontos MC, Chen AY, et al. Utilization and Impact of Pre-Hospital Electrocardiograms for Patients With Acute ST-Segment Elevation Myocardial Infarction. *J Am Coll Cardiol* 2009; 53: 161--166.
9. Leur RR van de, Boonstra MJ, Bagheri A, et al. Big Data and Artificial Intelligence: Opportunities and Threats in Electrophysiology. *Arrhythmia Electrophysiol Rev* 2020; 9: 146–154.
10. Krasteva V, Christov I, Naydenov S, et al. Application of Dense Neural Networks for Detection of Atrial Fibrillation and Ranking of Augmented ECG Feature Set. *Sensors* 2021; 21: 6848.
11. Yang T, Gregg RE, Babaeizadeh S. Detection of strict left bundle branch block by neural network and a method to test detection consistency. *Physiol Meas* 2020; 41: 025005.
12. Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med* 2019; 25: 65–69.
13. Kashou AH, Ko W-Y, Attia ZI, et al. A comprehensive artificial intelligence–enabled electrocardiogram interpretation program. *Cardiovasc Digital Heal J* 2020; 1: 62--70.
14. Leur RR van de, Blom LJ, Gavves E, et al. Automatic Triage of 12-Lead Electrocardiograms Using Deep Convolutional Neural Networks. *J Am Heart Assoc*; 9. Epub ahead of print 2020. DOI: 10.1161/jaha.119.015138.
15. Moons KGM, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012; 98: 691.
16. GE Healthcare - 2012 - Marquette 12SL ECG Analysis Program Physician's Guide.pdf.
17. Calster BV, Vergouwe Y, Looman CWN, et al. Assessing the discriminative ability of risk models for more than two outcome categories. *Eur J Epidemiol* 2012; 27: 761--770.
18. Hand DJ, Till RJ. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Mach Learn* 2001; 45: 171--186.
19. Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med* 2015; 162: 55.
20. Smith SW, Walsh B, Grauer K, et al. A deep neural network learning algorithm outperforms a conventional algorithm for emergency department electrocardiogram interpretation. *J Electrocardiol* 2019; 52: 88--95.
21. Hravnak M, Pellathy T, Chen L, et al. A call to alarms: Current state and future directions in the battle against alarm fatigue. *J Electrocardiol* 2018; 51: S44–S48.
22. Vranken JF, Leur RR van de, Gupta DK, et al. Uncertainty estimation for deep learning-based automated analysis of 12-lead electrocardiograms. *European Hear J - Digital Heal*. Epub ahead of print 2021. DOI: 10.1093/ehjdh/ztab045.
23. Leur RR van de, Bos MN, Taha K, et al. Improving explainability of deep neural network-based electrocardiogram interpretation using variational auto-encoders. *European Hear J - Digital Heal*. Epub ahead of print 2022. DOI: 10.1093/ehjdh/ztac038.



4

Supplemental material

Uncertainty estimation for deep learning-based automated analysis of 12-lead electrocardiograms

European Heart Journal Digital Health

Jeroen F Vranken*, Rutger R van de Leur*, Deepak K Gupta, Luis E Juarez Orozco, Rutger J Hassink, Pim van der Harst, Pieter A Doevendans, Sadaf Gulshad and René van Es

Abstract

Aims

Automated interpretation of electrocardiograms (ECGs) using deep neural networks (DNNs) has gained much attention recently. While the initial results have been encouraging, limited attention has been paid to whether such results can be trusted, which is paramount for their clinical implementation. This study aims to systematically investigate uncertainty estimation techniques for automated classification of ECGs using DNNs and to gain insight into its utility through a clinical simulation.

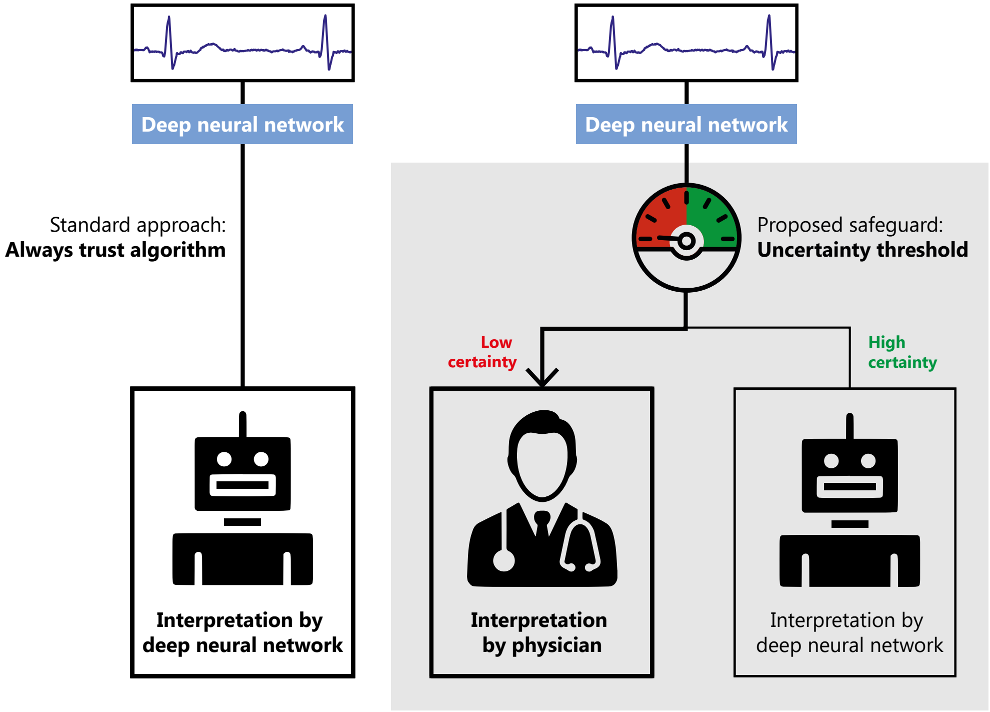
Methods and results

On a total of 526,656 ECGs from three different datasets, six different methods for estimation of aleatoric and epistemic uncertainty were systematically investigated. The methods were evaluated based on ranking, calibration and robustness against out-of-distribution data. Furthermore, a clinical simulation was performed where increasing uncertainty thresholds were applied to achieve a clinically acceptable performance. Finally, the correspondence between the uncertainty of ECGs and the lack of interpretational agreement between cardiologists was estimated. Results demonstrated the largest benefit when modelling both epistemic and aleatoric uncertainty. Notably, the combination of variational inference with Bayesian decomposition and ensemble with auxiliary output outperformed the other methods. The clinical simulation showed that accuracy of the algorithm increased as uncertain predictions were referred to the physician. Moreover, high uncertainty in DNN-based ECG classification strongly corresponded with lower diagnostic agreement in cardiologist's interpretation ($p < 0.001$).

Conclusion

Uncertainty estimation is warranted in automated DNN-based ECG classification and its accurate estimation enables intermediate quality control in the clinical implementation of deep learning. This is an important step towards clinical applicability of automated ECG diagnosis using DNNs.

Graphical abstract


Deep neural network
Deep neural network
Standard approach:
Always trust algorithm
Proposed safeguard:
Uncertainty threshold
Low certainty
High certainty
Interpretation by deep neural network
Interpretation by physician
Interpretation by deep neural network

Introduction

Worldwide, more than 300 million electrocardiograms (ECGs) are annually acquired, making it the most widespread cardiological diagnostic test in use. The ECG is utilized in daily clinical practice to diagnose a wide range of potentially life-threatening abnormalities and its correct interpretation requires expert knowledge from an experienced cardiologist, which might not always be directly available. Moreover, the massive number of ECGs acquired places a considerable logistic burden on the clinical routine.[1] Computerized interpretation of the ECG (CIE) has become increasingly important in supporting clinical practice. However, CIE has not yet been able to reach cardiologist-level accuracy and overreading automated ECG interpretations remains necessary.[2]

Substantial improvement in CIE is forthcoming with the development of deep learning algorithms that can learn abstract features from the raw ECG signal without the need for laborious hand-crafted feature extraction. Recent studies have shown encouraging results of deep neural networks (DNNs) applied to ECGs, ranging from detection of selected arrhythmias or conduction disorders to comprehensive interpretation for automatic triage.[3–5] While such reports have demonstrated the efficacy of deep learning in ECG analysis, there are additional challenges to be addressed before deep learning-based methods can be deployed in clinical practice.[6]

One such challenge is found in the fact that current deep learning models are architecturally forced to provide an output that translates to a diagnosis or prediction, while not reporting back to the user the degree to which such output might be uncertain (i.e. to which degree the model *does not know* the output is indeed correct). This output is provided even when the model has not seen the input before. Therefore, all prior deep learning models reported have been promoted without any evaluation or management of the *uncertainty* associated to their estimations.[7,8] It has been argued that the Softmax output (the probability distribution of predicted classes) of a regular DNN can also be interpreted as a measure of uncertainty. However, research has shown that this produces erroneous predictions with high confidence on unseen data and is therefore unsuitable for safety-critical applications.[9]

In clinical practice, expert clinicians consult colleagues or literature when confronted with complex cases that carry diagnostic uncertainty, which is then addressed through re-evaluation and consensus. Accordingly, it is highly desirable for deep learning algorithms employed in CIE to report some measure of uncertainty along with their diagnostic or predictive output so that equivocal cases can be re-evaluated by an experienced cardiologist.

For any diagnostic or predictive model, there are two distinct causes for the *uncertainty* of its prediction. These two are referred to as aleatoric and epistemic uncertainty (**Supplementary Figure 1**).[10] Aleatoric uncertainty arises from noise inherent in the data, such as high-frequency noise, lead reversals, baseline drift, or borderline cases present in the ECG recording, and can therefore not be reduced by further data collection. Alternatively, epistemic uncertainty is caused by a lack of knowledge from the algorithm, which for instance has not been exposed to a specific (disease) pattern during training. Epistemic uncertainty can therefore be reduced by further exposure of the model to additional data. Both types of uncertainty influence the confidence associated to a model's output and several different approaches exist to estimate aleatoric and epistemic uncertainty. However, to the best of our knowledge, none of these have been applied to DNN-based CIE.[11]

In this study, we aimed to systematically investigate the feasibility and performance of multiple uncertainty estimation methods for deep learning-based ECG analysis across different local and publicly available datasets and tasks. Additionally, we show which methods are the most useful to improve the clinical value of these algorithms through a clinical simulation.

Methods

Training data acquisition

Three 12-lead ECGs datasets were used to evaluate the uncertainty estimation methods. The UMCU-Triage and UMCU-Diagnose datasets were used to compare methods between an easier (UMCU-Diagnose) and challenging (UMCU-Triage) task. The publicly available CPSC2018 dataset was employed to increase the reproducibility of our experiments, and to compare results between a small (CPSC2018) and large (UMCU-Diagnose) dataset. The UMCU-Triage and UMCU-Diagnose datasets contain standard 12-lead ECGs acquired between January 2000 and August 2019 on all non-cardiology wards and outpatient clinics, the Intensive Care Unit and the Emergency Department of the University Medical Center Utrecht (UMCU, Utrecht, the Netherlands). The ECGs were acquired using a General Electric MAC 5500 (GE Healthcare, Chicago, IL, United States) and raw 10 second 12-lead ECG data waveforms were utilized. Extracted data were de-identified in accordance with the EU General Data Protection Regulation and written informed consent was waived by the ethical committee. All ECGs were interpreted by a cardiologist or cardiologist-in-training as part of the regular clinical workflow, and structured diagnosis labels were extracted from free-text interpretations using a text-mining algorithm described previously.[3] The CPSC2018 dataset was described in detail elsewhere and contains 12-lead ECGs acquired at 11 different hospitals across China.[12]

Training data labelling

The UMCU-Triage DNN performs a comprehensive ECG triage task and classifies ECGs into one of four distinct triage categories based on how promptly a cardiologist must be consulted: normal (no consultation necessary), abnormal not acute (low priority consultation), abnormal subacute (moderate priority consultation) and abnormal acute (high priority consultation). The ECG diagnoses and their corresponding triage categories were described before.[3] The CPSC2018 and UMCU-Diagnose datasets were used for a specific ECG diagnosis classification task and were annotated with 8 ECG diagnoses: normal, atrial fibrillation (AF), left bundle branch block (LBBB), right bundle branch block (RBBB), premature atrial contraction (PAC), premature ventricular contraction (PVC), ST-segment depression (STD) and ST-segment elevation (STE).

Validation data acquisition

The UMCU-Triage dataset was split into training and validation sets in a 95:5% ratio at the individual patient-level. The independent test set consisted of 984 randomly sampled ECGs from different patients, annotated by a panel of 5 practicing senior electrophysiologist-cardiologists.[3] All ECGs were interpreted by two blinded annotators, and, in case of disagreement, a third annotator was consulted. A majority vote policy was used to get the final triage class. All patients in the test set were excluded from the training and validation datasets. The UMCU-Diagnose dataset was trained and tested using a random train/validation/test split of 90:5:5% on the patient-level. The CPSC2018 data was divided according to a 90:10% train/validation split, and testing was performed with the official CPSC2018 test data which contains 300 ECGs.[11]

Deep neural network architecture

The base DNN architecture used in all experiments was based on an Inception Residual Network, which was described before by Van de Leur et al.[3] This model consists of 37 dilated single-dimensional convolutional layers, which convolve along the time-axis of the ECG (**Supplementary Figure 2**). The models were trained using the Adam optimizer with a learning rate of 0.0005.[13] Training was performed for 20 epochs, using mini-batches of size 128. To counteract class-imbalance in the data, the focal loss was used as the loss function with focusing parameter set to .[14] Complementary architecture details are provided in the Supplementary Methods.

Uncertainty estimation

Four methods for epistemic uncertainty, two methods for aleatoric uncertainty and their possible combinations were compared. The epistemic methods compared were: Monte Carlo dropout (MCD), Bayesian neural network with variational inference (VI), ensemble (ENS) and snapshot ensemble (SSE).[15–19] The aleatoric methods compared were: auxiliary output (AUX) and Bayesian decomposition (BD).[10,20,21] The estimation of epistemic uncertainty in all methods works in a similar way: [1] multiple predictions are obtained for a single ECG by training multiple networks (ENS and SSE) or by sampling from the same network (VI and MCD), [2] the class with the highest mean probability is selected and [3] the variance over the probabilities for that class is used as the measure for uncertainty. Aleatoric uncertainty is either modelled directly using an auxiliary output (either independently or combined with ENS, SSE and MCD) or Bayesian decomposition of the output of a Bayesian network (VI). Thus, for all methods, we get a new probability measure (the mean of the probabilities), referred to as the *confidence*, and an extra measure of uncertainty (the variance over the probabilities). An overview of the methods is given in **Table 1** and **Figure 1**, and the methods and implementation details are described extensively in the Supplementary Methods and **Supplementary Table 1**.

Next to regular evaluation on data the algorithm was trained on, the uncertainty methods were also evaluated for their ability to detect out-of-distribution (OOD) data, i.e. ECGs containing diagnoses that the network has never been seen before. This could happen when the algorithm is applied in a new setting with a different disease distribution than in the training dataset. OOD data was created by excluding ECGs of a specific class during training and adding those ECGs to the test set. The OOD classes were ECGs with acute arrhythmias (such as ventricular tachycardia) for UMCU-Triage (part of the abnormal acute class) dataset, and atrial fibrillation for UMCU-Diagnose and CPSC2018 datasets.

Clinical simulation

A clinical scenario was simulated where a DNN is applied in a clinical setting with different thresholds (**Figure 2**). In this simulation, every ECG is first classified by the DNN and the corresponding uncertainty estimate is obtained. Next, the results were split into a trusted and rejected group by applying a threshold based on the estimated uncertainty. This ensures that only ECGs with certain predictions are trusted, and uncertain ECGs can be then evaluated by a cardiologist. Performance of the trusted predictions was evaluated using the accuracy for every threshold. For the OOD setting, the influence of the threshold on the rate of rejection of the OOD class was visualized. The clinical simulation was performed using the same test sets as the other experiments.

Correspondence with cardiologist’s lack of agreement

We investigated whether predictions regarding ECGs which the uncertainty estimation methods marked as uncertain, corresponded with the ECGs on which cardiologist’s diagnoses differed. A unique opportunity to perform this evaluation was found in the UMCU-Triage test set since it contains annotations from multiple cardiologists. The agreement between the cardiologists was used as a proxy for their diagnostic certainty, which was then compared to the total estimated uncertainty of the DNN on the same ECGs.

Statistical analysis

For each base network, discriminatory performance was evaluated using the macro-averaged one-vs-one area under the receiver operating characteristic curve (AUC). Base network calibration was assessed using calibration plots. The uncertainty estimation methods were evaluated based on ranking, calibration and robustness against OOD data, followed by a clinical simulation in which uncertain predictions were excluded. The evaluation metrics are described below.

Ranking is concerned with the ordering of uncertainties and evaluates whether high certainty predictions align with high accuracy. Ranking was measured using the Area Under the Confidence-Oracle error (AUCO) metric (also referred to as Area Under the Sparsification Error curve).[22,23] The AUCO compares the theoretical best possible ordering based on the obtained Brier score to the ordering based on the estimated uncertainty, which are called oracle-error and confidence-error, respectively. The AUCO is then the area between the oracle-error and confidence-error curves, which measures the difference between the perfect ordering and the ordering made by the uncertainty estimation method.

In contrast to ranking, calibration looks at the actual value of the estimated confidence individually, and tests whether the estimates are over- or underconfident. To measure calibration, a calibration plot was created by splitting the mean maximum Softmax probabilities into ten bins, and calculating the accuracy over each bin. A perfectly calibrated model outputs probabilities that match up with the accuracy and would therefore lie on the diagonal. Probabilities above or below the diagonal are referred to as overconfident or underconfident, respectively. Calibration was measured using the Expected Calibration Error (ECE), which quantifies the difference on the calibration plot between the model's confidence and the perfect diagonal.[24]

The difference in the estimated uncertainty between the ECG where the cardiologists agreed and disagreed was assessed using the median and interquartile range (IQR) and Mann Whitney U test, as the data was not normally distributed. These were evaluated for the total UMCU-Triage test set and in a per-class fashion. A p-value below 0.05 was considered statistically significant.

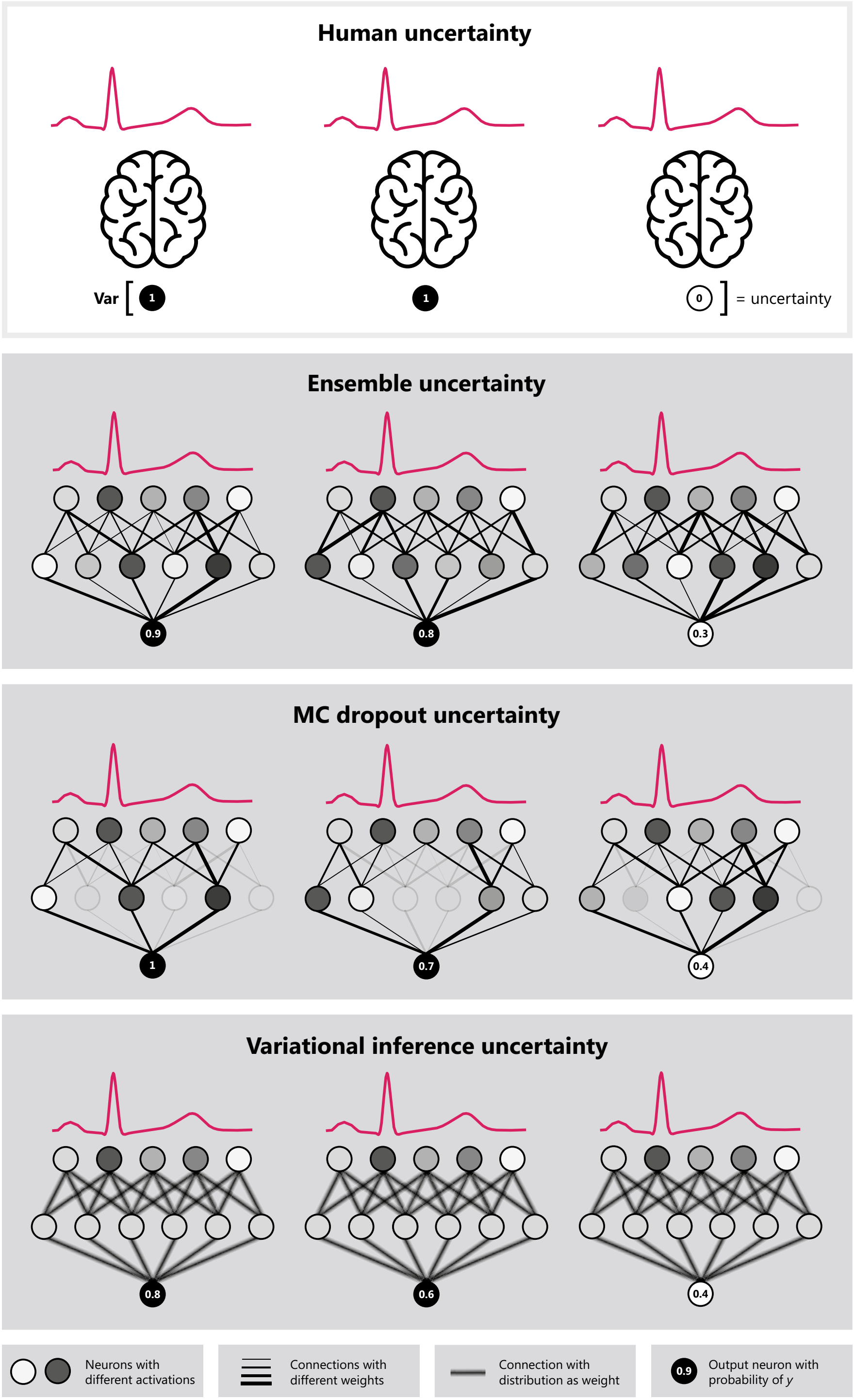

Human uncertainty
Var
1
1
0
= uncertainty
Ensemble uncertainty
0.9
0.8
0.3
MC dropout uncertainty
1
0.7
0.4
Variational inference uncertainty
0.8
0.6
0.4
Neurons with different activations
Connections with different weights
Connection with distribution as weight
0.9
Output neuron with probability of y

Figure 1. *Overview of the uncertainty estimation concept and the epistemic uncertainty estimation methods. All methods work similarly to human uncertainty (in the top box, illustrated as several brains), where there are multiple reviewers interpreting the same ECG.* The uncertainty is then calculated as the variance over these different predictions for the same ECG. With DNNs multiple predictions can be achieved using ensembles (i.e. training the same network multiple times), MC dropout (i.e. removing some nodes randomly during prediction) or variational inference (i.e. sampling from the same network with distributions as weights multiple times).

METHOD		DESCRIPTION
Monte Carlo dropout (MCD)	EPISTEMIC	Dropout is kept on during test time, thereby creating a different dropout mask of the network every time a prediction is made. Through making multiple predictions on the same input ECG with differing dropout masks, varying predictions are obtained. The variance within these predictions is the estimated epistemic uncertainty. **Simple to implement and can be applied to all existing models without retraining given that dropout was used.**
Variational Inference (VI)		Weights of the neural network are replaced by distributions, creating a Bayesian neural network. These distributions can be sampled to obtain a set of weights, which can be used to make predictions. Once trained, the distributions are sampled multiple times to obtain multiple sets of weights, which are used to make multiple predictions on the same input ECG. The variance within these predictions functions as the estimated epistemic uncertainty. **Theoretically sound approach to uncertainty but requires adjustment of network and training logic and training can be difficult and time-intensive.**
Ensemble (ENS)		Multiple the same neural networks are randomly initialized and trained on the same data, resulting in an ensemble of neural networks. After training, each ensemble member predicts on the same input ECG. The predictions are averaged, and the variance within the predictions is the estimated epistemic uncertainty. **Simple to implement and can be applied to all existing models but training logic needs slight changes and training demands more time.**
Snapshot ensemble (SSE)	EPISTEMIC	Dropout is kept on during test time, thereby creating a different dropout mask of the network every time a prediction is made. Through making multiple predictions on the same input ECG with differing dropout masks, varying predictions are obtained. The variance within these predictions is the estimated epistemic uncertainty. **Simple to implement and can be applied to all existing models without retraining given that dropout was used.**
Auxiliary output (AUX)	ALEATORIC	The auxiliary output method adds an additional output neuron to the last layer of the neural network for each class. These neurons are tasked with estimating the aleatoric uncertainty. The neurons are incorporated into the loss function during training, and thereby directly learn the aleatoric uncertainty present in the data. Once trained, the value of the auxiliary output neuron corresponding to the predicted class is the estimated aleatoric uncertainty. **Possibility to add aleatoric uncertainty estimation to non-Bayesian networks. Simple to implement, requires changing the last layer of the architecture.**
Bayesian decomposition (BD)		The Bayesian decomposition method works with the variational inference method. It decomposes the predictive distribution of a Bayesian neural network into an epistemic and aleatoric part directly. **Possibility to add aleatoric uncertainty estimation to Bayesian networks. Simple to implement when the network is already Bayesian.**

Table 1. *Description of evaluated uncertainty estimation methods.*

Results

Data distribution

The UMCU-Triage and -Diagnose datasets contained 316.987 and 194.880 ECGs, respectively, while the CPSC2018 dataset contained 6.877 ECGs. The class distribution in the different datasets is shown in **Table 2**. The UMCU-Triage test set consisted of 984 ECGs of unique patients of which 418 were normal, 410 abnormal not acute, 80 abnormal subacute and 76 abnormal acute. The UMCU-Diagnose and CPSC2018 test set consisted of 10.089 and 300 ECGs respectively, with similar distribution to **Table 2**.

Base network comparison

The mean AUCs of the base DNN and models with uncertainty estimation methods on in-distribution setting were 0.95 ± 0.0044 for the UMCU-Triage dataset, 0.99 ± 0.0016 for the UMCU-Diagnose dataset and 0.92 ± 0.0159 for CPSC2018 dataset. This shows that the models have similar performance and can therefore be compared fairly. In **Figure 3**, the calibration of the base network on all datasets is shown. The base network's probability was up to 15% underconfident on the UMCU-Triage and UMCU-Diagnose datasets in the in-distribution setting and up to 30% overconfident on both the in- and OOD setting for the CPSC2018 dataset.

Ranking

The VI model obtained the best ranking score among the models with a single uncertainty estimation method on the in-distribution setting of UMCU-Triage (**Table 3**). When combined with BD, ranking improved significantly, and VI+BD obtained the best ranking scores on both in- and OOD setting. The best performing uncertainty estimation methods for UMCU-Diagnose were VI, ENS, VI+BD and ENS+AUX for the in-distribution, and MCD for the OOD setting. For CPSC2018 the ENS model obtained the lowest AUCO on in-distribution setting, and VI+BD on the OOD setting. When comparing between in and OOD setting, the AUCO for OOD data was generally higher than in in-distribution setting. In **Table 3,** all AUCO scores are displayed. The ranking plots for all datasets are displayed in **Supplementary Figures 3-5**.

Calibration

The ECEs for all uncertainty estimation methods were lower than the base network, with the auxiliary output method on the CPSC2018 dataset being the only exception (**Table 4**). On UMCU-Triage, the best calibrated method was the SSE+AUX for both in-distribution and OOD setting. For UMCU-Diagnose, the lowest ECEs were obtained by the VI, AUX and VI+BD methods on in-distribution setting, and MCD+AUX, ENS+AUX and SSE+AUX on OOD setting. For the CPSC2018 dataset, SSE, ENS+AUX and SSE+AUX were the most calibrated methods on in-distribution, and the ENS+AUX model obtained the lowest ECE on OOD setting. **Table 4** shows the calibration results and calibration plots for all methods and datasets are shown in **Supplementary Figures 6-8.**

Clinical simulation

The clinical simulation uncertainty threshold plot for the UMCU-Triage dataset in the in-distribution setting is displayed in **Figure 4**. The results show that exclusion of uncertain ECGs improves the accuracy of all models. The VI+BD model had the steepest upward slope, and thus excluded the uncertain ECGs the fastest, thereby increasing overall model accuracy at the highest rate. Within **Table 5**, the accuracies of the models with uncertainty thresholds applied at 25%, 50% and 75% are displayed for the in-distribution setting, and in **Table 6** for the OOD setting. The accuracy of all models increased when estimated uncertain samples were removed.

In **Figure 5**, the normalized per-class thresholding plots for the VI+BD and ENS+AUX models on the UMCU-Diagnose dataset are shown. The ECGs containing atrial fibrillation are of average uncertainty in the in-distribution setting, but in the OOD setting where the models have never seen atrial fibrillation before, the ECGs with atrial fibrillation are marked with high uncertainty, and thereby removed at the fastest rate. Plots for the other datasets are shown in **Supplementary Figures 9 and 10**.

Correspondence with cardiologist’s lack of agreement

The cardiologists showed moderate agreement on the triage class in the UMCU-Triage expert test set and agreed on 736 of the 984 ECGs (75%, Cohen’s kappa 0.60, $p < 0.001$). The highest agreement was observed in the normal class (77%) and the lowest in the abnormal acute class (61%). The total certainty was lower for ECGs in which cardiologists’ annotations did not agree (median 39%, IQR 43%) as compared to ECGs in which cardiologists did agree (median 55%, IQR 50%, overall $p < 0.001$). The certainty was the highest for the normal class (median 73%, IQR 40%) and the lowest for the abnormal acute class (median 22%, IQR 26%). The consensus of the panel of cardiologists is plotted against the median total uncertainty per class for the VI+BD method in **Figure 6**.

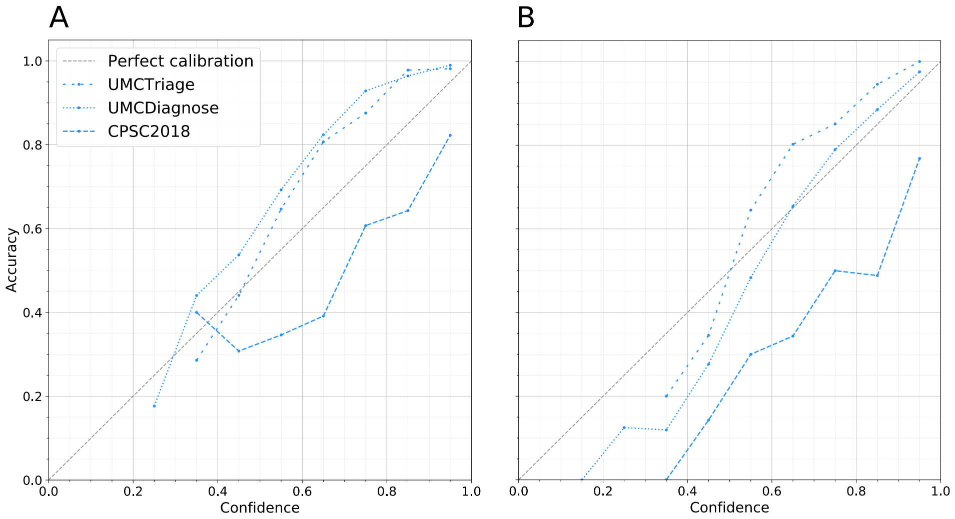


Figure 4.
Clinical simulation with accuracies of predictions as a function of excluding uncertain ECGs on the in-distribution setting of the UMCU-Triage dataset.
The threshold percentage corresponds to the percentage of data that needs to be evaluated by a physician after exclusion. Through excluding uncertain ECGs, accuracy of all models improved. The VI+BD model had the steepest upward slope, and thus excluded the uncertain ECGs the fastest, thereby increasing overall model accuracy at the highest rate. MCD: Monte-Carlo Dropout, VI: Variational Inference, ENS: Ensemble, SSE: Snapshot Ensemble, AUX: Auxiliary output, BD: Bayesian decomposition.

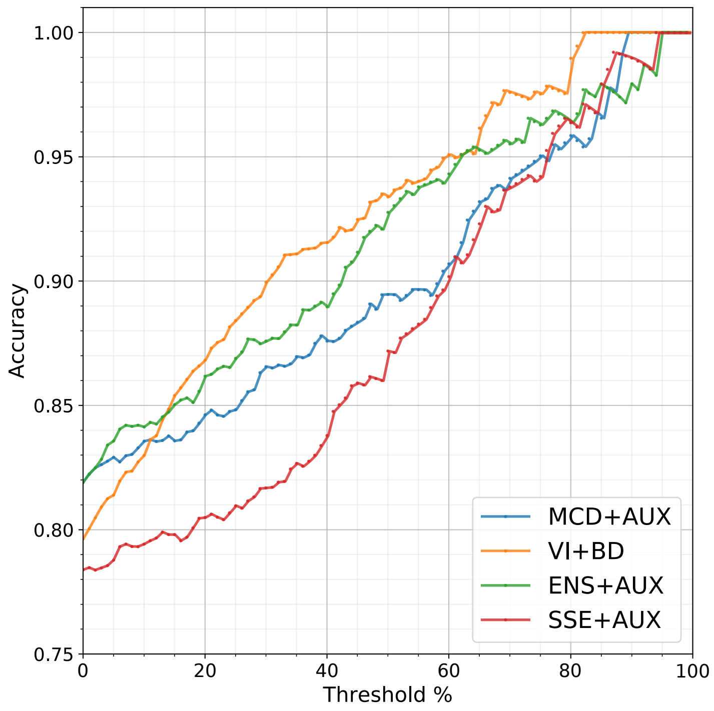


Figure 3.
Calibration of the base network in the in-distribution (A) and out-of-distribution (B) setting for all datasets.
In a calibration plot the predicted probability or confidence of the network is grouped into ten bins from low (i.e. 20-30%) to high (i.e. 90-100%). For all these bins the accuracy in that bins is calculated. A perfectly calibrated model outputs confidences that match up exactly with the accuracy. A model which predicts higher probabilities than the accuracy is overconfident, which can be observed by a line that falls under the diagonal. An underconfident model is the opposite and lies above the diagonal. The base models without uncertainty estimation are op to 30% over- or underconfident.

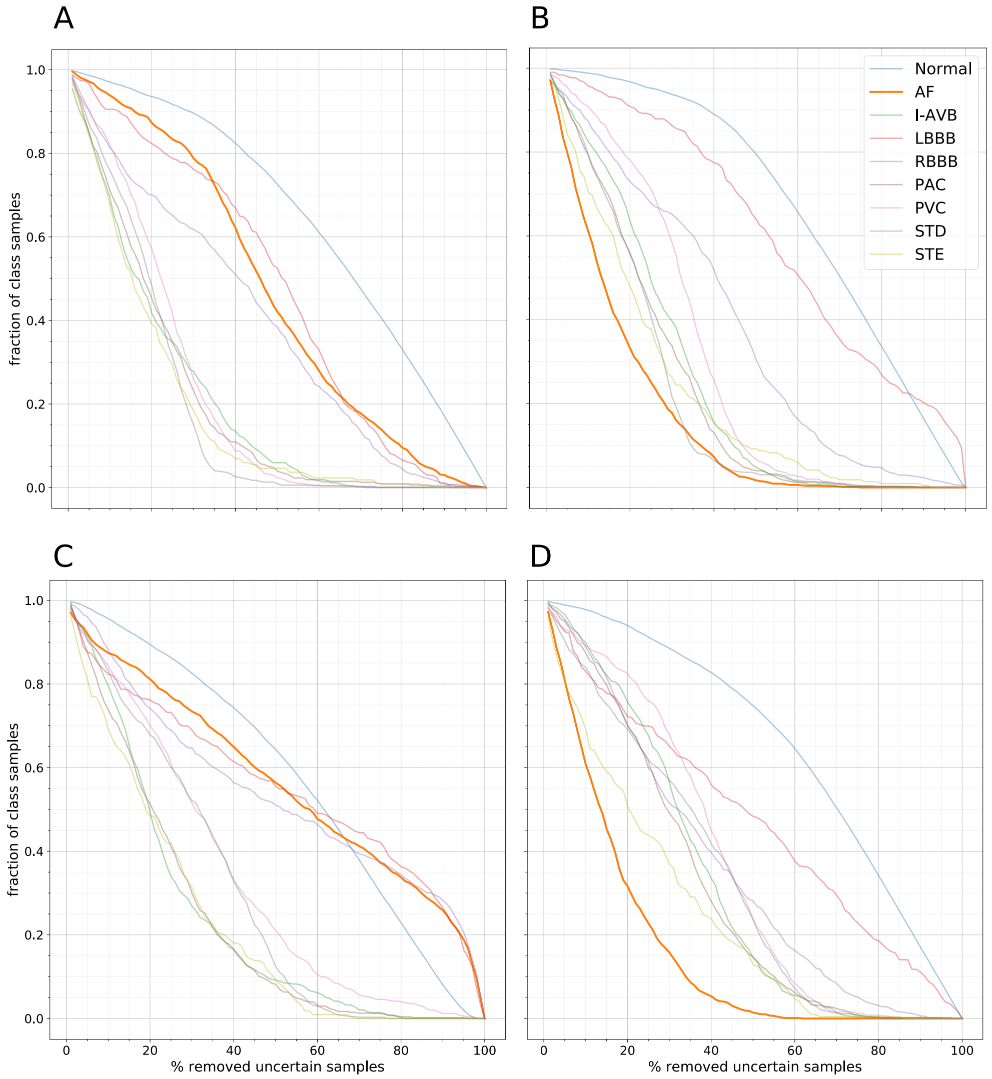


Figure 5.
***Normalized per-class thresholding plots of the VI+BD (A, B) and ENS+AUX (C, D) models on the UMCU-Diagnose dataset*.**
The first column (A, C) is for the in-distribution setting, the second column (B, D) is for the out-of-distribution setting. Classes with high uncertainty are removed first and have a steep downward slope. In the in-distribution plots (A, C), the model was trained on all classes, including atrial fibrillation (AF). These plots show that the algorithm is certain about prediction atrial fibrillation, as these samples are excluded slower than other classes. In the out-of-distribution plots (B, D), the algorithm was trained on a dataset that contained no atrial fibrillation ECGs. These plots show that the model is now very uncertainty about this unseen class, as it excludes the atrial fibrillation ECGs first. AF: atrial fibrillation, I-ABV: first degree atrioventricular block, LBBB: left bundle branch block, RBBB: right bundle branch block, PAC: premature atrial contraction, PVC: premature atrial contraction, STD: ST-depression, STE: ST-elevation.

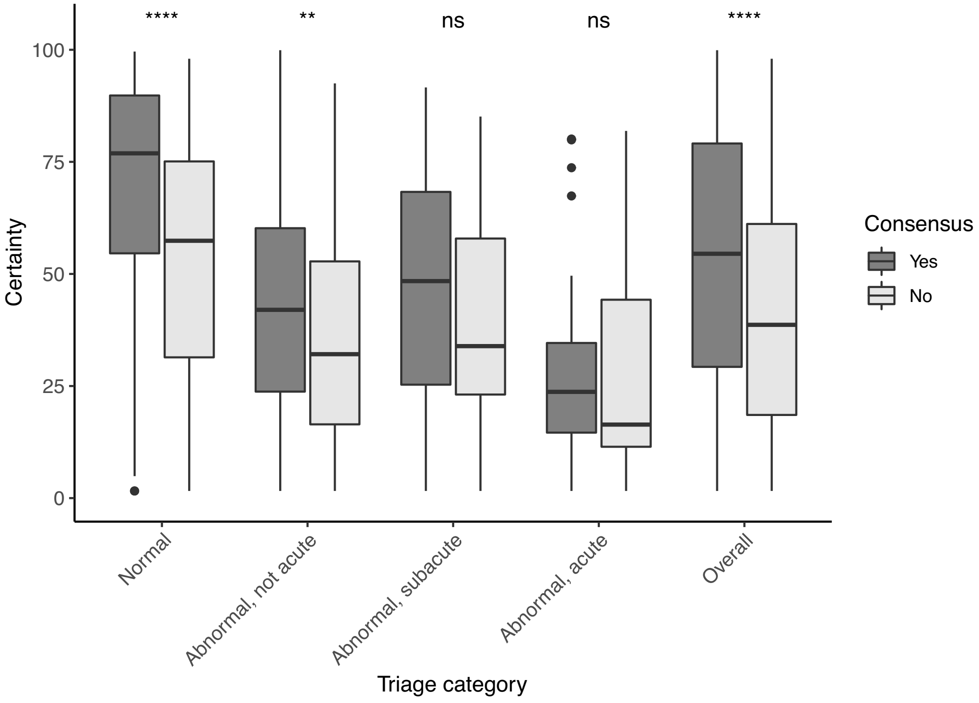


Figure 6.
Correspondence of uncertainty with cardiologist's lack of agreement.
The ECGs in the expert test set of UMCU-Triage are grouped by consensus between two cardiologists and compared with the estimated uncertainty for these ECGs, both per-class and overall. The algorithm is more certain about ECGs where the cardiologists agreed. Moreover, the algorithm is most certain about the normal ECGs and least certain about the abnormal, acute ECGs, which is also the smallest class. ** $p < 0.01$, **** $p < 0.0001$, ns: not significant.

	CLASS	#	%
UMCU TRIAGE	**Normal**	138774	43.78
	Abnormal, not acute	139656	44.06
	Abnormal, subacute	23113	7.29
	Abnormal, acute	15444	4.87
	Total	**316987**	
UMCU-DIAGNOSE	**Normal**	109787	56.35
	Atrial fibrillation	20073	10.30
	First-degree atrioventricular block	8411	4.32
	Left bundle branch block	6290	3.23
	Right bundle branch block	13568	6.96
	Premature atrial contraction	9258	4.75
	Premature ventricular contraction	9580	4.75
	ST-segment depression	13538	6.85
	ST-segment elevation	4375	2.24
	Total	**194880**	

	CLASS	#	%
CPSC2018	**Normal**	918	13.35
	Atrial fibrillation	1098	15.97
	First-degree atrioventricular block	704	10.24
	Left bundle branch block	207	3.01
	Right bundle branch block	1695	24.65
	Premature atrial contraction	556	8.08
	Premature ventricular contraction	672	9.77
	ST-segment depression	825	12.00
	ST-segment elevation	202	2.94
	Total	**6877**	

Table 2.
Overview of dataset characteristics.

METHOD	UNCERTAINTY TYPE	UMCU-TRIAGE		UMCU-DIAGNOSE		CPSC 2018	
		in-dist.	**OOD**	**in-dist.**	**OOD**	**in-dist.**	**OOD**
None	-	0.05	0.07	0.02	0.04	0.21	0.28
MCD	epistemic	0.11	0.15	0.03	0.03	0.15	0.20
VI	epistemic	0.08	0.10	0.02	0.04	0.20	0.17
ENS	epistemic	0.11	0.10	0.02	0.04	0.14	0.22
SSE	epistemic	0.11	0.10	0.03	0.04	0.24	0.28
AUX	aleatoric	0.10	0.09	0.07	0.08	0.18	0.23
MCD + AUX	total	0.09	0.12	0.03	0.04	0.18	0.31
VI + BD	total	0.06	0.07	0.02	0.04	0.20	0.15
ENS + AUX	total	0.08	0.10	0.02	0.04	0.16	0.26
SSE + AUX	total	0.12	0.10	0.05	0.04	0.26	0.42

Table 3. ***Ranking performance measured using area under the confidence-oracle error (AUCO).*** The AUCO of individual epistemic uncertainty estimation methods is improved when combined with a method for estimating aleatoric uncertainty. OOD: out-of-distribution. MCD: Monte-Carlo Dropout, VI: Variational Inference, ENS: Ensemble, SSE: Snapshot Ensemble, AUX: Auxiliary output, BD: Bayesian decomposition.

METHOD	**UNCERTAINTY TYPE**	UMCU-TRIAGE		UMCU-DIAGNOSE		CPSC 2018	
		in-dist.	**OOD**	**in-dist.**	**OOD**	**in-dist.**	**OOD**
None	-	0.11	0.10	0.09	0.03	0.17	0.25
MCD	epistemic	0.08	0.05	0.06	0.04	0.07	0.12
VI	epistemic	0.07	0.08	0.02	0.06	0.09	0.11
ENS	epistemic	0.06	0.07	0.03	0.04	0.09	0.18
SSE	epistemic	0.05	0.04	0.04	0.04	0.06	0.09
AUX	aleatoric	0.04	0.05	0.02	0.05	0.18	0.26
MCD + AUX	total	0.07	0.07	0.03	0.04	0.14	0.24
VI + BD	total	0.07	0.08	0.02	0.06	0.09	0.11
ENS + AUX	total	0.08	0.07	0.04	0.04	0.06	0.06
SSE + AUX	total	0.03	0.03	0.04	0.04	0.06	0.07

Table 4.
Calibration measured in Expected Calibration Error.
The acquired Expected Calibration Errors are lower for models with uncertainty estimation compared to the base model without uncertainty estimation. MCD: Monte-Carlo Dropout, VI: Variational Inference, ENS: Ensemble, SSE: Snapshot Ensemble, AUX: Auxiliary output, BD: Bayesian decomposition.

METHOD	UMCU-TRIAGE				UMCU-DIAGNOSE				CPSC 2018			
	0%	25%	50%	75%	0.92	25%	50%	75%	0%	25%	50%	75%
MCD	0.82	0.86	0.88	0.90	0.91	0.96	0.98	1.00	0.69	0.76	0.88	0.93
VI	0.80	0.85	0.91	0.98	0.92	0.97	0.99	1.00	0.70	0.74	0.77	0.87
ENS	0.81	0.85	0.88	0.90	0.91	0.97	0.99	1.00	0.73	0.82	0.91	0.93
SSE	0.81	0.83	0.88	0.93	0.92	0.96	0.99	1.00	0.66	0.65	0.71	0.73
AUX	0.81	0.84	0.88	0.96	0.92	0.95	0.95	0.95	0.71	0.75	0.87	0.93
MCD + **AUX**	0.82	0.85	0.89	0.95	0.92	0.96	0.99	1.00	0.72	0.79	0.85	0.93
VI + **BD**	0.80	0.88	0.93	0.98	0.91	0.98	0.99	1.00	0.70	0.73	0.79	0.89
ENS + **AUX**	0.82	0.87	0.93	0.96	0.92	0.98	0.99	1.00	0.73	0.77	0.83	0.95
SSE + **AUX**	0.78	0.81	0.87	0.94	0.91	0.95	0.97	0.99	0.66	0.67	0.75	0.81

Table 5.
Accuracy scores for non-thresholded (0%) and thresholded (25%, 50%, 75%) predictions on all datasets on in-distribution setting. Predictions are thresholded by removing 25%, 50% and 75% of the estimated most uncertain samples. Model accuracy increases for all methods and dataset when uncertain samples are removed. MCD: Monte-Carlo Dropout, VI: Variational Inference, ENS: Ensemble, SSE: Snapshot Ensemble, AUX: Auxiliary output, BD: Bayesian decomposition.

METHOD	UMCU-TRIAGE				UMCU-DIAGNOSE				CPSC 2018			
	0%	25%	50%	75%	0%	25%	50%	75%	0%	25%	50%	75%
MCD	0.80	0.82	0.83	0.87	0.82	0.94	0.99	1.00	0.59	0.66	0.80	0.88
VI	0.81	0.86	0.90	0.93	0.81	0.93	0.98	1.00	0.61	0.70	0.79	0.93
ENS	0.82	0.87	0.89	0.93	0.83	0.92	0.98	1.00	0.62	0.69	0.83	0.88
SSE	0.80	0.85	0.90	0.93	0.82	0.92	0.98	1.00	0.56	0.56	0.61	0.77
AUX	0.80	0.85	0.91	0.96	0.82	0.92	0.95	0.96	0.60	0.68	0.78	0.85
MCD + **AUX**	0.80	0.84	0.89	0.92	0.82	0.94	0.99	0.99	0.57	0.63	0.71	0.81
VI + **BD**	0.81	0.88	0.93	0.97	0.81	0.93	0.99	1.00	0.61	0.72	0.84	0.93
ENS + **AUX**	0.81	0.88	0.90	0.93	0.83	0.93	0.99	1.00	0.61	0.60	0.62	0.84
SSE + **AUX**	0.79	0.83	0.90	0.95	0.82	0.93	0.98	0.99	0.58	0.57	0.57	0.53

Table 6.
Accuracy scores for non-thresholded (0%) and thresholded (25%, 50%, 75%) predictions on all datasets on out-of-distribution setting. Predictions are thresholded by removing 25%, 50% and 75% of the estimated most uncertain samples. Model accuracy increases for all methods and dataset when uncertain samples are removed, except for the model with SSE+AUX. MCD: Monte-Carlo Dropout, VI: Variational Inference, ENS: Ensemble, SSE: Snapshot Ensemble, AUX: Auxiliary output, BD: Bayesian decomposition.

Discussion

This study is the first to systematically investigate the feasibility and performance of uncertainty estimation methods for the automated classification of ECGs using DNNs. Our calibration results documented that the regular DNN is up to 30% either over- or underconfident, stressing the need for adequate uncertainty estimation (**Figure 3**). We demonstrated how implementing uncertainty estimation improves both calibration and ranking across datasets with differing sizes and tasks. The proposed methods therefore provide an improved and better calibrated probability measure together with an additional uncertainty measure. While pressure testing this new uncertainty measure in a safe and insightful clinical simulation, we showed that by thresholding the uncertainty estimates and thereby rejecting uncertain ECGs markedly improves accuracy in the remaining data. Furthermore, out-of-distribution ECG diagnoses that the algorithm has not seen during training are rejected faster. Finally, these uncertainties were shown to significantly correlate with the disagreement that exists between cardiologists in clinical ECG interpretation.

When implementing new technologies into clinical practice, knowing its limitations is of the utmost importance, especially if the technology concerns ‘black-box’ algorithms such as DNNs. Surprisingly, while there is a rapid growth of publications on DNNs that perform ECG analyses, we found none that included uncertainty estimations. When training a DNN for a specific task such as ECG interpretation, the algorithm is constitutionally forced to accept every input and assign it to an output, even in the cases where the algorithm’s estimations carry great uncertainty. The results from the DNNs without uncertainty estimation in this study showed that the network was underconfident on the large UMCU-Triage and UMCU-Diagnose dataset, while overconfident on the relatively small CPSC2018 dataset. Such discrepancies, if left unchecked, could potentially lead to unfavourable or potentially dangerous situations when applied in a clinical setting where a patient could be wrongly diagnosed by a DNN prediction with high Softmax probability. These findings demonstrate that out-of-the-box DNN predictions should not be blindly trusted without estimating their prediction uncertainty. In our opinion, incorporating the estimation of the uncertainty of DNN predictions is therefore an essential prerequisite when applying an algorithm into clinical practice.

Uncertainty estimation techniques

The variety of estimation methods employed (and their combinations) allowed us to extensively investigate their comparative performance. In the ranking results for the UMCU-Triage dataset, we demonstrated that when comparing models with only epistemic uncertainty estimation methods to models with both epistemic and aleatoric uncertainty estimation, the ranking improves for the latter. Therefore, it seems that aleatoric uncertainty is beneficial to the ranking score on a large dataset such as UMCU-Triage. This is in line with earlier work stating that aleatoric uncertainty is more important for large datasets because all the epistemic uncertainty has been taken away through providing the network with enough training data.[10] It is therefore important to model aleatoric uncertainty when dealing with large datasets. Regarding the calibration results, we found that the ECEs for all networks with uncertainty estimation methods were lower than the ECEs of the baseline network. Thus, calibration performance improved noticeably in all the networks that employed uncertainty estimation. These findings clearly demonstrate the benefits of modelling uncertainty for the calibration of a DNN. When comparing the calibration scores on the small CPSC2018 dataset, we observed that the ECE of the AUX model was the largest out of all models. The only model with uncertainty estimation that performed worse than the base network was thus a model that only modelled aleatoric uncertainty, whereas all other models that have epistemic uncertainty estimation improved upon the base network. This strongly suggests that it may be more important to model epistemic uncertainty for small datasets because there is still much epistemic uncertainty present after training, which is confirmed in earlier work.[10]

Therefore, through our experiments, we found that epistemic uncertainty should be modelled for small datasets and aleatoric uncertainty for large datasets. Preferably however, both should be modelled, which is why we only consider models that estimated both types of uncertainty. From these models, the MCD+AUX model displayed large overconfidence on the OOD setting of CPSC2018 (as shown in **Table 4**) and is therefore not recommended. The SSE+AUX model's showed poor ranking in all datasets in both the in-distribution and OOD datasets and this model is therefore also not recommended. Overall, the VI+BD and ENS+AUX models performed best for improving ranking and calibration across datasets and tasks in both the in-distribution and OOD setting and are therefore recommended as a starting point in similar ECG diagnosis settings. However, further research is needed to confirm the generalizability of our results in other settings.

Our findings for the ENS method align with recent research where this method also performed best out of the tested uncertainty estimation methods. However, the results for the VI method differ with these studies, that found VI to perform best on small datasets, but was outperformed by other methods on the large ImageNet dataset.[11,25] We believe the difference in outcomes is due to the fact that all our datasets are an order of magnitude smaller than ImageNet, and we therefore do not observe the same effect. Finally, one study recommended the MCD method, however they did not perform testing on OOD data, which is where we found the method to be overconfident.[11]

Clinical simulation

In our analyses, accuracy of all models increased when estimated uncertain samples were removed (**Tables 5 and 6**, **Figure 4**). These findings show that the estimated uncertainty can be used as a threshold, so only certain samples are ultimately classified by an actually accurate model. Such an implementation is highly attractive in a clinical setting, so that the ECGs with high estimated uncertainty (which the network is prone to misdiagnose), can be passed on to a cardiologist for further analysis. The thresholds for when to trust the network and when to consult a cardiologist can be set according to the required accuracy for the specific task or setting. Employing a clinical workflow with such an intermediate "quality control" structure is envisioned to greatly reduce clinical workload, while maintaining or improving the quality of diagnoses.

Both the recommended VI+BD and ENS+AUX methods perform well in quickly increasing accuracy in the group with trusted classifications when the threshold increases, both in an in-distribution and OOD setting (**Tables 5 and 6, Figure 4**). For the UMCU-Diagnose dataset's in-distribution setting, an uncertainty threshold of only 25% results in a near-perfect accuracy of 98% in the trusted group (**Table 5**). For the UMCU-Triage task, which is more difficult than predicting a single diagnostic statement, we observed that 75% of the ECGs needed to be excluded to gain the same near-perfect accuracy. This indicates that the network is more uncertain about this task. The same holds for the CPSC2019 dataset, where the high uncertainty is likely due to the small sample size. The thresholded OOD results revealed that after excluding 25% of uncertain samples, most of the obtained accuracies returned to normal in-distribution levels, hinting that the bulk of the OOD data had been excluded (**Table 6**). This exhibits the possibility of excluding new or rare diseases present in the ECG which the DNN had not seen before.

After training, an uncertainty estimation method is expected to ascribe high uncertainty to predictions on the unseen OOD class. When focussing on the OOD class specifically, the per-class thresholding plots (**Figure 5**) for the UMCU-Diagnose dataset show that the VI+BD and ENS+AUX methods estimated a higher uncertainty for the OOD ECGs compared to in-distribution ECGs. This finding suggests that the uncertainty estimation methods correctly detected the OOD ECGs, by ascribing them high uncertainty. However, the uncertainty does not increase further when the OOD ECGs already belong to the most uncertain class, as observed for the UMCU-Triage dataset in **Supplementary Figure 9**. Furthermore, the OOD ECGs are not always identified as most uncertain, as is the case for the ENS+AUX method on the CPSC2018 dataset shown in **Supplementary Figure 10**. However, when comparing the obtained AUCO scores between in-distribution and OOD setting, it was also observed that the AUCOs for OOD setting are generally higher, suggesting that introducing OOD data can degrade ranking. The tested uncertainty estimation methods are therefore not completely robust against OOD data and this remains a point of improvement.

Correspondence with cardiologist's lack of agreement

Most interestingly, uncertainty was shown to significantly correlate with the lack of diagnostic agreement encountered even between experienced cardiologists when interpreting an ECG. This seems to suggest that the cardiologists and DNNs may struggle with the similar complex patterns in challenging ECGs, either due to aleatoric uncertainty caused by noise or borderline cases inherent in the data, or through epistemic uncertainty of ECGs with rare abnormalities. This notion represents a solid step towards confident clinical deployment based on the assurance that *uncertainty* estimation methods function as expected and align with cardiologists on what is most worth their restricted clinical time.

Limitations

This study has several limitations to address. First of all, the test-sets of UMCU-Triage and CPSC2018 were small, and results on these datasets are therefore prone to stochasticity. Secondly, the OOD class on the UMCU-Triage test set only constituted 1.8% of the data, which complicated the interpretation of the thresholding results. Thirdly, the ECGs in the CPSC2018 dataset are of varying length between 6 and 60 seconds. We extracted only the first 10-seconds and zero-padded ECGs which were shorter, which could potentially lead to missing features in the ECG. Fourthly, experiments were performed on a single DNN architecture, which reduces the generalizability of the results towards other DNN architectures. Residual convolutional neural networks are, however, the most commonly used in DNN-based analysis of ECGs. [26]

Clinical perspectives and future work

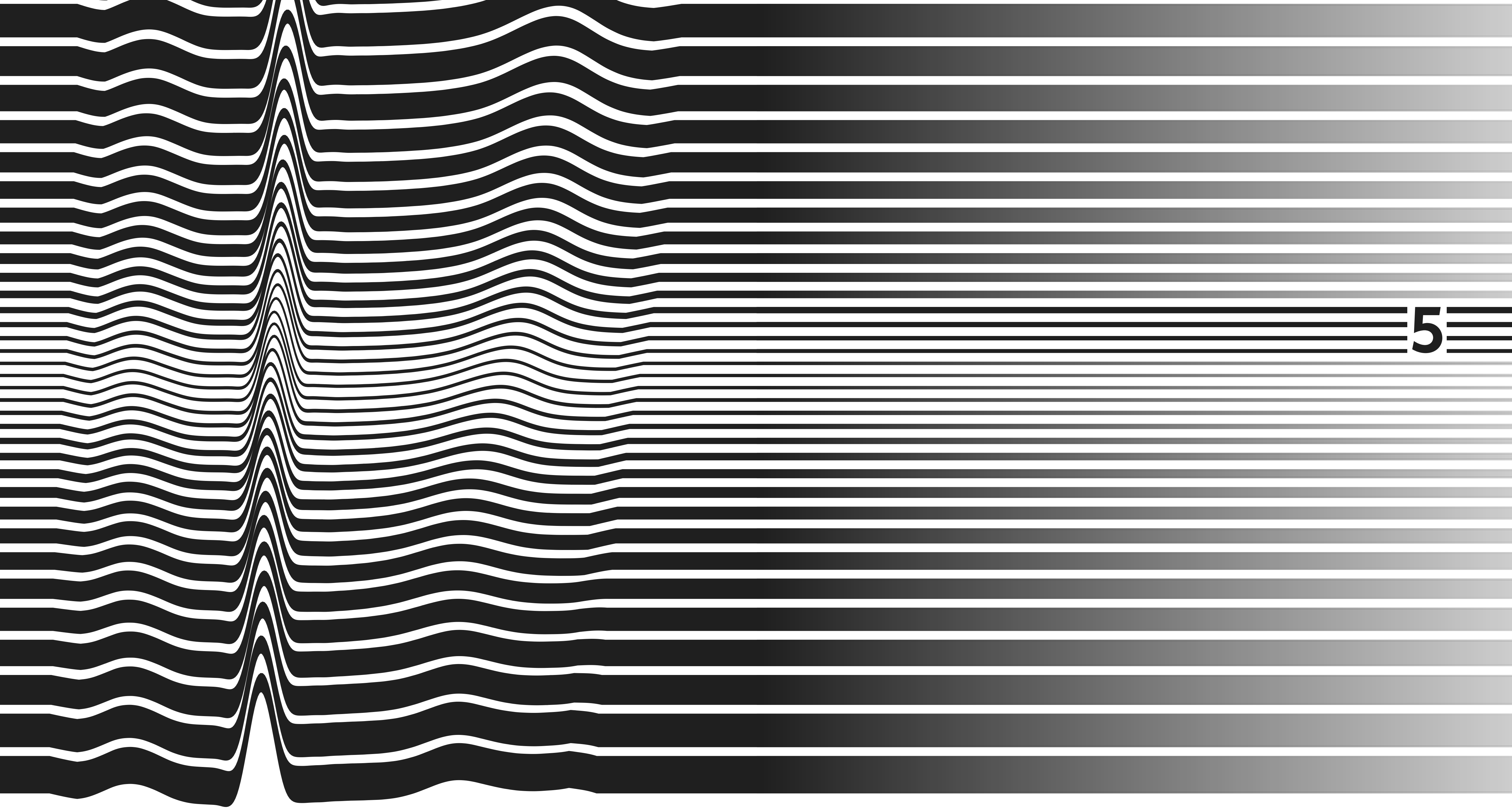
Our study demonstrated that through uncertainty estimation, we are coming one step closer to applying DNNs in a clinical setting. Firstly, our study dealt with multi-class classification, where only a single class is present in the ECG. However, in the real-world it often occurs that multiple diseases are present within the same ECG. Therefore, it might be interesting to investigate uncertainty estimation for networks that accommodate for multi-label classification too. Secondly, we observed that the average estimated uncertainty differs per class, as displayed in **Figure 6**. This allows for the setting of class-specific thresholds, because the estimated certainty for a common class lies much higher than for an uncommon class. Future studies should investigate whether novel uncertainty estimation methods could account for these different uncertainty thresholds per class, as this might be necessary in specific clinical problems. Moreover, the effect of pretraining, oversampling or data augmentation on uncertainty in imbalanced or small datasets should be investigated. Thirdly, a visualization of the estimated uncertainty could guide cardiologists into understanding why a DNN had difficulties interpreting ECGs. This could be performed using a technique such as Guided Grad-CAM.[27] Finally, the estimated uncertainties could also be used to improve DNNs, which can be achieved in two phases. Firstly, it might be used during training as a guide towards parts of the data that the DNN is uncertain about, where cleaning or additional data is necessary. Secondly, during use in clinical practice an active learning workflow is possible, where uncertain ECGs are interpreted by a cardiologist and the DNN continuously improves by learning from these ECGs.

Conclusion

In conclusion, this is the first study to apply and systematically investigate uncertainty estimation techniques on DNN-based CIE. We demonstrated the need for uncertainty estimation and showed that through its implementation, ECGs that a DNN would otherwise classify incorrectly can be excluded and passed on to a cardiologist for further review. Furthermore, we found a strong correlation between estimated uncertainty and disagreement between cardiologists. This study shows the possibility of strengthening the application of DNNs in practice through uncertainty estimation and is an important step towards the clinical applicability of automated ECG diagnosis through deep learning.

REFERENCES

1. Cook DA, Oh S-Y, Pusic MV. Accuracy of Physicians' Electrocardiogram Interpretations. Jama Intern Med 2020; 180: 1461.
2. Schläpfer J, Wellens HJ. Computer-Interpreted Electrocardiograms Benefits and Limitations. J Am Coll Cardiol 2017; 70: 1183–1192.
3. Leur RR van de, Blom LJ, Gavves E, et al. Automatic Triage of 12-Lead Electrocardiograms Using Deep Convolutional Neural Networks. J Am Heart Assoc; 9. Epub ahead of print 2020. DOI: 10.1161/jaha.119.015138.
4. Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 2019; 25: 65–69.
5. Ribeiro AH, Ribeiro MH, Paixao GMM, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat Commun 2020; 11: 1760.
6. Leur RR van de, Boonstra MJ, Bagheri A, et al. Big Data and Artificial Intelligence: Opportunities and Threats in Electrophysiology. Arrhythmia Electrophysiol Rev 2020; 9: 146–154.
7. Ghahramani Z. Probabilistic machine learning and artificial intelligence. Nature 2015; 521: 452--459.
8. Leibig C, Allken V, Ayhan MS, et al. Leveraging uncertainty information from deep neural networks for disease detection. Sci Rep-uk 2017; 7: 1--14.
9. Louizos C, Welling M. Multiplicative Normalizing Flows for Variational Bayesian Neural Networks. In: Precup, Teh D and, Whye Y (eds) Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: PMLR, 2017, pp. 2218–2227.
10. Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? Advances in Neural Information Processing Systems 2017; 2017-Decem: 5575--5585.
11. Filos A, Farquhar S, Gomez AN, et al. A Systematic Comparison of Bayesian Deep Learning Robustness in Diabetic Retinopathy Tasks, http://arxiv.org/abs/1912.10481 (2019).
12. Liu F, Liu C, Zhao L, et al. An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection. J Med Imag Health In 2018; 8: 1368--1373.
13. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. In: Bengio Y, LeCun Y (eds) 3rd International Conference on Learning Representations. San Diego, CA, USA: Conference Track Proceedings. Epub ahead of print 2015. DOI: 10.1063/1.4902458.
14. Lin TY, Goyal P, Girshick R, et al. Focal Loss for Dense Object Detection. 2017 Ieee Int Conf Comput Vis Iccv 2017; 2017-Octob: 2999--3007.
15. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. 33rd International Conference on Machine Learning, ICML 2016 2016; 3: 1651--1660.
16. Blundell C, Cornebise J, Kavukcuoglu K, et al. Weight Uncertainty in Neural Networks. 37, http://arxiv.org/abs/1505.05424 (2015).
17. Kingma DP, Salimans T, Welling M. Variational dropout and the local reparameterization trick. Advances in Neural Information Processing Systems 2015; 2015-Janua: 2575--2583.
18. Huang G, Li Y, Pleiss G, et al. Snapshot ensembles: Train 1, get M for free. 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings 2019; 1--14.
19. Beluch WH, Genewein T, Nurnberger A, et al. The Power of Ensembles for Active Learning in Image Classification. 2018 Ieee Cvf Conf Comput Vis Pattern Recognit 2018; 9368--9377.
20. Depeweg S, Hernandez-Lobato JM, Doshi-Velez F, et al. Uncertainty Decomposition in Bayesian Neural Networks with Latent Variables.
21. Kwon Y, Won JH, Kim BJ, et al. Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation. Comput Stat Data An; 142. Epub ahead of print 2020. DOI: 10.1016/j.csda.2019.106816.
22. Scalia G, Grambow CA, Pernici B, et al. Evaluating Scalable Uncertainty Estimation Methods for DNN-Based Molecular Property Prediction. 2019; 1--52.
23. Ilg E, Cicek O, Galesso S, et al. Uncertainty estimates and multi-hypotheses networks for optical flow. Lect Notes Comput Sc 2018; 11211 LNCS: 677--693.
24. Naeini MP, Cooper GF, Hauskrecht M. Obtaining well calibrated probabilities using Bayesian Binning. Proc Aaai Conf Artif Intell Aaai Conf Artif Intell 2015; 4: 2901--2907.
25. Ovadia Y, Fertig E, Ren J, et al. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift.
26. Hong S, Zhou Y, Shang J, et al. Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. Comput Biol Med 2020; 122: 103801.
27. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017 Ieee Int Conf Comput Vis Iccv 2017; 2017-Octob: 618--626.



5

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

European Heart Journal Digital Health

Rutger R van de Leur*, Max N Bos*, Karim Taha, Arjan Sammani, Ming W Yeung, Stefan van Duijvenboden, Pier D Lambiase, Rutger J Hassink, Pim van der Harst, Pieter A Doevendans, Deepak K Gupta and René van Es

Supplemental material

Abstract

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 (AUROC 0.94 vs 0.96), detection of reduced EF (AUROC 0.90 vs 0.91) and prediction of one-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.

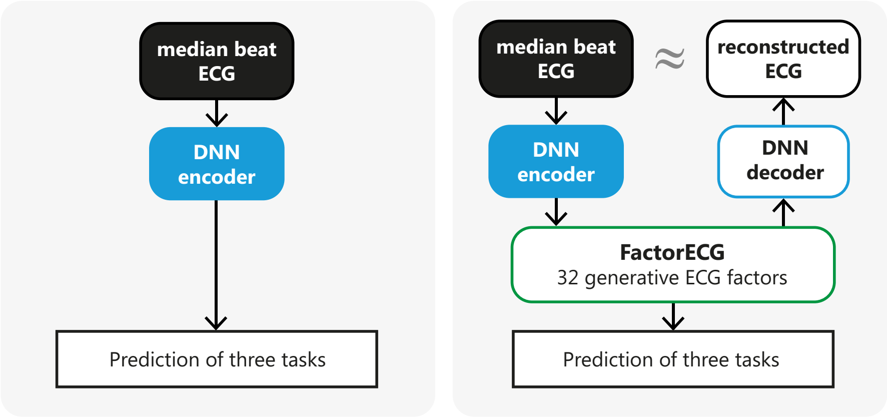
Conclusions

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.

Graphical abstract

Current approaches for explaining deep neural networks

Novel explainable variational autoencoder-based pipeline

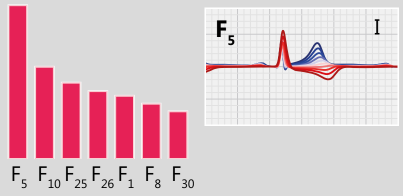


Task	AUROC
Diagnostic ECG statements	0.96
Reduced ejection fraction	0.91
One-year mortality	0.75

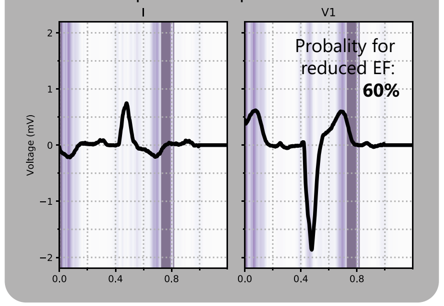
Task	AUROC
Diagnostic ECG statements	0.94
Reduced ejection fraction	0.89
One-year mortality	0.76

No model-level explainability

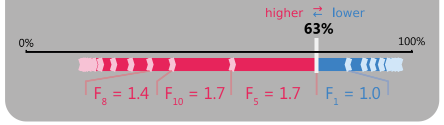
Model-level explainability
combining overall importance of ECG factors with vizualization of the factors using the decoder



Individual-level explainability
only temporal location of ECG feature important for prediction



Individual-level explainability
qualitative assessment of which ECG factors (morphological ECG changes) caused the prediction



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 and one-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 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, that 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 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 (VAE) 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 ejection fraction and prediction of one-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 VAE

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 and Taha et al (2021).[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' DNNs 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, a complete overview can be found in the Supplementary 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 5 practicing electrophysiologists or cardiologists for all diagnostic statements as described by Van de Leur et al (2020).[3] A reduced set of the 35 diagnostic statements was tested, as some abnormalities did not occur in the test dataset. Moreover, the myocardial ischemia labels in different locations were combined.

To train and validate the algorithms to detect reduced ejection fraction (below 40%) and predict one-year mortality, we selected patients using the same approaches as Attia et al. and Raghunath et al., respectively.[4,5] For the reduced ejection fraction model, patients with an ECG-echocardiogram pair (acquired within 14 days) were retrieved, the ejection fraction (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 overrepresentation of sicker patients with multiple pairs. For the one-year mortality model, all patients with at least one 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 selection 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 Methods.

For external validation of the VAE and the performance of the models for detection of reduced ejection fraction, 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 74395). All 10 second 12-lead resting ECGs were acquired using a GE CardioSoft device at 500Hz and converted to median beats by the GE algorithm. Only individuals where the left ventricular ejection fraction was determined on the CMR using a manual analysis protocol by Petersen et al were included (UK Biobank return number 2541).[16,17] Details on the UK Biobank cohort, the CMR protocol and the manual CMR analysis protocol have been described before.[17–19]

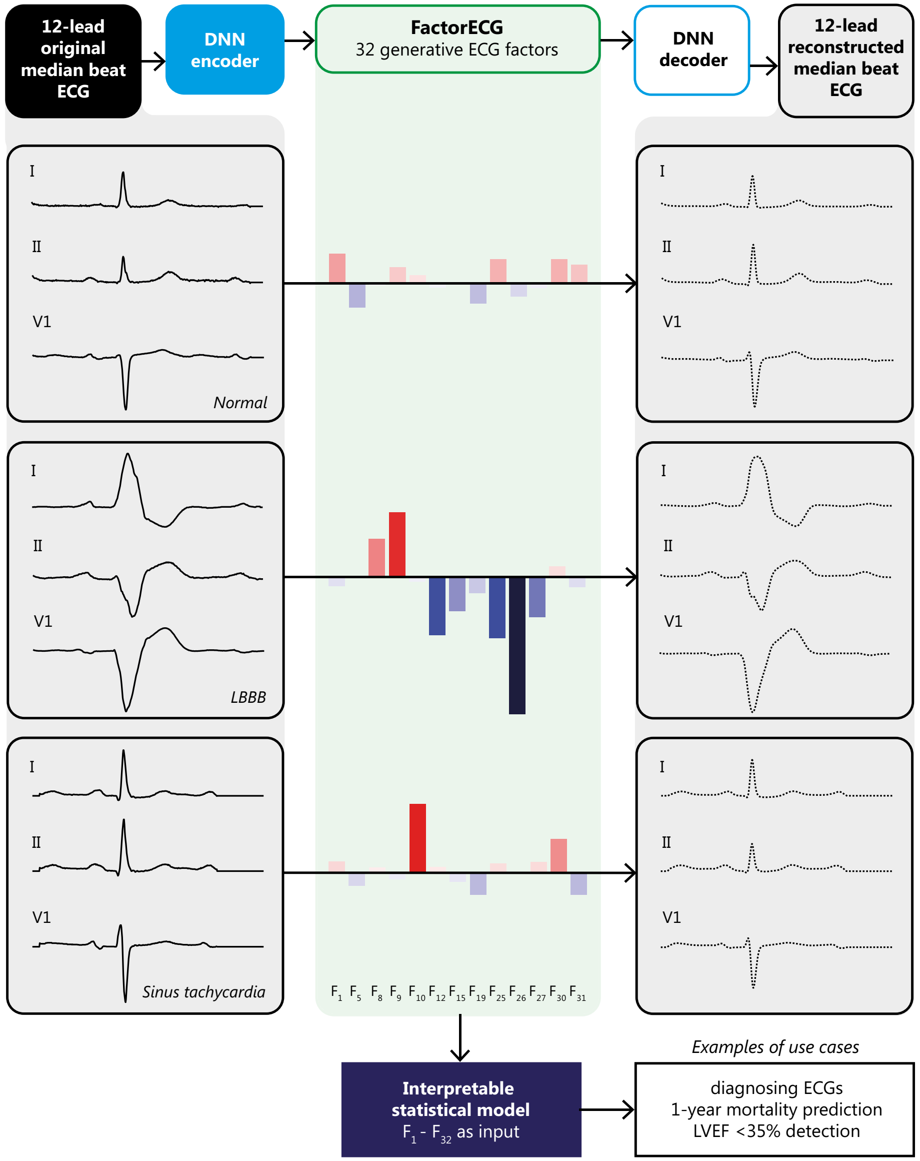

12-lead original median beat ECG
DNN encoder
FactorECG
32 generative ECG factors
DNN decoder
12-lead reconstructed median beat ECG
I
II
V1
Normal
LBBB
Sinus tachycardia
F1 F5 F8 F9 F10 F12 F15 F19 F25 F26 F27 F30 F31
Examples of use cases
Interpretable statistical model
F1 - F32 as input
diagnosing ECGs
1-year mortality prediction
LVEF <35% detection

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

Training and architecture of the VAE

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 variational autoencoder 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. Detailed information on the training and architecture of the VAE can be found in the Supplementary Methods.

Training and explainability of the pipeline

To obtain an explainable pipeline for prediction or diagnosis, we combined the following steps: (1) the median beat 12-lead ECGs were encoded in their FactorECG using the pretrained VAE encoder, (2) the 21 significant ECG factors were entered into common interpretable statistical models to perform the prediction or diagnosis task, and (3) 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 to current state-of-the-art 'black box' DNNs in three tasks: conventional ECG interpretation, detection of reduced ejection fraction and prediction of one-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 ejection fraction and prediction of one-year mortality, as the aim was maximum performance, we trained two extreme gradient boosting decision tree (XGBoost) models.[20] For this model, interpretability was obtained using Shapley Additive exPlanations (SHAP), 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 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 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. left bundle branch block or sinus tachycardia).

Statistical analysis

All data are presented as mean ± 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 diagnostic models was followed.[23]

Results

Development of the VAE 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 Table 1**.

By analyzing the factor traversals (**Supplementary Figure 2**), 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 to 0.09 (**Supplementary Figure 3**).

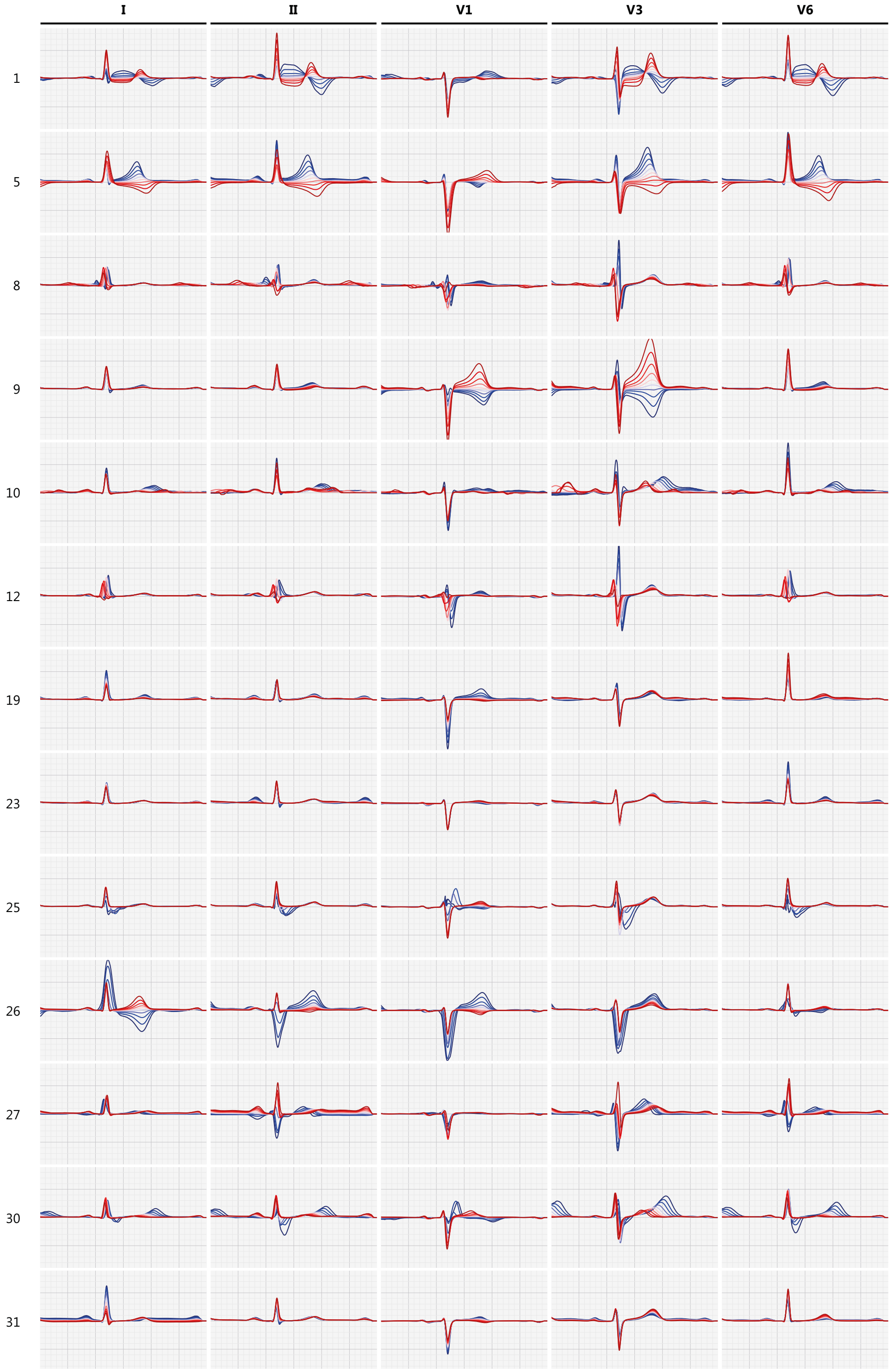


Figure 2. *Factor traversals of a subset of the ECG factors for leads I, II, V1, V3, V6. Factor traversals of a subset of the 21 ECG factors that hold significant information for correctly reconstructing ECGs.* Each row corresponds to the factor traversal for one ECG factor and the columns to a subset 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 ECGs are then plotted where red represents high values for that factor and blue low values.

Performance and explainability for conventional ECG interpretation

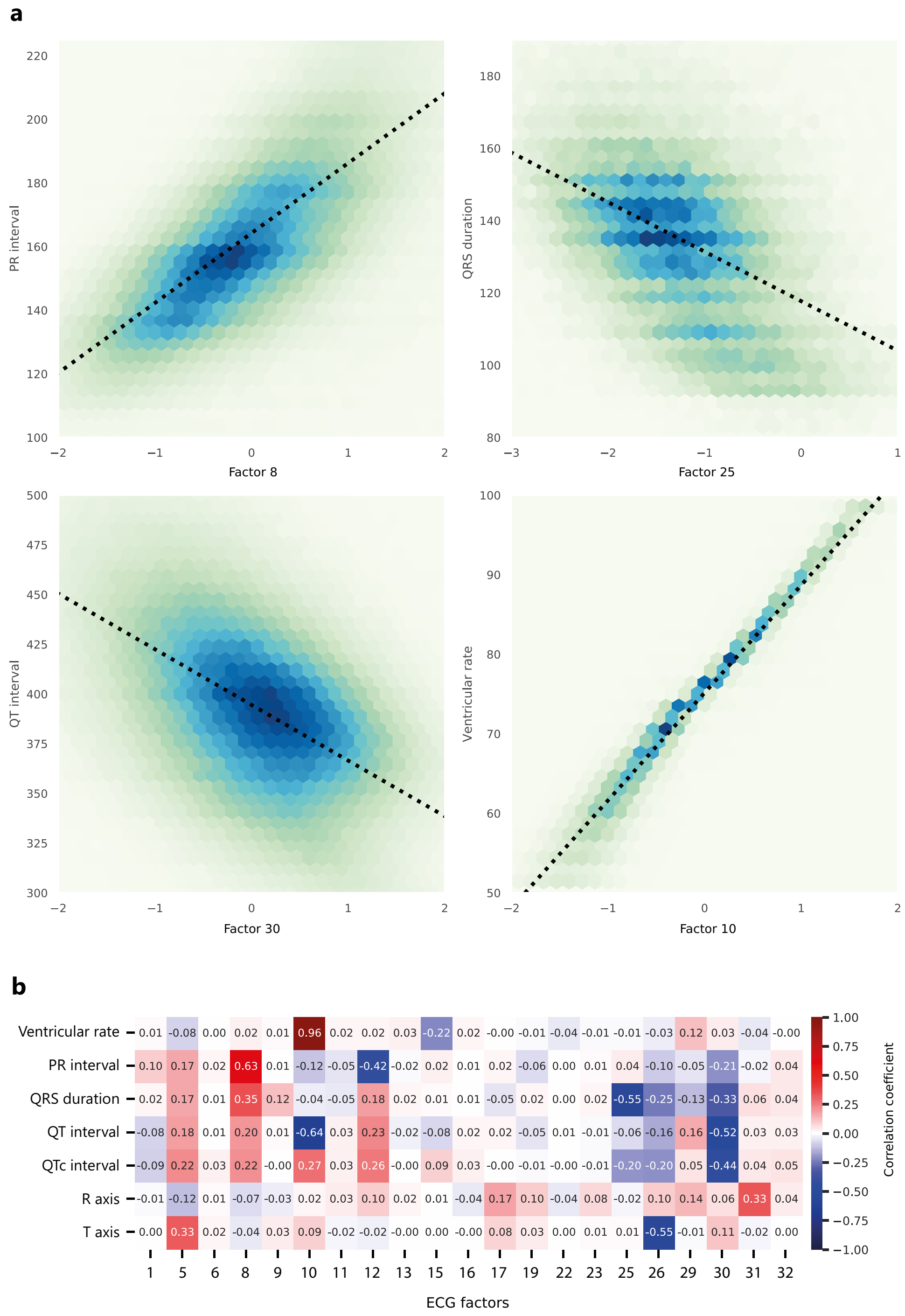
The dataset for training the algorithms to perform conventional ECG interpretation consisted of 369,216 ECGs of 152,831 patients, while for validation the expert-annotated dataset was used, containing 965 ECGs (of 965 patients) of adequate quality. 343 (36%) of the ECGs had more than one diagnostic statement 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 to 0.73 [95% CI 0.65–0.81] for the rule-based MUSE algorithm and 0.96 [95% 0.94–0.98] for the 'black box' DNN. The explainable pipeline performed similarly for most diagnostic statements but was outperformed for 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 practice, 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 Figure 2**.

Performance and explainability for detection of reduced ejection fraction

For the algorithms to detect reduced ejection fraction, 39,603 matched ECG-echocardiogram pairs of 22,676 patients were available, of which 25% (5669 unique patients, first pair per patient used) was used for validation. 713 patients (13%) in the validation set had an ejection fraction below 35%. 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 ejection fraction 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, and left bundle branch block, respectively. **Figure 6** shows a model- and individual patient-level explanation for the detection of reduced ejection fraction using the novel pipeline, in comparison to the post-hoc explainability methods used up until now.

Performance and explainability for prognosis of one-year mortality

For the models to predict one-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. 5334 patients (7.5%) in the validation set deceased within one 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 to 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 prediction of one-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 one-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.



	1	5	6	8	9	10	11	12	13	15	16	17	19	22	23	25	26	29	30	31	32
Ventricular rate	0.01	-0.08	0.00	0.02	0.01	0.96	0.02	0.02	0.03	-0.22	0.02	-0.00	-0.01	-0.04	-0.01	-0.01	-0.03	0.12	0.03	-0.04	-0.00
PR interval	0.10	0.17	0.02	0.63	0.01	-0.12	-0.05	-0.42	-0.02	0.02	0.01	0.02	-0.06	0.00	0.01	0.04	-0.10	-0.05	-0.21	-0.02	0.04
QRS duration	0.02	0.17	0.01	0.35	0.12	-0.04	-0.05	0.18	0.02	0.01	0.01	-0.05	0.02	0.00	0.02	-0.55	-0.25	-0.13	-0.33	0.06	0.04
QT interval	-0.08	0.18	0.01	0.20	0.01	-0.64	0.03	0.23	-0.02	-0.08	0.02	0.02	-0.05	0.01	-0.01	-0.06	-0.16	0.16	-0.52	0.03	0.03
QTc interval	-0.09	0.22	0.03	0.22	-0.00	0.27	0.03	0.26	-0.00	0.09	0.03	-0.00	-0.01	-0.00	-0.01	-0.20	-0.20	0.05	-0.44	0.04	0.05
R axis	-0.01	-0.12	0.01	-0.07	-0.03	0.02	0.03	0.10	0.02	0.01	-0.04	0.17	0.10	-0.04	0.08	-0.02	0.10	0.14	0.06	0.33	0.04
T axis	0.00	0.33	0.02	-0.04	0.03	0.09	-0.02	-0.02	-0.00	0.00	-0.00	0.08	0.03	0.00	0.01	0.01	-0.55	-0.01	0.11	-0.02	0.00

Figure 3.* *Relationship of the ECG factors with conventional ECG measurements. a. Hexagon plots where datapoints of ECG factor-ECG measurements pairs over all samples in the VAE 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 ECG measures of ventricular rate, PR interval, QRS duration, QT interval, Bazett corrected QT interval, R-axis, and T-axis and ECG factor values over all samples in the VAE dataset.

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

Manually analyzed CMR imaging and 12-lead ECG recordings were available for 4855 individuals, of which 28 had a reduced ejection fraction (0.62%). The VAE, that 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 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).

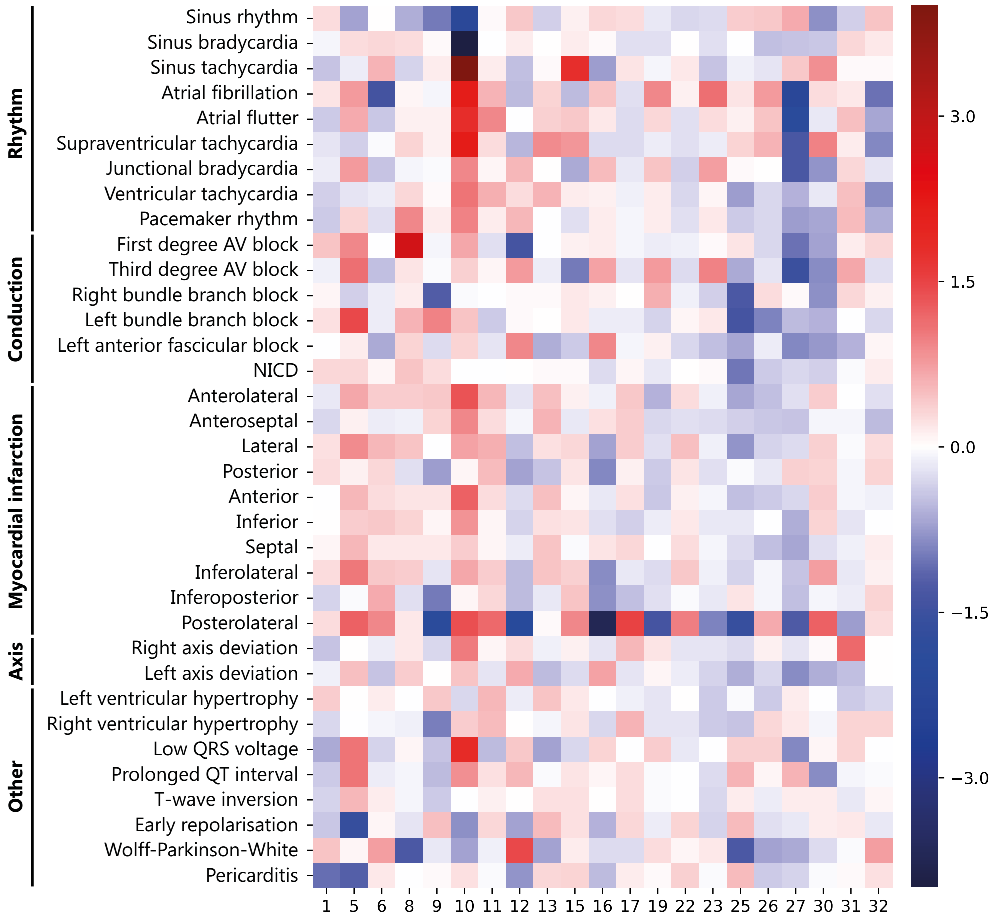


Figure 4.

Importance score for each of the 32 factors in predicting 35 diagnostic ECG statements. Importance scores of each of the 32 factors in the logistic regression for all 35 diagnostic ECG 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: nonspecific intraventricular conduction delay.

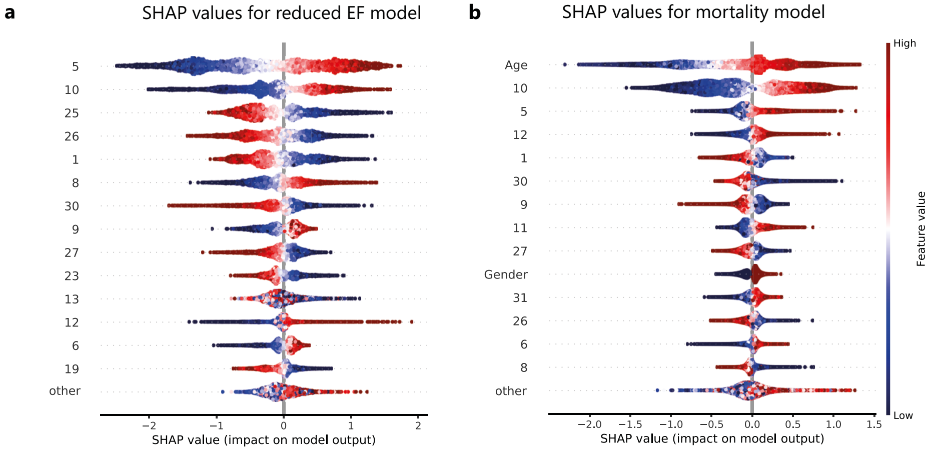


Figure 5.

Explanations for the one-year mortality and reduced EF models using SHAP values. a. The most important model-level ECG factors for detecting reduced ejection fraction computed using SHAP values. Importance is ordered from top-to-bottom and coloring corresponds to the reconstructed ECGs in Figure 2. b. The most important global ECG factors for predicting one-year mortality. Importance is ordered from top-to-bottom and coloring corresponds to the reconstructed ECGs in Figure 2.

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 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 is this work consists of three parts: firstly, the ECGs is encoded in its FactorECG using the pretrained VAE encoder, then the 21 significant ECG factors are entered into interpretable sta-tistical 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.
Factor ECG	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 one 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 ejec-tion 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 val-ues, 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.

Table 1. *Glossary of terms used throughout the manuscript.*

DIAGNOSTIC STATEMENT	PREVALENCE	MUSE 12 SL		EXPLAINABLE PIPELINE		BLACK BOX DNN	
	n (%)	AUROC [95% CI]	AU-PRC	AUROC [95% CI]	AU-PRC	AUROC [95% CI]	AU-PRC
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
Ventricular tachycardia	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
Junctional bradycardia	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

DIAGNOSTIC STATEMENT	PREVALENCE	MUSE 12 SL		EXPLAINABLE PIPELINE		BLACK BOX DNN	
	n (%)	AUROC [95% CI]	AU-PRC	AUROC [95% CI]	AU-PRC	AUROC [95% CI]	AU-PRC
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 repolarisation	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

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

FACTOR	HIGH VALUES		LOW VALUES	
	ECG Morphology	**Associations**	**ECG Morphology**	**Associations**
1	Inferolateral horizonal ST depression	Left ventricular hypertrophy	Inferolateral horizonal ST depression	Pericarditis, reduced EF and one-year mortality
5	Inferolateral T-wave inversion	T-wave axis, LBBB, inferior and lateral ischemia, low QRS voltage, reduced EF and one-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 ischemia, T-wave inversion and one-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 one-year mortality	Longer QT-interval and TP-interval	
11	Subtle QRS- and T-wave changes	One-year mortality	Subtle QRS- and T-wave changes	Increased PR-interval and first-degree AV block
12	Earlier onset of depolarisation	Reduced PR-interval, WPW pattern, LAFB and one-year mortality	Later onset of depolarization	Reduced EF
13	Anterior horizon-tal ST-elevation	Anterior and septal ischemia	Anterior horizontal ST-depression	Third degree AV-block and junctional brady-cardia.
15	P/T overlap	Sinus tachycardia	Reduced P-wave amplitude	Posterior and lateral ischemia
16	Subtle T-wave changes	LAFB	Subtle T-wave changes	Inferior ischemia
17	Lateral horizontal ST-elevation	Lateral ischemia and right ventricular hypertrophy	Lateral horizontal ST-depression	
19	Slower R-wave progression		Faster R-wave progression	
22	Baseline shift		Baseline shift	
23	Reduced P-wave amplitude	Atrial fibrillation, junctional bradycar-dia, third degree AV block	Increased P-wave amplitude	

FACTOR	HIGH VALUES		LOW VALUES	
	ECG Morphology	**Associations**	**ECG Morphology**	**Associations**
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 flut-ter, SVT, junctional bradycardia, reduced EF and one-year mortality
30	Shorter QT-interval		Longer QT-interval	Prolonged QT interval, reduced EF and one-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 pre-cordial QRS-amplitude		Increased precordial QRS amplitude	LVH

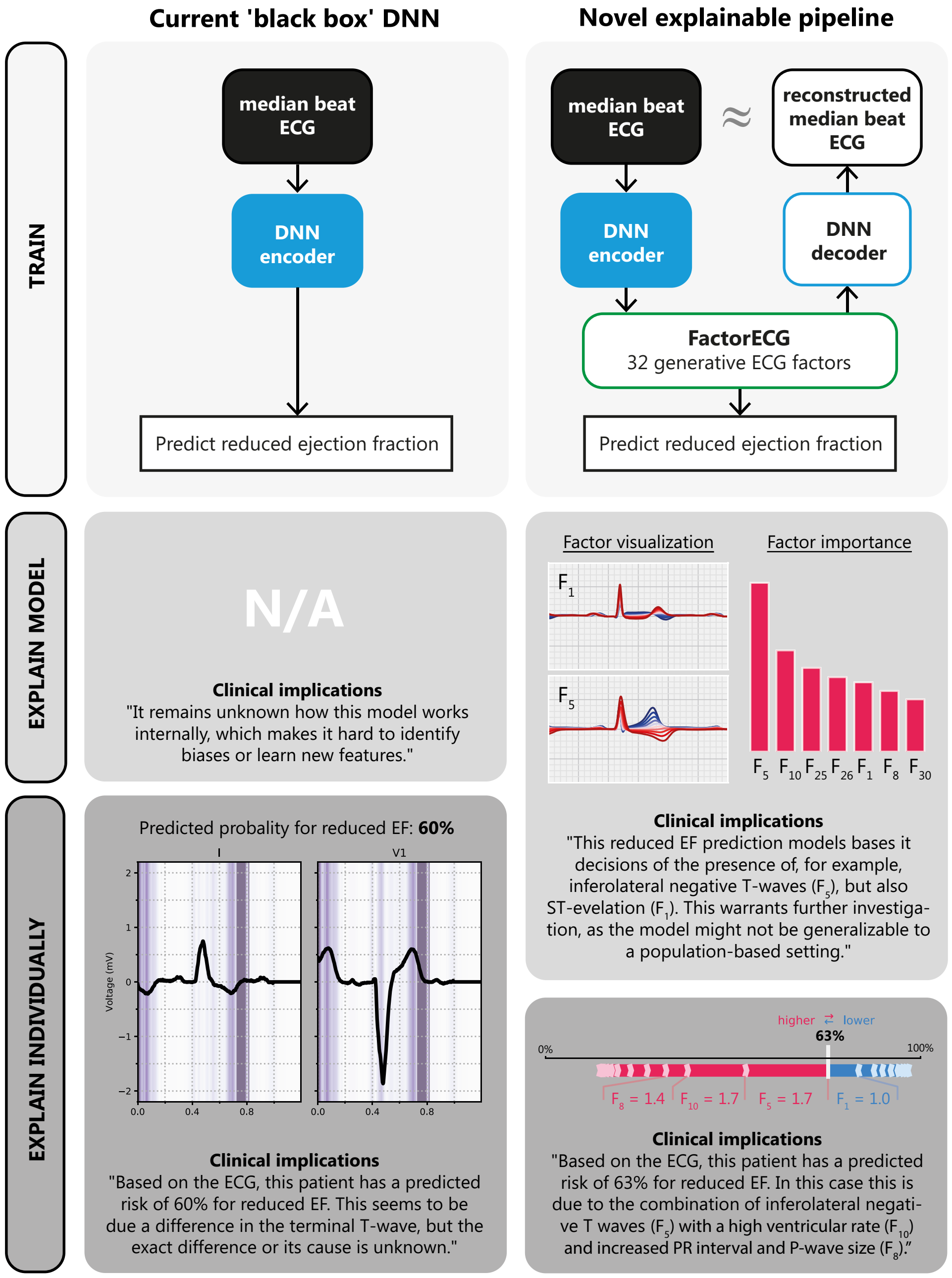
Table 3.
Summarizing description of ECG morphology and associations of the 21 significant ECG factors. The influence of an ECG factor on median beat ECG morphology is determined using visual inspection of the factor traversals (Figure 2). A summary of the most important associations of every ECG factor with conventional ECG measurements, ECG diagnostic statements, reduced EF and one-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: nonspecific intraventricular conduction delay, RBBB: right bundle branch block, RVH: right ventricular hypertrophy, SVT: supraventricular tachycardia, WPW: Wolff-Parkinson-White.

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 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 represents 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 ejection fraction, 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 quantitively specify the 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 also determined that approximately 20-25 factors are needed to encode a single or median beat ECG.[28] Our work makes the latent space of a VAE (i.e. the FactorECG) clinically useful and explainable to physicians, by (a) linking the ECG factors with known ECG measurements and diagnostic statements (**Figures 3 and 4 and Table 2 and 3**), (b) providing extensive visualizations offline (**Figure 2**) and using an online tool (https://decoder.ecgx.ai) and (c) 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 practice. 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://encoder.ecgx.ai and https://github.com/rutgervandeleur/ecgxai).



Clinical implications

"It remains unknown how this model works internally, which makes it hard to identify biases or learn new features."

Clinical implications

"This reduced EF prediction models bases it decisions of the presence of, for example, inferolateral negative T-waves (F_5), but also ST-evelation (F_1). This warrants further investigation, as the model might not be generalizable to a population-based setting."

Clinical implications

"Based on the ECG, this patient has a predicted risk of 60% for reduced EF. This seems to be due a difference in the terminal T-wave, but the exact difference or its cause is unknown."

Clinical implications

"Based on the ECG, this patient has a predicted risk of 63% for reduced EF. In this case this is due to the combination of inferolateral negative T waves (F_5) with a high ventricular rate (F_{10}) and increased PR interval and P-wave size (F_8)."

Figure 6.
Comparison of architecture and model- and individual patient-level explainability using the novel inherently explainable approach as compared to post-hoc heatmap-based explainability for detection of reduced ejection fraction. The conventional 'black box' DNN contains only a single encoder to interpret the ECG. Afterwards, Guided Grad-CAM is applied to show what segments of the ECG 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 ECG features. By combining factor SHAP importance scores and factor traversals, we obtain model-level explainability. Individual patient-level explainability is achieved using individual SHAP importance score.

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 Figure 2**) 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 left bundle branch block, 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 ejection fraction 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 importance of this latter feature illustrates that the DNN also picks up reduced ejection fraction due to acute ischemia. 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 one-year mortality performs worse than in the original paper by Raghunath et al., it does perform similarly to the 'black box' DNN on our dataset.[5] The difference in performance is likely 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 one-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 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 ischemia classes and lower correlation coefficients of the reconstructed ECGs (**Supplementary Table 1**). 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 (**Supplementary Table 1**). 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]

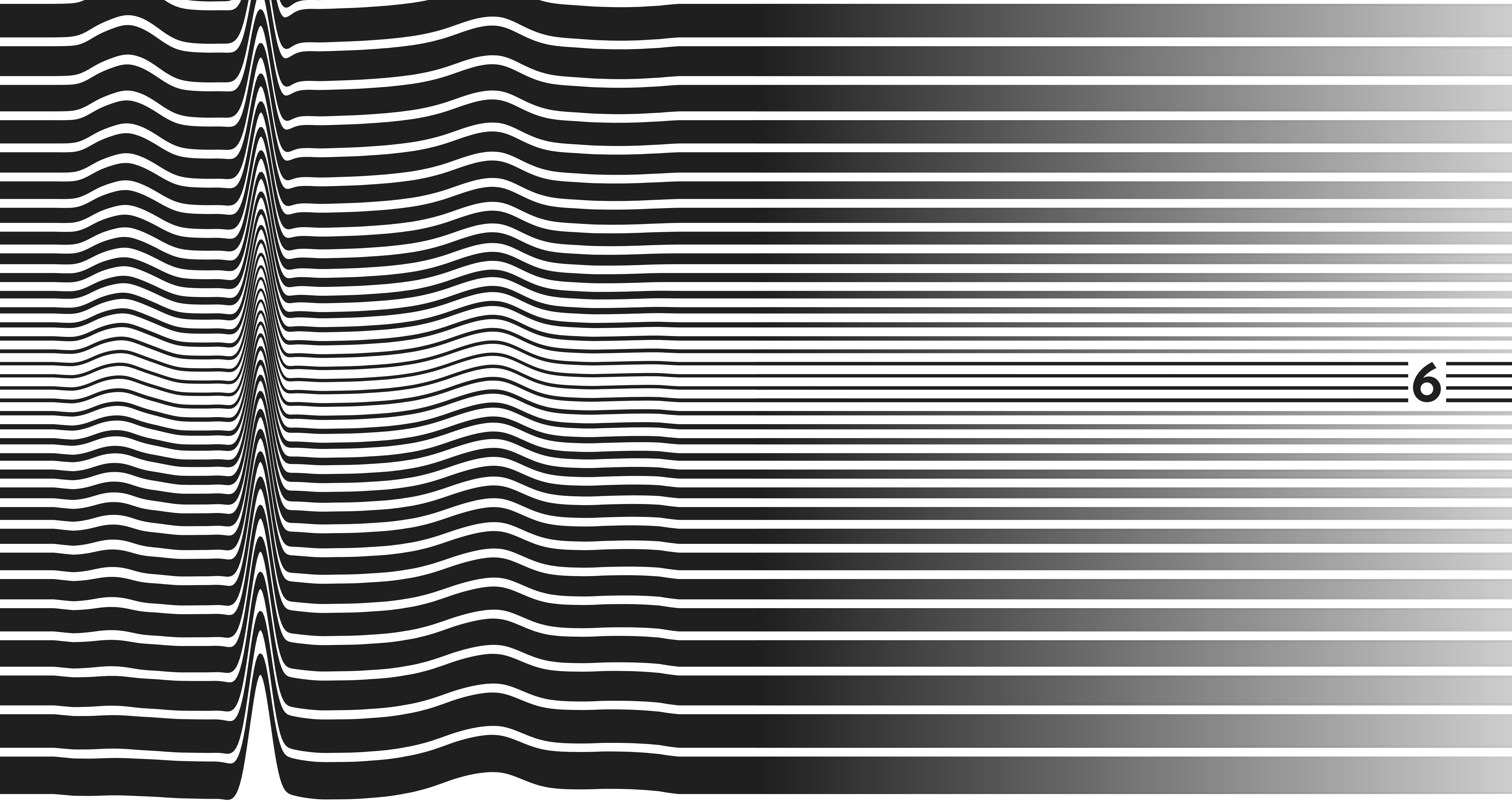
Future studies should focus on evaluating the use of inherently explainable pipelines on other ECG tasks, as the dimensionality reduction of our algorithm to 21 factors broadens the usability of DNNs greatly to much smaller labeled datasets than before. Another important perspective is using the approach on full 10-second 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.

Conclusion

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.

REFERENCES

1. Leur RR van de, Boonstra MJ, Bagheri A, et al. Big Data and Artificial Intelligence: Opportunities and Threats in Electrophysiology. *Arrhythmia Electrophysiol Rev* 2020; 9: 146–154.
2. Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med* 2019; 25: 65–69.
3. Leur RR van de, Blom LJ, Gavves E, et al. Automatic Triage of 12-Lead Electrocardiograms Using Deep Convolutional Neural Networks. *J Am Heart Assoc*; 9. Epub ahead of print 2020. DOI: 10.1161/jaha.119.015138.
4. Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. *Nat Med* 2019; 25: 70--74.
5. Raghunath S, Cerna AEU, Jing L, et al. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nat Med* 2020; 26: 886--891.
6. 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.
7. Kundu S. AI in medicine must be explainable. *Nat Med* 2021; 1–1.
8. 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.
9. Goodman B, Flaxman S. European Union Regulations on Algorithmic Decision-Making and a “Right to Explanation.” *Ai Mag* 2017; 38: 50–57.
10. Leur RR van de, Taha K, Bos MN, et al. Discovering and Visualizing Disease-Specific Electrocardiogram Features Using Deep Learning: Proof-of-Concept in Phospholamban Gene Mutation Carriers. *Circulation Arrhythmia Electrophysiol*; 14. Epub ahead of print 2021. DOI: 10.1161/circep.120.009056.
11. Kwon J, Cho Y, Jeon K-H, et al. 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, et al. Sanity Checks for Saliency Maps. In: *Advances in Neural Information Processing Systems 31*, pp. 9505--9515.
13. Hooker S, Erhan D, Kindermans P-J, et al. 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., pp. 9737–9748.
14. Kingma DP, Welling M. Auto-Encoding Variational Bayes. In: Bengio Y, LeCun Yann (eds) *2nd International Conference on Learning Representations*. Banff, AB, Canada: Conference Track Proceedings, http://arxiv.org/abs/1312.6114 (2014).
15. Higgins I, Matthey L, Pal A, et al. 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, et al. 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, et al. Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UK Biobank population cohort. *J Cardiov Magn Reson* 2017; 19: 18.
18. Petersen SE, Matthews PM, Francis JM, et al. UK Biobank’s cardiovascular magnetic resonance protocol. *J Cardiov Magn Reson* 2016; 18: 8.
19. Sudlow C, Gallacher J, Allen N, et al. 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, pp. 785–794.
21. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg U. V., Bengio S., et al. (eds) *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., pp. 4765--4774.
22. Bos MN, Leur RR van de, Vranken JF, et al. Automated Comprehensive Interpretation of 12-lead Electrocardiograms Using Pre-trained Exponentially Dilated Causal Convolutional Neural Networks. *2020 Comput Cardiol* 2020; 00: 1–4.
23. Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med* 2015; 162: 55.
24. Kwon J, Lee SY, Jeon K, et al. Deep Learning–Based Algorithm for Detecting Aortic Stenosis Using Electrocardiography. *J Am Heart Assoc* 2020; 9: e014717.
25. Jang J-H, Kim TY, Lim H-S, et al. 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, et al. 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, Gribanov DV, et al. Interpretable Feature Generation in ECG Using a Variational Autoencoder. *Frontiers Genetics* 2021; 12: 638191.
29. O’Neal WT, Mazur M, Bertoni AG, et al. Electrocardiographic Predictors of Heart Failure With Reduced Versus Preserved Ejection Fraction: The Multi‐Ethnic Study of Atherosclerosis. *J Am Heart Assoc*; 6. Epub ahead of print 2017. DOI: 10.1161/jaha.117.006023.
30. Yao X, McCoy RG, Friedman PA, et al. 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, et al. Heart rate and cardiovascular mortality: The Framingham study. *Am Heart J* 1987; 113: 1489–1494.
32. Porthan K, Viitasalo M, Jula A, et al. 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, et al. A comprehensive artificial intelligence–enabled electrocardiogram interpretation program. *Cardiovasc Digital Heal J* 2020; 1: 62--70.
34. Siegersma KR, Leur RR van de, Onland-Moret NC, et al. Deep neural networks reveal novel sex-specific electrocardiographic features relevant for mortality risk. *European Hear J - Digital Heal*. Epub ahead of print 2022. DOI: 10.1093/ehjdh/ztac010.



6

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

European Heart Journal Digital Health

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

To the Editor,

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 (VAE), which use deep neural networks (DNNs) to learn the underlying factors of variation in the median beat 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. have shown that even untrained DNNs provide heatmaps that look reassuring.[3] Moreover, Hooker et al. have shown that when you remove the regions deemed important by many saliency-based methods, performance of the classifier does not decrease after retraining.[4] 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 2-dimensional images of the ECG into the deep neural network 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 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 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 XGBoost with SHAP) 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 AI, allow for bias detection and broaden the application of AI to many other (rare) diseases.

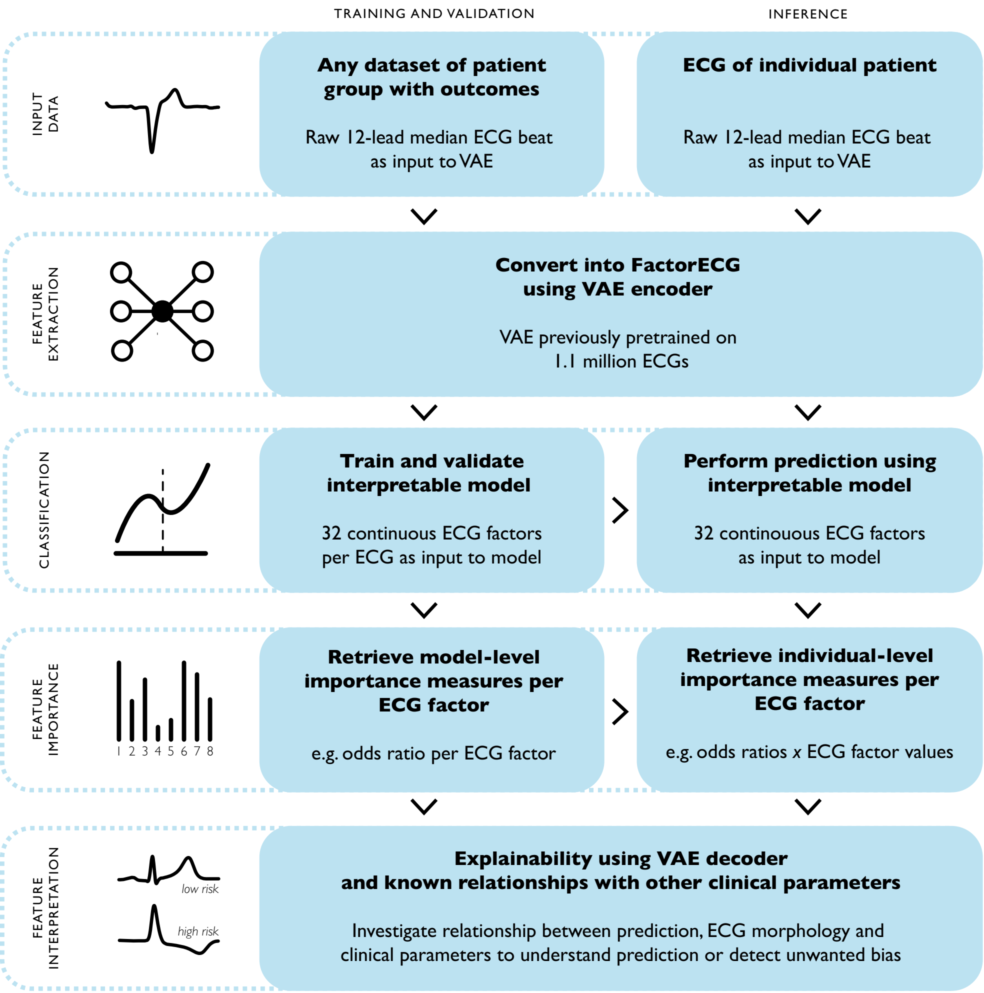
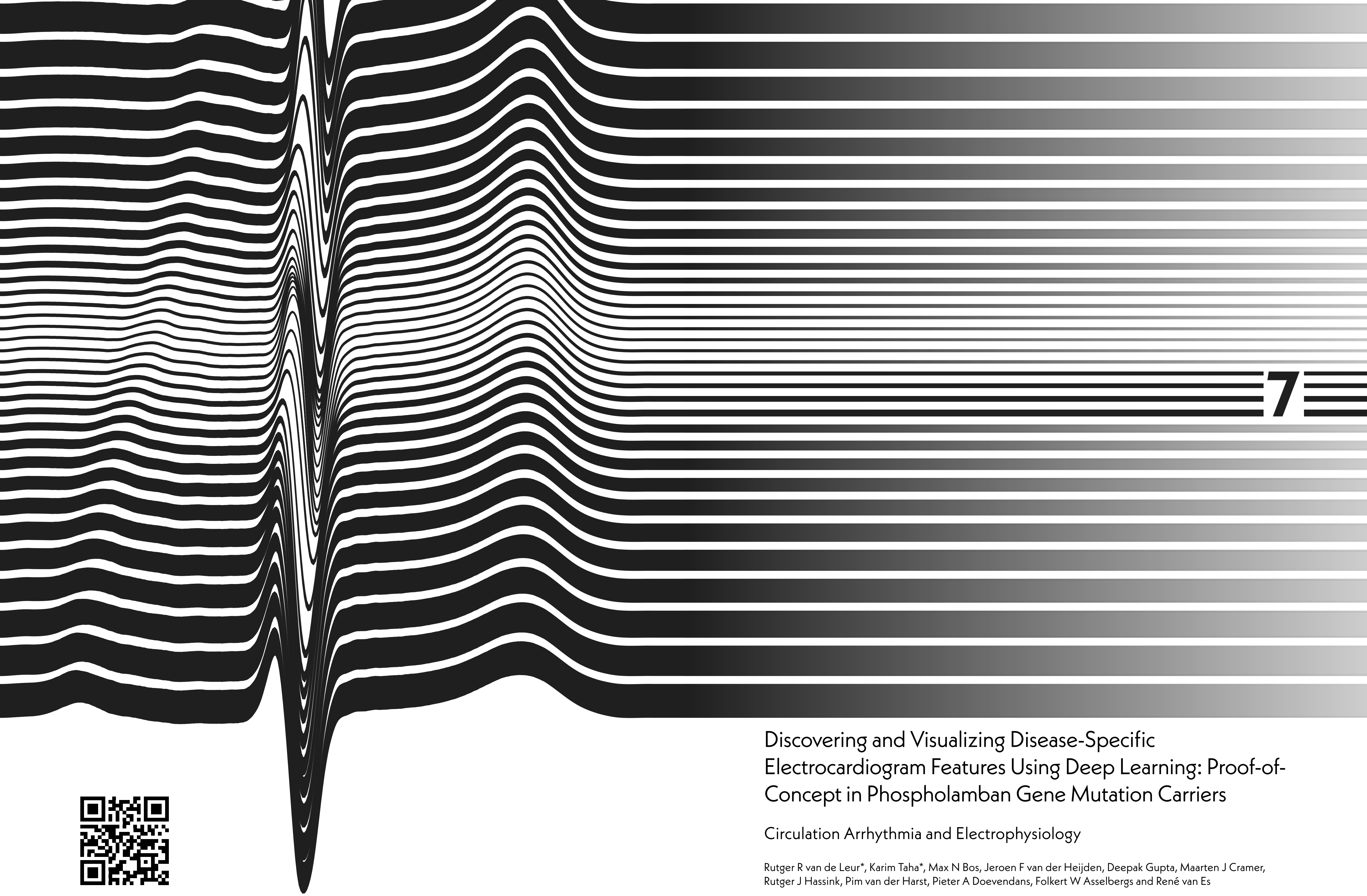

TRAINING AND VALIDATION
INFERENCE
INPUT DATA
Any dataset of patient group with outcomes
Raw 12-lead median ECG beat as input to VAE
ECG of individual patient
Raw 12-lead median ECG beat as input to VAE
FEATURE EXTRACTION
Convert into FactorECG using VAE encoder
VAE previously pretrained on 1.1 million 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 ECG dataset with available outcomes can be used as input to the pipeline.* As the ECGs are converted to only 32 factors, datasets can be relatively small. The VAE encoder, that was pretrained on 1.1 million ECGs, is subsequently used to convert every single ECG into its FactorECG (32 continuous values that represent that ECG). These 32 factors per ECG 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 ECG factor individually. As we are able to visualize the influence of the individual ECG factors on the ECG morphology using the VAE decoder, a direct relationship between ECG morphology and the prediction can be obtained on the model-level. During inference (right), an individual ECG can be entered into the pretrained VAE encoder. Prediction is performed using the previously trained interpretable, and individual-level importance measures per ECG factor are obtained. These individual importance measures can subsequently be related to the ECG 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).

REFERENCES

1. Leur RR van de, Bos MN, Taha K, et al. Improving explainability of deep neural network-based electrocardiogram interpretation using variational auto-encoders. *European Hear J - Digital Heal*. Epub ahead of print 2022. DOI: 10.1093/ehjdh/ztac038.
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, et al. Sanity Checks for Saliency Maps. In: *Advances in Neural Information Processing Systems 31*, pp. 9505--9515.
4. Hooker S, Erhan D, Kindermans P-J, et al. 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., pp. 9737–9748.
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, et al. Performance of a convolutional neural network derived from an ECG database in recognizing myocardial infarction. *Sci Rep-uk* 2020; 10: 8445.
7. Leur RR van de, Taha K, Bos MN, et al. Discovering and Visualizing Disease-Specific Electrocardiogram Features Using Deep Learning: Proof-of-Concept in Phospholamban Gene Mutation Carriers. *Circulation Arrhythmia Electrophysiol*; 14. Epub ahead of print 2021. DOI: 10.1161/circep.120.009056.
8. Hughes JW, Olgin JE, Avram R, et al. Performance of a Convolutional Neural Network and Explainability Technique for 12-Lead Electrocardiogram Interpretation. *Jama Cardiol*; 6. Epub ahead of print 2021. DOI: 10.1001/jamacardio.2021.2746.
9. Sammani A, Leur RR van de, Henkens MTHM, et al. Life-threatening ventricular arrhythmia prediction in patients with dilated cardiomyopathy using explainable electrocardiogram-based deep neural networks. *Ep Europace*. Epub ahead of print 2022. DOI: 10.1093/europace/euac054.
10. Wouters P, Leur R van de, Vessies M, et al. PO-658-01 EXPLAINABLE DEEP LEARNING OUTPERFORMS GUIDELINE CRITERIA FOR PREDICTION OF CARDIAC RESYNCHRONIZATION THERAPY OUTCOME. *Heart Rhythm* 2022; 19: S274–S275.



7

Discovering and Visualizing Disease-Specific Electrocardiogram Features Using Deep Learning: Proof-of-Concept in Phospholamban Gene Mutation Carriers

Circulation Arrhythmia and Electrophysiology

Rutger R van de Leur*, Karim Taha*, Max N Bos, Jeroen F van der Heijden, Deepak Gupta, Maarten J Cramer, Rutger J Hassink, Pim van der Harst, Pieter A Doevendans, Folkert W Asselbergs and René van Es

Abstract

Background

Electrocardiogram (ECG) interpretation requires expertise and is mostly based on physician recognition of specific patterns, which may be challenging in rare cardiac diseases. Deep neural networks (DNN) can discover complex features in ECGs and may facilitate the detection of novel features which possibly play a pathophysiological role in relatively unknown diseases. Using a cohort of phospholamban (PLN) p.Arg14del mutation carriers, we aimed to investigate whether a novel DNN-based approach can identify established ECG features, but moreover we aimed to expand our knowledge on novel ECG features in these patients.

Methods

A DNN was developed on 12-lead median beat ECGs of 69 patients and 1380 matched controls and independently evaluated on 17 patients and 340 controls. Differentiating features were visualized using Guided Grad-CAM++. Novel ECG features were tested for their diagnostic value by adding them to a logistic regression model including established ECG features.

Results

The DNN showed excellent discriminatory performance with a c-statistic of 0.95 (95% confidence interval 0.91-0.99) and sensitivity and specificity of 0.82 and 0.93, respectively. Visualizations revealed established ECG features (low QRS voltages and T-wave inversions), specified these features (e.g. R and T-wave attenuation in V2/V3) and identified novel PLN-specific ECG features (e.g. increased PR-duration). The logistic regression baseline model improved significantly when augmented with the identified features ($p<0.001$).

Conclusions

A DNN-based feature detection approach was able to discover and visualize disease-specific ECG features in PLN mutation carriers and revealed yet unidentified features. This novel approach may help advance diagnostic capabilities in daily practice.

Introduction

Interpretation of the electrocardiogram (ECG) requires expertise and is mainly based on physician recognition of patterns that are known to belong to a particular disorder. However, for rare and relatively unknown cardiac diseases, this may be challenging since ECG features are often unknown and require expert knowledge to recognize. By automating the discovery and expanding the knowledge on disease-specific ECG features, interpretation of ECGs by physicians could be improved. Such a support tool could be of particular importance when expert knowledge is not readily available or in research settings to automate the detection of disease-specific ECG features.

Recently, ECGs have been analyzed using deep neural networks (DNNs), which are computer algorithms that are based on the structure and functioning of the human brain.[1] Their layers can be trained to discover complex patterns in ECGs, without requiring hand-crafted feature extraction. Several studies have applied DNNs for automated predictions from ECGs, and one recent study showed that it is feasible to diagnose hypertrophic cardiomyopathy (HCM) on the ECG.[2–4] However, the methods used in these studies all require very large datasets, which are often not available for rare diseases. Furthermore, these previous studies all focus on prediction, but specific ECG patterns used by DNNs are rarely visualized.[3,5–8] Visualization of such features takes advantage of the feature discovery embedded in DNNs and will help clinicians to interpret ECGs more accurately, and possibly facilitate discovery of novel features.

Cardiomyopathy-related genetic mutations are rare but are often associated with typical ECG features. An example is the deletion of three base pairs (c.40_42delAGA) in the phospholamban (*PLN*) gene, leading to the deletion of Arginine 14 in the PLN protein (p.Arg14del).[9–11] Prevalence of the PLN p.Arg14del mutation is estimated to be 0.07% in the northern regions of the Netherlands and is present in 12% of Dutch patients developing a phenotype of arrhythmogenic right ventricular cardiomyopathy (ARVC) and in 15% of patients developing dilated cardiomyopathy (DCM).[11–13] With regard to ECG characteristics in these mutation carriers, typical features that have previously been described are attenuated QRS-amplitudes and inverted T-waves in the right and left precordial leads.[12,14,15]

Beside using DNNs merely for prediction or diagnosis, we hypothesize that DNNs can also be used for feature visualization itself. This will potentially enable discovery of novel ECG features that belong to a particular disease. In this study, we used a cohort of PLN mutation carriers to investigate whether a novel DNN-based approach can (i) identify the already well-established ECG features in these mutation carriers and (ii) possibly expand our knowledge on ECG features in these mutation carriers.

Methods

Data availability

The data used in this study are not publicly available due to privacy restrictions. The code for training the DNN and for generating the visualizations and tables in this paper is available upon request from the corresponding author.

Data source and study participants

The dataset consisted of 12-lead ECGs from patients between 18 and 85 years old acquired in the University Medical Center Utrecht (UMCU) from January 2000 to August 2019. All extracted data were de-identified in accordance with the EU General Data Protection Regulation and written informed consent was therefore not required by the ethical committee. All ECGs were interpreted by a physician as part of the clinical workflow and these free text annotations were structured using a text mining algorithm described before.[3] We excluded all ECGs of insufficient quality and all ECGs with supraventricular and ventricular arrhythmias (excluding premature atrial and ventricular complexes), paced rhythms, undefined rhythms and signs of acute ischemia.

All index patients in the dataset who carry the genetic PLN p.Arg14del mutation and their relatives that tested positive, were identified. ECGs acquired after the implantation of a left ventricular assist device (LVAD) or heart transplantation were excluded. Only the first acquired ECG of each mutation carrier was used for development of the model.

The control group was derived from the remaining dataset and consisted of 365,173 ECGs of 147,098 unique patients. Per mutation carrier, 20 controls were matched using propensity score matching on age and sex. This number was chosen to have sufficient samples to train the DNN without having a too severe class imbalance. Only one ECG per control subject, sampled without replacement, was used to make sure every subject was only used once. The matched groups were randomly split in an 80:20 manner to training and test sets.

Data acquisition

For all ECGs the median beats were exported from the MUSE ECG system (MUSE version 8, GE Healthcare, Chicago, IL, United States). The median beat data is constructed by aligning all QRS-complexes of the same shape (e.g. excluding premature ventricular complexes) and generating a representative QRS-complex by taking the median voltage.[16] Acquisition and feature extraction of the included ECGs is described in more detail in the Supplemental Material.

Baseline logistic regression model

To demonstrate the capability of DNN in identifying novel relevant features, we first developed a baseline logistic regression model, only based on the established ECG features of PLN mutation carriers. The matching variables, age and sex, and the established PLN-specific ECG features (low QRS voltage and right (V2-V3) and left (V4-V6) precordial T-wave inversion) were included as predictors in the model.[17] The model was trained on the training dataset and evaluated on the test set.

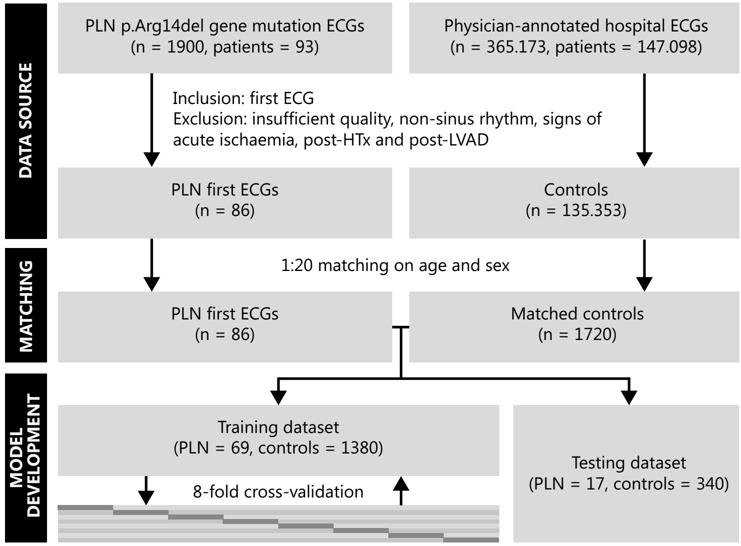

DATA SOURCE
PLN p.Arg14del gene mutation ECGs
(n = 1900, patients = 93)
Physician-annotated hospital ECGs
(n = 365.173, patients = 147.098)
Inclusion: first ECG
Exclusion: insufficient quality, non-sinus rhythm, signs of acute ischaemia, post-HTx and post-LVAD
PLN first ECGs
(n = 86)
Controls
(n = 135.353)
MATCHING
1:20 matching on age and sex
PLN first ECGs
(n = 86)
Matched controls
(n = 1720)
MODEL DEVELOPMENT
Training dataset
(PLN = 69, controls = 1380)
8-fold cross-validation
Testing dataset
(PLN = 17, controls = 340)

Figure 1. Flowchart of the patient selection and model development process.
ECG: electrocardiogram, HTx: heart transplantation, LVAD: left ventricular assist device, PLN: phospholamban.

Deep neural network development

We constructed a deep convolutional neural network with exponentially dilated causal convolutions. The proposed architecture, inspired by the method described by Van Oord *et al*. and Franceschi *et al*., compromises of several 1-dimensional dilated causal convolution blocks.[18,19] Eight-fold cross validation on the training dataset was used for optimization of the hyperparameters of the network. The simplest network with the highest geometric mean of area under the receiver operating curve and F2 score averaged over all folds was chosen and trained on the complete training dataset. The performance of this network was estimated on the test subset. Network training was performed using the PyTorch package (version 1.3).[20] A detailed description of the architecture of the DNN can be found in the Expanded Methods and an overview of the network architecture is shown in **Supplemental Figure 1**.

Feature visualization

To identify the parts of the ECG that are important for the DNNs prediction, we applied Guided Gradient Class Activation Mapping ++ (Guided Grad-CAM++), a technique for explanations in convolutional neural networks, to 1-dimensional data.[5,6] Guided Grad-CAM++ combines the fine-grained and lead-specific visualizations of guided backpropagation with the class-discriminative and global Grad-CAM technique. The median beat visualization methodology is described in more detail in the Supplemental Material.

Validation of newly identified features in an updated model

Based on inspection of the visualization output, we identified distinctive features with an arbitrary prevalence above 25%. The detected important features were translated to quantitative features (e.g. R-wave amplitude) and added to the baseline logistic regression model, starting with the most prevalent. If multiple similar features were found in leads belonging to the same region, the most prevalent feature in that region was used. Leads I, aVL and V4-V6 were grouped as lateral leads and II, III and aVF as inferior leads. To evaluate the added value of the newly identified ECG features, we determined if the nested baseline logistic regression model fit improved using the likelihood ratio test (LRT) and Akaike's information criterion (AIC).

Subgroup analyses

In subgroup analyses, we analyzed whether predictive performance and detected features differed between subsets of patients. Due to the small sample size, these exploratory subgroup analyses were performed on the combined training and test datasets. We investigated the performance in presymptomatic PLN p.Arg14del mutation carriers. Presymptomatic was defined as: no cardiac symptoms as per judgement of the treating physician, no history of (non-)sustained ventricular arrhythmia, premature ventricular complex burden of <500 beats per 24 hours and left ventricular ejection fraction ≥45%.

Statistical analysis

The baseline characteristics were expressed as mean ± standard deviation or median with interquartile range (IQR), where appropriate. Categorical variable differences were tested using the chi-square test or Fisher's exact test and continuous variables using the Student's t-test or Mann Whitney U test. Multiple testing correction was performed for the baseline characteristics using Bonferroni's method. The overall discriminatory performance of the DNN, baseline and updated models were assessed in the test set with the concordance-statistic (c-statistic) or area under the receiver operating characteristic (ROC) curve, sensitivities, specificities, positive and negative predictive values. The models were compared at a prespecified specificity of 94%. The 95% confidence intervals (CI) around the performance measures and odd's ratios were obtained using 2000 bootstrap samples. All statistical analyses were performed using R version 3.5 (R Foundation for Statistical Computing, Vienna, Austria).

Results

Study population

A total of 93 PLN p.Arg14del mutation carriers were identified, of which 86 were eligible for this study. Four patients were excluded as all their ECGs were acquired after LVAD or heart transplantation and three patients as all their ECGs were non-sinus rhythm. The control group consisted of 135,353 patients after exclusions, of which 1,720 patients were matched. The flowchart is shown in **Figure 1** and the baseline characteristics in **Table 1**.

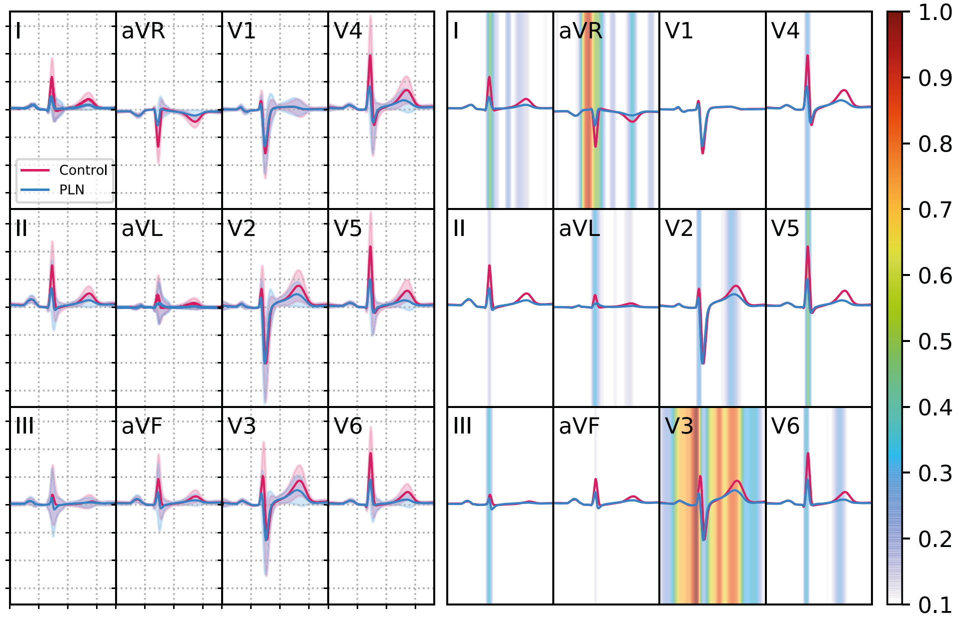


Figure 2. *Output of the Guided Grad-CAM visualization algorithm for all PLN mutation carriers and their controls.* ¬Left: Mean of temporally normalized median 12-lead ECGs of both the PLN mutation carriers (blue) and control patients (red) with their respective standard deviations. Right: the same median ECG beat with the Guided Grad-CAM output of the DNN superimposed to indicate the importance of a specific temporal segment for the classification of the DNN. The colormap represents the proportion of patients where that region was important (i.e. had a Guided Grad-CAM value above the threshold). Guided Grad-CAM: Guided Gradient Class Activation Mapping, PLN: phospholamban.

Baseline logistic regression performance

The discriminative performance (by c-statistic) of the baseline logistic regression model was 0.84 (95% CI 0.73-0.92) in the test set. The most important predictor of the PLN mutation was the presence of low QRS voltage, followed by left precordial inverted T-waves. No significant effect of age, gender or right precordial negative T-waves was found.

Deep neural network performance

The cross-validated mean c-statistic, sensitivity, specificity and F2 score obtained in the training dataset were 0.86±0.07, 0.73±0.13, 0.91±0.04 and 0.56±0.05, respectively. The c-statistic of the DNN, trained on the complete training dataset, was 0.95 (95% CI 0.91-0.99) in the independent test set. The mean ECG beats for the complete dataset with a superimposed Guided Grad-CAM visualization can be found in **Figure 2**. **Figure 3** shows a representative example of a mutation carrier and a control subject that shows similar pre-established features (low QRS voltage and inverted T-waves) but is correctly identified by the DNN.

Feature detection

Based on the Guided Grad-CAM maps (**Figure 2**), we identified the following six most prevalent combined ECG segments: (i) R-waves in V2/V3 (58-99%), (ii) PR-interval (98%), (iii) T-waves in V2/V3 (36-89%), (iv) R-waves in I/aVL/V4-V6 (34-59%), (v) R-waves in II/III/aVF (22-46%) and (vi) T-waves in I/aVL/V6 (22-36%). Figure 4 shows correlation between the Grad-CAM maps and the human interpretation, on an individual level.

After inspection of the median beat and its standard deviation at these locations, the following most prevalent features per region were extracted from the ECG and added to the baseline logistic regression model: (i) maximum R-wave amplitude in V3, (ii) PR interval, (iii) T-wave peak voltage in V3, (iv) maximum R-wave amplitude in V6, (v) maximum R-wave amplitude in III and (vi) T-wave peak voltage in I.

The updated logistic regression model's c-statistic was 0.91 (95% CI 0.83 - 0.97). The significantly associated baseline variables low QRS voltage and inverted left precordial T-waves remained significant in the updated model. The newly identified features were maximum R-wave amplitude in V3 and V6, the T-wave amplitude in I and V3 and the PR interval. The updated model had a better fit than the baseline model with an AIC of 388, compared to 461 for the baseline model (LRT p<0.001). The performance measures of all three models are shown in **Table 2**. The odds ratios of the variables in the baseline and updated models are appreciated in **Table 3**. The summary measures for the quantitative translations of the newly identified features, that are added to

Table 1. Baseline demographics and electrocardiogram characteristics of all patients and patients in the training and test splits, stratified by phospholamban mutation carriers and their matched controls. PLN: phospholamban

	Overall		**Train**		**Test**	
	Controls	***PLN***	***Controls***	***PLN***	***Controls***	***PLN***
n	1720	86	1380	69	340	17
Age, years, mean (SD)	44 (15)	44 (15)	44 (14)	44 (14)	42 (16)	42 (17)
Female sex, n (%)	1040 (61)	52 (61)	820 (59)	41 (59)	220 (65)	11 (65)
PR interval, ms, mean (SD)	151 (24)	162 (28)	151 (24)	161 (27)	149 (20)	164 (34)
QRS interval, ms, mean (SD)	93 (15)	93 (19)	94 (15)	93 (18)	93 (15)	94 (20)
QTc interval, ms, mean (SD)	422 (29)	429 (40)	422 (29)	427 (39)	423 (26)	434 (45)
Maximum voltage extremity leads, mV, mean (SD)	1.2 (0.38)	0.79 (0.43)	1.2 (0.38)	0.81 (0.45)	1.2 (0.40)	0.72 (0.39)
Maximum voltage precordial leads, mV, mean (SD)	2.2 (0.80)	1.8 (0.74)	2.2 (0.77)	1.8 (0.75)	2.3 (0.89)	1.5 (0.69)
Low QRS voltage, n (%)	41 (2.4)	31 (36)	27 (2.0)	22 (32)	14 (4.1)	9 (53)
T-wave morphology, n (%)						
Aspecific abnormalities	66 (3.8)	27 (31)	48 (3.5)	20 (29)	18 (5.3)	7 (41)
Inverted in the extremity leads	33 (1.9)	14 (16)	25 (1.8)	12 (17)	8 (2.4)	2 (12)
Inverted in the right precordial leads	34 (2.0)	10 (12)	28 (2.0)	9 (13)	6 (1.8)	1 (5.9)
Inverted in the left precordial leads	49 (2.8)	22 (26)	41 (3.0)	20 (29)	8 (2.4)	2 (12)

the baseline logistic regression model, are shown in **Table 4**.

Subgroup analyses

Performance was higher for symptomatic than presymptomatic patients, with c-statistics of 0.97 (95% CI 0.95-0.98) and 0.95 (95% CI 0.91-0.98), respectively. Sensitivity was 86% for symptomatic patients (n = 75) and 64% for presymptomatic patients (n = 11), at a similar specificity of 94%. Guided Grad-CAM maps showed a difference in features between symptomatic and asymptomatic patients, where the prolonged PR-interval, attenuated R- and T-wave in V3 and attenuated T-wave in V6 were more important in presymptomatic patients while the overall attenuated R-waves were more prominent in symptomatic patients. The Guided Grad-CAM maps for presymptomatic and symptomatic patients are shown in Supplemental **Figures 2 and 3**.

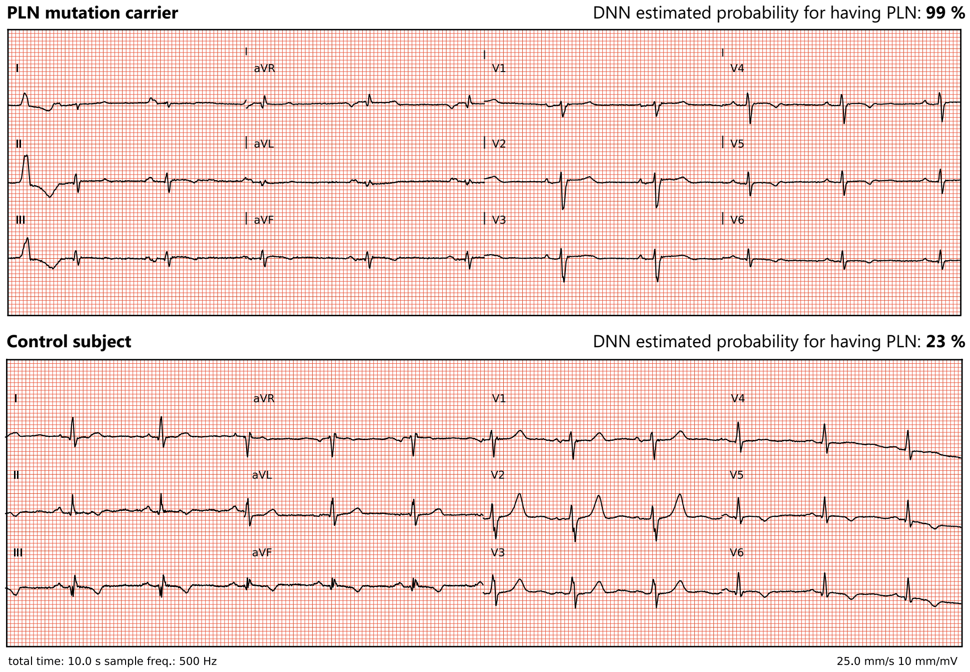

PLN mutation carrier
DNN estimated probability for having PLN: 99 %
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
Control subject
DNN estimated probability for having PLN: 23 %
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
total time: 10.0 s sample freq.: 500 Hz
25.0 mm/s 10 mm/mV

Figure 3. ***Representative examples of an ECG of a PLN mutation carrier (top panel) and a control subject (bottom panel) with their respective DNN probability score for having the PLN mutation.*** Note that the control subject ECG also exhibits the established PLN features (low QRS voltages and the presence of inverted T-waves in the left precordial leads) but is classified correctly as a control subject. The features as detected by the DNN (decreased R- and T-wave voltage in V3) can be used to distinguish the PLN mutation carriers and control subject. DNN: deep neural network, PLN: phospholamban.

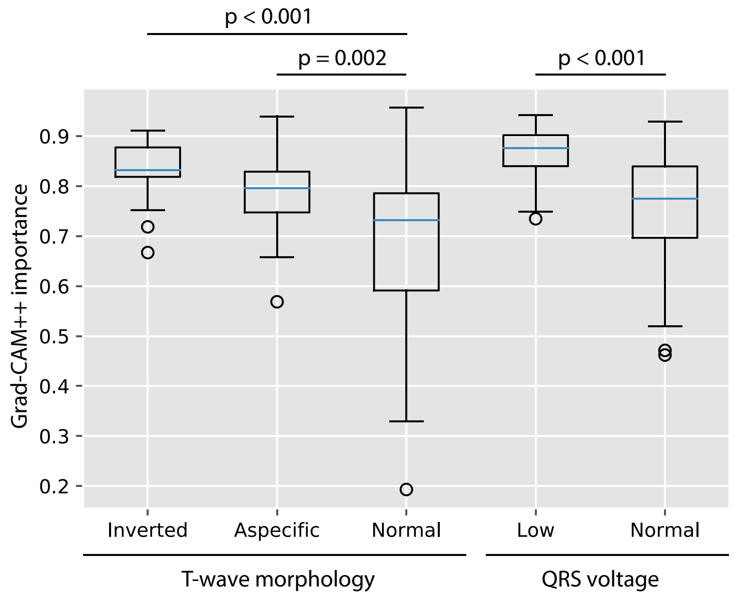


Figure 4. ***Relationship of the mean Grad-CAM++ importance value of the T-wave area with the human interpretation of the T-wave and of the QRS-complex area with the human classification of low QRS voltage in PLN patients.*** In the temporally aligned Grad-CAM++ curves, the mean is taken for the area of the QRS-complex and the T-wave. A boxplot of the importance values (between 0 and 1) of that region for the network for predicting PLN are shown in relationship with the human interpretation of the corresponding segments. Grad-CAM++: gradient class activation mapping ++.

Table 3. *Odds ratios and 95% confidence intervals for the variables in the baseline and updated logistic regression models for prediction of PLN mutation carrier status in the training dataset.* The baseline model includes the currently established electrocardiogram features of phospholamban mutation carriers. Features identified by the deep neural network, translated into quantitative measures, are added in the updated model for validation of these features.

	Baseline	**Updated**
Age per year increase	0.84 [0.73-0.92]	0.91 [0.83-0.97]
Male sex	0.96 [0.55 – 1.7]	1.17 [0.61 – 2.2]
Low QRS voltage	16.6 [8.1 – 34]	3.7 [1.5 – 9.0]
Left precordial inverted T-waves	7.4 [2.9 – 18]	4.5 [1.6 – 12]
Right precordial inverted T-waves	1.3 [0.37 – 4.2]	1.3 [0.33 – 4.9]
Right precordial inverted T-waves	-	0.37 [0.15 – 0.86]
Right precordial inverted T-waves	-	1.2 [1.1 – 1.3]
Right precordial inverted T-waves	-	1.13 [0.31 – 3.7]
Right precordial inverted T-waves	-	0.073 [0.0025 – 0.20]
Right precordial inverted T-waves	-	0.0013 [0.000013 – 0.06]
Right precordial inverted T-waves	-	3.5 [1.4 – 9.0]

Table 2. *Discriminatory performance of the baseline and updated logistic regressions models and the deep neural network in the independent test set.* The baseline model includes the currently established electrocardiogram features of phospholamban mutation carriers. Features identified by the deep neural network, translated into quantitative measures, are added in the updated model for validation of these features. DNN: deep neural network, CI: confidence interval.

	LOGISTIC REGRESSION MODELS		DNN
	Baseline	**Updated**	
C-statistic [95% CI]	0.84 [0.73-0.92]	0.91 [0.83-0.97]	0.95 [0.91-0.99]
Sensitivity	53%	76%	82%
Specificity	94%	93%	93%
Positive predictive value	33%	34%	37%
Negative predicitive value	98%	99%	99%

Table 4. *Summary measures of the quantitative translations of the newly identified electrocardiogram features of phospholamban mutation carriers.* Most prevalent newly identified features for predicting the phospholamban mutation, as identified by the visualizations of the deep neural network, were translated into quantitative measures and tested in the updated logistic regression model for validation. PLN: phospholamban, SD: standard deviation, IQR: interquartile range, mV: millivolt, ms: millisecond..

	Controls	**PLN**	**p-value**
R-wave voltage in V3, mV, median [IQR]	0.72 [0.47 – 1.1]	0.31 [0.19 – 0.61]	<0.001
PR interval, ms, mean (SD)	151 (24)	162 (28)	<0.001
T-wave voltage in V3, mV, mean (SD)	0.46 (0.29)	0.28 (0.29)	<0.001
R-wave voltage in V6, mV, median [IQR]	0.66 [0.46 – 0.91]	0.28 [0.15 – 0.45]	<0.001
T-wave voltage in I, mV, mean (SD)	0.25 (0.15)	0.11 (0.15)	<0.001
R-wave voltage in III, mV, me-dian [IQR]	0.38 [0.18 – 0.72]	0.22 [0.09 – 0.56]	<0.001

Discussion

In this study we demonstrate a novel DNN-based end-to-end approach that allows for detection and visualization of disease-specific ECG features. To the best of our knowledge, this is the first time DNNs have successfully been applied as an ECG feature detector, in contrast to previously developed prediction algorithms. Using a unique combination of median ECG beats and visualizations, the algorithm was able to automatically reveal established ECG features in PLN p.Arg14del mutation carriers (low QRS voltages and T-wave inversions), specify these features (R and T-wave attenuation in V2 and V3) and find novel features (increased PR-duration). Applying this promising concept in more cardiac diseases (especially rare or unknown ones) can potentially support physicians while reviewing ECGs, thereby improving ECG interpretation in daily clinical practice.

Previous literature

Several studies showed that DNNs can be used to make predictions from ECGs with a high performance.[2–4,7,8] An example is the recent study by Ko *et al.*, who developed a DNN to detect HCM, resulting in an AUC of 0.96.[4] From a clinical point of view, this network is very attractive, because this would allow the clinician to easily and automatically distinguish HCM in a screening setting. However, clinical implementation of such a network is still challenging for several reasons. Firstly, such networks are often seen as “black boxes”, and, secondly, the validity of these high-dimensional networks in external datasets is still unproven.

Similarly, we developed a DNN that recognizes ECGs of a specific patient population (i.e. PLN mutation carriers) with high diagnostic performance.[4] A different architecture was chosen, as it has an increased diagnostic performance in PLN mutation carriers and allows for more detailed visualizations. Unique to our study is the use of hard outcome data and the focus on feature detection, which may directly support clinicians with ECG interpretation in daily practice. Moreover, we show that these features can be used in a relatively simple logistic regression model, which might be easier generalizable.

Disease-specific ECG features in PLN mutation carriers

This novel approach was validated in PLN mutation carriers, because typical ECG characteristics in these subjects have been described extensively before.[10,12,14,15] PLN mutation carriers are at risk of developing an often biventricular phenotype of ARVC and/or DCM, and are typically characterized by subepicardial fibro-fatty replacement.[21] This leads to an ECG with low QRS voltages, which can be seen both in the limb leads and in precordial leads.[10,14] In addition, negative T-waves were previously described in both the right precordial leads and in the left precordial leads.[12,15]

Using this novel approach, we could correctly identify all of these previously described ECG features (**Figure 2**) and show that the network also uses the pre-established features for diagnosis (**Figure 4**). In addition, we could specify the leads in which these features are typically present. With the visualization tool, we found attenuated R-waves to be particularly present in the lateral leads I, aVL and V6, and in the right precordial leads V2 and V3. While the low voltages in these mutation carriers are often measured as QRS peak-to-peak amplitude, we observed that these low voltages were only based on R-wave attenuation, while the S-wave seemed unaltered. Furthermore, we found attenuated/inverted T-waves to be typically present in leads V2, V3 and V6 (as described previously), but also in leads I and aVL. Besides the ECG characteristics that were already identified before in PLN mutation carriers, we also found an ECG feature, the PR-interval, that was not described before in these subjects. This was confirmed in the updated logistic regression model. Interestingly, a recent meta-analysis of genome-wide association studies also showed an association between a locus in the PLN gene and PR-interval, which already suggested that PLN plays a role in atrio-ventricular conduction.[22]

In an exploratory analysis, the DNN performed well in both presymptomatic and symptomatic mutation carriers. Our approach also suggested that particular features were more important in presymptomatic mutation carriers (PR-interval and R- and T-wave attenuation in V2 and V3), as compared to symptomatic carriers. This might indicate that our approach can be used in subgroups who are in different stages of a disease, to gain knowledge on the sequence in which ECG abnormalities naturally occur. In particular for PLN mutation carriers, it is important to gain knowledge on the first electrical changes, because this may improve early screening and risk stratification of presymptomatic mutation-carrying family members.

Employed methodology

The use of DNN for the analysis of data generally requires large amounts of balanced data but the group of PLN mutation carriers studied in this investigation contained only 86 patients. The focus on features detection instead of prediction in this paper, however, allowed the use of such small datasets, as we were able to reduce the highly dimensional DNN to a few important features. Moreover, to allow training on this extremely imbalanced dataset, while also correcting for age and sex differences between mutation carriers and controls, we applied propensity score matching.

In the present study we used ECG median beats as input for the DNN model, which allowed the network to focus on morphology rather than rhythm. The use of median beats prohibits detection of rhythm specific ECG features (e.g. premature contractions or heart rate variability) and can also not be used for detection of beat-to-beat ECG variations. To our best knowledge, this is the first study in which median beats are used for deep learning.

Limitations

Firstly, although the proposed approach is feasible in small datasets, care should be taken while interpreting results derived from small cohort studies as findings may not hold up when evaluated on other cohorts. Especially the number of patients in the test set is a major limitation. To show clinical applicability of the ECG features and algorithm as described in this study, external validation studies should be performed. Secondly, for the PLN mutation carriers the clinical phenotype may be variable among mutation carriers. Therefore, it should be noted that this approach helps to visualize the most common ECG features on a group level, but important ECG features that are present in small subgroups may be missed. Subgroup analyses in more homogeneous subgroups (e.g. presymptomatic relatives) can be used to reveal important features in these specific subgroups. Thirdly, the ECGs of the control group were extracted from a large database in which additional patient specific characteristics are not available. Therefore, no comparisons or matching between both groups were possible to correct for other influencing factors. However, the case-control matching ratio of 1:20 used in this study presumably equalized the groups, and the detected features align with literature on other PLN mutation carriers. Fourthly, the conduction intervals and P-, QRS- and T-wave boundaries are based on the automated GE algorithm, which might cause inaccuracies. Boundary measurements on median beats have proven to be very accurate, however.[23] Fifthly, the proposed approach is not possible for ECGs with arrhythmias or acute ischemia, as these (temporary) conditions have a large influence on the morphology of the median beat. In this study, the algorithm is not intended to be used in these situations and only 3 patients were excluded for this reason. Finally, the visualization technique used in this paper, Guided Grad-CAM++, is one of the most frequently used techniques for fine-grained heatmaps but has limitations of its own.[6] For example, guided backpropagation might be independent on the choice of the model or data generating process.[24] Therefore, we validated the detected features in a logistic regression model and showed that Grad-CAM++ values agree with the pre-established PLN ECG features. Feature visualization in DNNs is a new and developing field and future research should focus on improving visualization techniques for DNNs and applying them in ECGs.

Future perspectives

Future studies should be conducted applying this novel approach to other less well characterized diseases, such as new genetic mutations, to discover novel ECG characteristics. The visualizations provide the end-user with feedback on the importance and location of detected ECG features. Moreover, future studies should consider elucidating the pathophysiological mechanisms of newly identified ECG features by using other experimental methods such as (non-)invasive electrophysiological mapping.

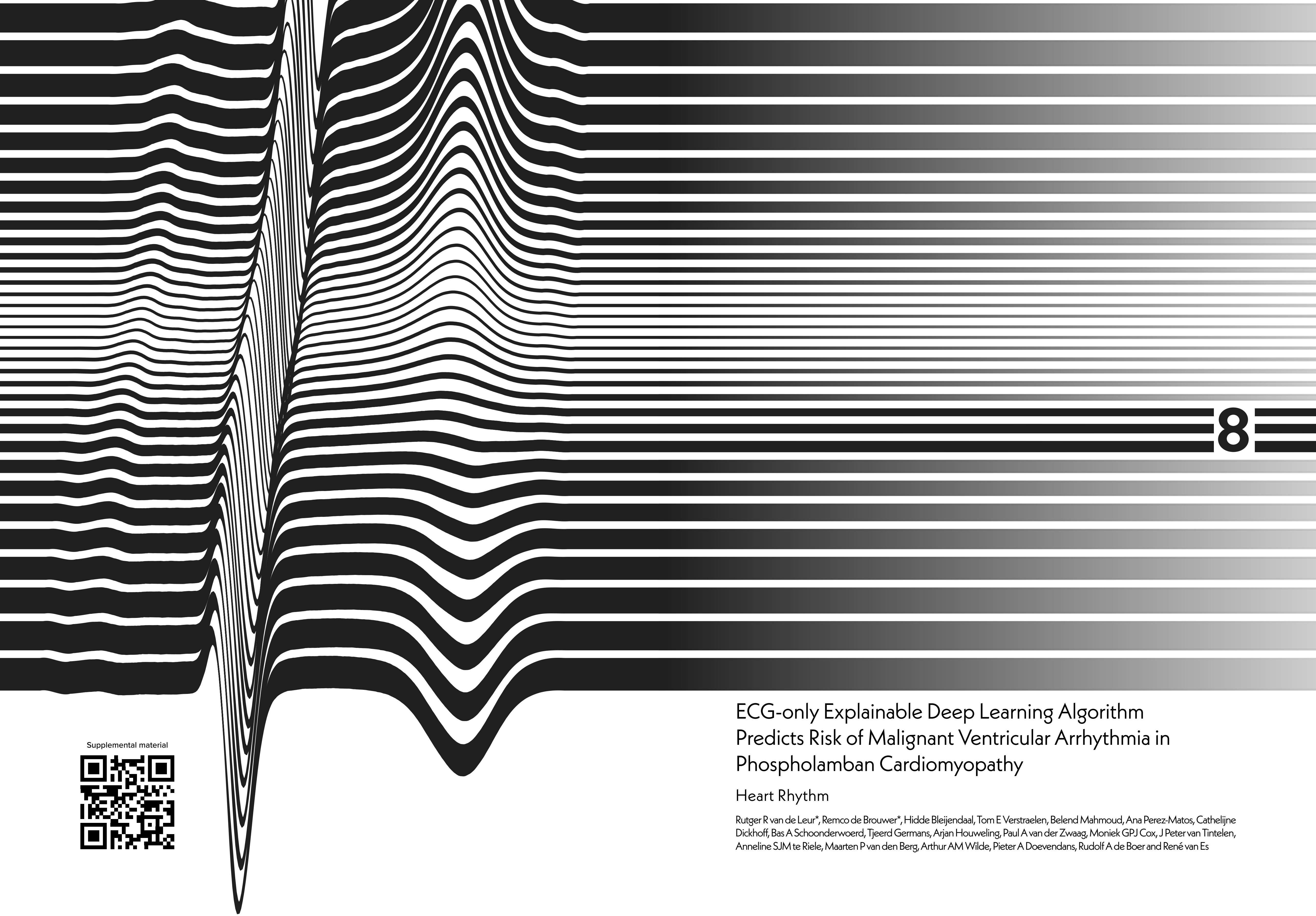
The influence of the discovered ECG features on disease penetrance in asymptomatic carriers or progression of disease in symptomatic carriers should be examined with longitudinal ECG or outcome data. Finally, combining our approach and a DNN trained on other cohorts with a focus on screening, such as family members of mutation carriers or large healthy population cohorts, might be of interest in clinical practice. Detection and visualization of possible carrier status in the ECG even before the genetic diagnosis is done could determine which family members or healthy individuals require genetic testing or follow-up.

Conclusion

This study demonstrated a novel DNN-based end-to-end approach that allows for detection and visualization of disease-specific ECG features. In a cohort of PLN p.Arg14del mutation carriers, the algorithm showed excellent diagnostic performance and revealed already established ECG features. Moreover, we were able to specify these features and to detect novel features. This novel way to use DNNs may help advance diagnostic capabilities in daily practice, especially in rare and new cardiac diseases.

REFERENCES

1. Goodfellow I, Bengio Y, Courville A. Deep Learning. The MIT Press; 2016.
2. Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med*. 2019;25.
3. Van de Leur RR, Blom LJ, Gavves E, et al. Automatic Triage of 12-Lead Electrocardiograms Using Deep Convolutional Neural Networks. *J Am Heart Assoc*. 2020;9.
4. Ko W-Y, Siontis KC, Attia ZI, et al. Detection of Hypertrophic Cardiomyopathy Using a Convolutional Neural Network-Enabled Electrocardiogram. *J Am Coll Cardiol*. 2020;75:722–733.
5. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proc IEEE Int Conf Comput Vis*. 2017;2017-Octob:618–626.
6. Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. *Proc - 2018 IEEE Winter Conf Appl Comput Vision, WACV 2018*. 2018;839–847.
7. Raghunath S, Ulloa Cerna AE, Jing L, et al. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nat Med*. 2020;26:886-891.
8. Kwon JM, Lee SY, Jeon KH, et al. Deep Learning-Based Algorithm for Detecting Aortic Stenosis Using Electrocardiography. *J Am Heart Assoc*. 2020;9:e014717.
9. Van der Zwaag PA, van Rijsingen IAW, de Ruiter R, et al. Recurrent and founder mutations in the Netherlands-Phospholamban p.Arg14del mutation causes arrhythmogenic cardiomyopathy. *Netherlands Hear J*. 2013;21:286–293.
10. Haghighi K, Kolokathis F, Gramolini AO, et al. A mutation in the human phospholamban gene, deleting arginine 14, results in lethal, hereditary cardiomyopathy. *Proc Natl Acad Sci U S A*. 2006;103:1388–1393.
11. Hof IE, van der Heijden JF, Kranias EG, et al. Prevalence and cardiac phenotype of patients with a phospholamban mutation. *Netherlands Hear J*. 2019;27:64–69.
12. Van Der Zwaag PA, Van Rijsingen IAW, Asimaki A, et al. Phospholamban R14del mutation in patients diagnosed with dilated cardiomyopathy or arrhythmogenic right ventricular cardiomyopathy: Evidence supporting the concept of arrhythmogenic cardiomyopathy. *Eur J Heart Fail*. 2012;14:1199–1207.
13. Milano A, Blom MT, Lodder EM, et al. Sudden cardiac arrest and rare genetic variants in the community. *Circ Cardiovasc Genet*. 2016;9:147–153.
14. Posch MG, Perrot A, Geier C, et al. Genetic deletion of arginine 14 in phospholamban causes dilated cardiomyopathy with attenuated electrocardiographic R amplitudes. *Hear Rhythm*. 2009;6:480–486.
15. Groeneweg JA, Van Der Zwaag PA, Olde Nordkamp LRA, et al. Arrhythmogenic right ventricular dysplasia/cardiomyopathy according to revised 2010 task force criteria with inclusion of non-desmosomal phospholamban mutation carriers. *Am J Cardiol*. 2013;112:1197–1206.
16. GE Healthcare. Marquette 12SL ECG Analysis Program Physician's Guide. Chicago, United States: 2012.
17. Pearce N. Analysis of matched case-control studies. *BMJ*. 2016;352:1–4.
18. Oord A van den, Dieleman S, Zen H, et al. WaveNet: A Generative Model for Raw Audio. 2016;1–15.
19. Franceschi J-Y, Dieuleveut A, Jaggi M. Unsupervised Scalable Representation Learning for Multivariate Time Series. 2019.
20. Paszke A, Gross S, Massa F, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems. 2019.
21. Gho JMIH, Van Es R, Stathonikos N, et al. High resolution systematic digital histological quantification of cardiac fibrosis and adipose tissue in phospholamban p.Arg14del mutation associated cardiomyopathy. *PLoS One*. 2014;9.
22. Van Setten J, Verweij N, Mbarek H, et al. Genome-wide association meta-analysis of 30,000 samples identifies seven novel loci for quantitative ECG traits. *Eur J Hum Genet*. 2019;27:952–962.
23. Willems JL, Zywietz C, Arnaud P, et al. Influence of noise on wave boundary recognition by ECG measurement programs. Recommendations for preprocessing. *Comput Biomed Res*. 1987;20:543–562.
24. Adebayo J, Gilmer J, Muelly M, et al. Sanity checks for saliency maps. *Adv Neural Inf Process Syst*. 2018;2018-Decem:9505–9515.



8

Supplemental material

ECG-only Explainable Deep Learning Algorithm Predicts Risk of Malignant Ventricular Arrhythmia in Phospholamban Cardiomyopathy

Heart Rhythm

Rutger R van de Leur*, Remco de Brouwer*, Hidde Bleijendaal, Tom E Verstraelen, Belend Mahmoud, Ana Perez-Matos, Cathelijne Dickhoff, Bas A Schoonderwoerd, Tjeerd Germans, Arjan Houweling, Paul A van der Zwaag, Moniek GPJ Cox, J Peter van Tintelen, Anneline SJM te Riele, Maarten P van den Berg, Arthur AM Wilde, Pieter A Doevendans, Rudolf A de Boer and René van Es

Abstract

Introduction

Phospholamban (PLN) p.(Arg14del) variant carriers are at risk of developing malignant ventricular arrhythmias (MVA). Accurate risk stratification allows for timely implantation of intracardiac defibrillators (ICD) and is currently performed using a multimodality prediction model. This study aims to investigate whether an explainable deep learning-based approach allows for risk prediction using only electrocardiogram (ECG) data.

Methods

A total of 679 PLN p.(Arg14del) carriers without MVA at baseline were identified. A deep learning-based variational auto-encoder, trained on 1.1 million ECGs, was used to convert the 12-lead baseline ECG into its FactorECG, a compressed version of the ECG which summarizes it into 32 explainable factors. Prediction models were developed using Cox regression.

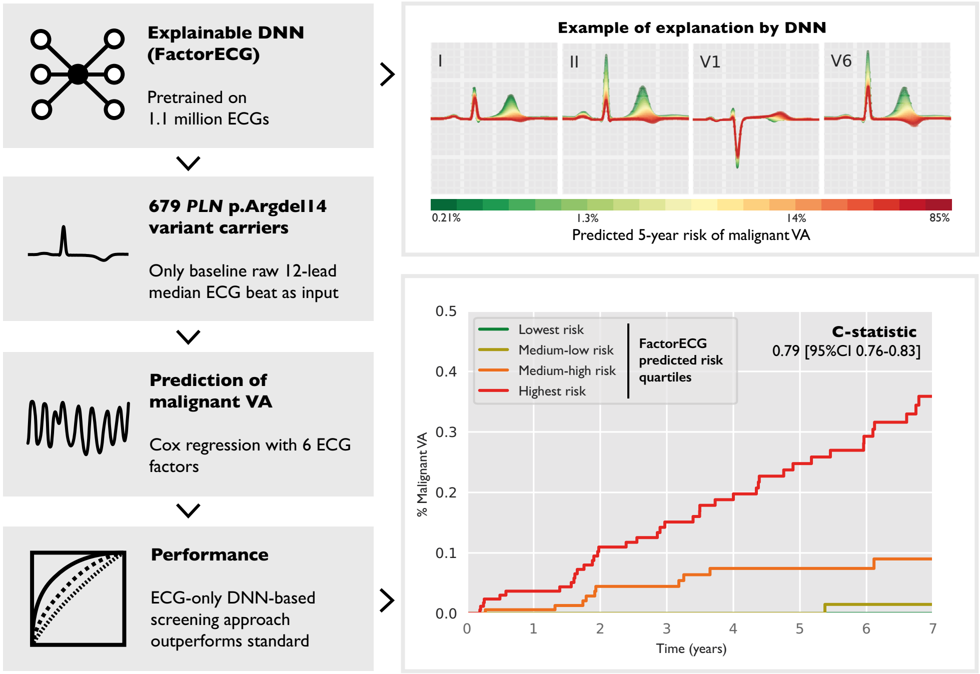
Results

The deep learning-based ECG-only approach was able to predict MVA with an AUC of 0.79 [95% CI 0.75 – 0.85], comparable to the current prediction model (AUC 0.83 [95% CI 0.79 – 0.88], p = 0.064) and outperforming a model based on conventional ECG parameters (low voltage ECG and negative T waves; 0.65 [95% CI 0.58 – 0.73], p < 0.001). Clinical simulations showed that a two-step approach, with ECG-only screening followed by a full work-up, resulted in 60% less additional diagnostics, while outperforming the use of the multimodal prediction model in all patients. A visualization tool was created to provide interactive visualizations (https://pln.ecgx.ai).

Conclusion

Our deep learning-based algorithm based on ECG data only accurately predicts the occurrence of MVA in PLN p.(Arg14del) carriers, enabling more efficient stratification of patients that need additional diagnostic testing and follow-up.

Graphical abstract



Graphical abstract. First, an artificial intelligence algorithm (variational auto-encoder) was pretrained on 1.1 million ECGs to learn the underlying continuous factors that generate the ECG (i.e., the FactorECG). In this process, the VAE learns to reconstruct ECGs as accurate as possible using only 21 continuous factors without any human input. In the training phase, median beat ECGs of 679 PLN p.(Arg14del) patients were each converted into their FactorECG. Six factors were subsequently used as input in a Cox model to predict malignant ventricular arrhythmia (MVA), and compared to the current standard. A two-step approach, where echocardiographs and Holter monitoring was only performed in the group with high predicted risk based on the FactorECG outperformed the the current multimodal model, while needing significantly less diagnostic tests. The algorithm is explainable by using the decoder to visualize the effect of the ECG factors that significantly predicted MVA on the median beat ECG morphology. DNN; deep neural network, ECG; electrocardiogram, PLN: phospholamban, VA: ventricular arrhythmia.

Introduction

Phospholamban (*PLN*) p.(Arg14del) cardiomyopathy is an inherited disease caused by a pathogenic genetic variant in the gene encoding the phospholamban protein.[1,2] This causes this protein to misfold, which in turn causes defects in the regulation of the sarcoplasmic reticulum Ca^{2+} pump.[3] This disturbance in the Ca^{2+} homeostasis of the cardiomyocyte eventually affects the composition of cardiac tissue, resulting in structural abnormalities such as cardiac fibrosis which cause, among others, distinct electrocardiographical changes (low QRS voltage in the extremity leads and negative T-waves).[4–6]

The pathogenic *PLN* p.(Arg14del) variant is associated with an arrhythmogenic or dilated cardiomyopathy characterized by progressive heart failure, malignant ventricular arrhythmias (VA) and sudden cardiac death (SCD).[7] All of these characteristics may already occur at a young age, but not all carriers of this genetic variant develop symptoms due to its incomplete penetrance. The *PLN* p.(Arg14del) genetic variant is a founder mutation in the Netherlands; its prevalence is estimated to be 1:500-1000 in large parts of the country. It has also been identified in several other countries including Spain, Greece, Vietnam, China, Japan, Canada, and the United States.[8,9] The relatively high prevalence in the Netherlands enables the compilation of uniquely large datasets.

There is no evidence-based disease-modifying therapy available for *PLN* p.(Arg14del) cardiomyopathy, though implantation of an implantable cardiac defibrillator (ICD) may improve outcomes. Currently, affected patients are treated according to general clinical guidelines, with risk score algorithms being used to identify carriers at particular risk of MVA. The latest validated risk score algorithm uses data from Holter registration, electrocardiography (ECG), echocardiography, and cardiac magnetic resonance imaging (CMR).[1]

Current prediction models use manual interpretation of the ECG, but recent reports have shown that deep neural networks (DNNs), a type of artificial intelligence (AI), can be trained to discover more complex patterns in ECGs in order to diagnose *PLN* p.(Arg14del) cardiomyopathy.[10,11] Although the need for very large data sets and the lack of interpretability were former common drawbacks of deep learning, a novel technique that uses a variational auto-encoder (VAE; the FactorECG) broadens the applicability of DNNs to much smaller data sets while also providing improved explainability (i.e. explaining which ECG morphology is associated with the outcome).[12–14] The aim of the current study is to evaluate whether this explainable deep learning based approach could be implemented to assess the risk of MVA using only ECG data, allowing clinicians to make more informed decisions regarding patient management while simultaneously reducing the total health care burden of this disease.

Methods

Study population and clinical data acquisition

All index patients and relatives carrying the *PLN* p.(Arg14del) variant were identified from a large nationwide registry. Patients that were genetically evaluated in the University Medical Center Utrecht, University Medical Center Groningen and Amsterdam University Medical Center between 2009 and 2020 were included in the current study. Clinical data were collected using chart review from the first clinical contact until last follow-up in both the university as well as non-university medical centers. Data acquired within one year of the first clinical contact and before the first event of MVA were used as baseline. Design and detailed data collection of the nationwide registry were described in detail before.[15] This study followed the Code of Conduct and the Use of Data in Health Research and was approved by local ethics and/or institutional review boards.

Electrocardiographic data acquisition

All raw 10-second 12-lead ECGs of the included patients were extracted from the MUSE ECG system (MUSE version 8, GE Healthcare, Chicago, IL, United States) from the three university medical centers and resampled to 500Hz using linear interpolation, if necessary. All ECGs were converted into median beats by aligning all primary QRS complexes (e.g. excluding premature ventricular complexes) and taking the median voltage.[16]

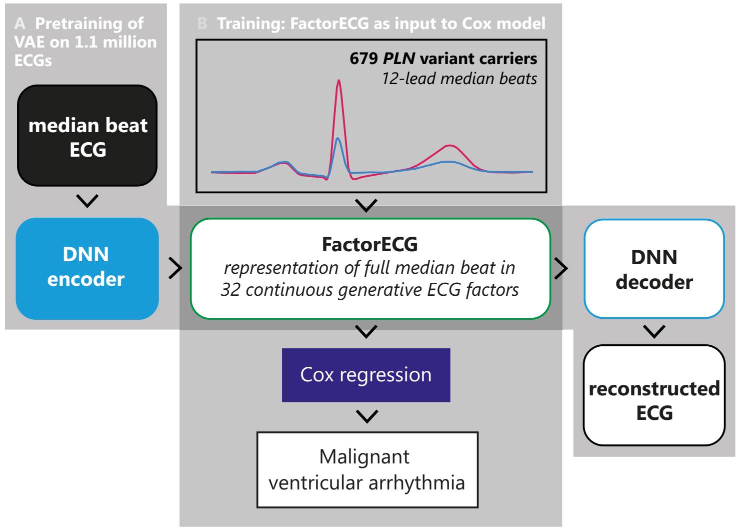
Clinical outcomes

The primary outcome of MVA was defined as done previously as a composite of sustained ventricular tachycardia (>30 seconds or terminated electrically or pharmacologically), ventricular fibrillation, appropriate ICD intervention or (aborted) sudden cardiac death.[1]

Explainable deep neural network

A recently developed approach that uses a DNN to learn explainable features from the 12-lead median beat ECG was employed. These features are explainable in the sense that the clinician obtaining an output from the DNN can visualize the ECG morphology that was associated with the outcome.[12] In this approach, a generative deep neural network, called variational auto-encoder (VAE), is used to learn the underlying generative factors of the ECG without any assumptions. This VAE consists of three parts, an encoder, the FactorECG (32 continuous factors) and a decoder and was pretrained by learning to reconstruct 1,144,331 ECGs of 251,473 patients using only the 32 factors. After training, the pretrained encoder can be used to convert any median beat ECG into its FactorECG, the distinctive set of 32 factors that represent that ECG. In the current analysis, we used these 32 continuous factors as input to the Cox and logistic regressions models (**Figure 1**).

Figure 1. ***Schematic overview of the applied deep learning-based strategy. In the pretraining phase, the variational auto-encoder (VAE) is trained a data set of 1.1 million median beat electrocardiograms (ECG) to learn to reconstruct the ECG as accurately as possible in 32 variables (the FactorECG; A).*** In the training phase, the pretrained VAE encoder is used to convert the PLN variant carrier ECGs into their FactorECG (B). Of these, six ECG factors that were associated with reduced ejection fraction in a previous study were selected, and used in a Cox regression model to predict malignant ventricular arrhythmia. The pretrained decoder can be used to visualize which ECG features were important for prediction. DNN: deep neural network; ECG: electrocardiogram; VAE: variational auto-encoder.


A Pretraining of VAE on 1.1 million ECGs
B Training: FactorECG as input to Cox model
median beat ECG
679 PLN variant carriers
12-lead median beats
DNN encoder
FactorECG
representation of full median beat in 32 continuous generative ECG factors
DNN decoder
Cox regression
reconstructed ECG
Malignant ventricular arrhythmia

The individual ECG factors can be made explainable on both the model-, and individual patient-level. This was done on the model-level by varying the values of the factors individually between -3 and 3, while generating the median beat ECG using the decoder. As the other factors are kept constant, the individual influence of that factor on the ECG morphology can be visualized. Patient-level explanations can be obtained by investigating the FactorECG values of that specific ECG and the coefficients of the prediction model. This way, we could determine which factors were important in a specific patient to make the prediction. Interactive visualizations of the model are available on https://pln.ecgx.ai. The architecture and training procedures for the FactorECG have been described in detail before.[12]

Predictor variables

Three different sets of predictors were evaluated and compared. Two ECG-only predictor sets, one baseline with the accepted conventional ECG criteria (number of leads with negative T-waves and presence of low QRS voltage) and one with the standardised FactorECG values, were compared to the predictor set used in the multimodal prediction model (the two conventional ECG criteria, number of premature ventricular complexes on Holter and left ventricular ejection fraction (LVEF)).[1] Given the low number of events in this cohort, we selected the six ECG factors most associated with a reduced LVEF in a previous study to achieve at least 10 events per predictor.[12] Detailed definitions of all predictor variables were described before.[1]

Clinical utility

Potential consequences of using the different prediction models to determine ICD implantation with different thresholds for 5-year risk of MVA were explored. For each model and threshold false positives (ICD but no MVA), true positives (ICD and MVA), false negatives (no ICD but MVA) and true negatives (no ICD and no MVA) were calculated.

In addition to the three predictor sets, we evaluated a two-step approach where only patients with a high predicted risk using the ECG-only FactorECG model were referred for additional diagnostics. In that subgroup, we simulated that an echocardiogram and 24h Holter monitoring was performed and if a carriers had an LVEF below 50% or more than 500 PVCs/24 hours on Holter monitoring an ICD was implanted. The risk threshold for additional diagnostics was chosen at the best trade-off of positive and negative predictive value in the current cohort.

Statistical analysis

Multivariable Cox proportional hazards models were used to evaluate the effect of the three different predictor sets on the risk of MVA, while taking the time-to-event into account. For all models, the proportional hazards assumption was verified and non-linear relationships were investigated using natural cubic splines. Multivariable hazards ratios (HR) were reported to investigate the effect of the different predictors on MVA. As the ECG factors were standardized, the HR was also used as a measure of importance for the individual ECG factors. Model fit was assessed and compared using Akaike's Information Criterion (AIC).

As a result of the retrospective design, there were missing values in some predictor variables. Missing data was considered missing at random and multiple imputation using chained equations was performed (using all characteristics from **Table 1** and the prespecified six ECG factors). Given a mean proportion of missing values of approximately 30%, we generated 30 imputed datasets.[17] Results on the imputed datasets were pooled using Rubin's rules.

Internal validation of the discriminatory performance (as measured by Harrell's C-statistic) was performed using a bootstrap-based optimism estimation technique. Here, all model development steps (including multiple imputation and pooling using Rubin's rules) were repeated on 500 bootstrap samples.[18] Each new pooled model was tested on the original data and the optimism was defined as the mean difference between the C-statistic in the original and bootstrapped datasets. This value is subtracted from the apparent performance measure (i.e. the C-statistic in original data from a model fitted on the original data).[19] These optimism-corrected measures have been shown to be an unbiased estimate of the generalizability of the model, without losing any data for training.[20] The bootstrap samples were also used to determine the 95% confidence intervals (CIs) around the C-statistic. Permutations tests were used to compare the C-statistic from the different predictors sets. In addition, a net reclassification improvement (NRI) was computed.[21] The NRI, sensitivity, specificity, positive and negative predictive values were derived at three different prespecified clinically used probability cutoffs: 5%, 7.5% and 10%.

Baseline characteristics were expressed as mean ± standard deviation

(SD), or median with interquartile range (IQR), where applicable. All statistical analyses were performed using Python version 3.9. The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis Statement for the reporting of diagnostic models was followed, where applicable.[22]

	CHARACTERISTIC	MISSING, N (%)	OVERALL (N = 679)
Patient demographics	**Age (years), median [IQR]**	0 (0)	42 [27 – 55]
	Male sex, n (%)	0 (0)	294 (43)
	Proband, n (%)	0 (0)	113 (17)
History	**1st degree family member with MVA, n (%)**	0 (0)	91 (13)
	NYHA class > 1, n (%)	0 (0)	62 (9.1)
Electrocardiography	**Ventricular rate (bpm), median [IQR]**	260 (40)	70 [62 – 81]
	PR duration (ms), median [IQR]	260 (40)	146 [134 – 164]
	QRS duration (ms), median [IQR]	260 (40)	86 [80 – 98]
	Corrected QT duration (ms), median [IQR]	260 (40)	411 [398 – 430]
	Number of leads with negative T-waves, n (%)	120 (18)	1 [0 – 2]
	Low voltage ECG, n (%)	61 (9)	95 (15)
	NSVT on Holter, n (%)	0 (0)	67 (10)
	24 h PVC count >500, n (%)	273 (40)	125 (31)
Imaging	**LVEF, median [IQR]**	224 (33)	54 [48 – 60]
	RVEF, median [IQR]	146 (22)	65 [50 – 65]
	MRI LGE, n (%)	417 (61)	77 (29)
Outcomes	**MVA, n (%)**	0 (0)	72 (10)
	Duration of follow-up (years), median [IQR]	0 (0)	4.3 [1.7 – 7.4]

Table 1. ***Baseline characteristics of the study population. IQR: interquartile range, LGE: late gadolinium enhancement, LVEF: left ventricular ejection fraction, MRI: magnetic resonance imaging,*** NSVT: non-sustained ventricular arrhythmia, NYHA: New York Heart Association, RVEF: right ventricular ejection fraction, PVC: premature ventricular complex, VA: ventricular arrhythmia.

Results

Study population

The total cohort consisted of 1067 *PLN* p.(Arg14del) variant carriers. After exclusion of patients with MVA at baseline (n = 65, 6%), patients without follow-up data (n = 221, 21%) and patients without any baseline test in the participating centers (n = 102, 9.6%), 679 *PLN* carriers were included in the analysis. Raw 12-lead ECG waveforms within one year of first presentation were available for 472 (70%) patients and of these 419 (89%) were of adequate quality. Performance of the pretrained VAE for the included ECGs was good, with a Pearson correlation coefficient between original and reconstructed ECG of 0.89. A total of 72 patients (10%) reached the primary outcome of MVA during a follow-up of 4.3 years [IQR 1.7 – 7.4]. The composite consisted of appropriate ICD therapy, sustained VT/VF and SCD in 37, 26 and 9 patients, respectively. Additional baseline characteristics are shown in **Table 1**.

Model performance

The baseline ECG-only model (consisting of the number of negative T-waves and low QRS voltage as predictors) predicted MVA with an optimism-corrected c-statistic of 0.65 [95% CI 0.58 – 0.73]. The FactorECG model (consisting of 6 ECG factors) outperformed the baseline model with an optimism-corrected c-statistic of 0.79 [95% CI 0.75 – 0.85] ($p < 0.001$) and was comparable to the multimodal prediction model (optimism-corrected c-statistic 0.83 [95% CI 0.79 – 0.88] ($p = 0.064$). The overall NRI for the FactorECG model compared to the baseline ECG-only was 32 [95% CI 14 – 51], with 44% [95% CI 28 - 69] more patients with MVA correctly moved upwards to the group with a risk over 7.5%. When comparing the FactorECG model to the multimodal prediction model, the NRI was 0.06 [95% CI -0.05 – 0.16], with 4.5% [95% CI -4.2 – 14.1] more patients with MVA moved upwards to the group with a risk over 7.5%. This indicates that the

FactorECG model identifies more patients with MVA than the baseline ECG-only model, without missing cases compared to the multimodal model. An overview of the NRI, sensitivity, specificity, positive and negative predictive values at different probability thresholds for all predictor sets can be found in **Table 2**.

Table 2. ***Prognostic performance measures for the different predictor sets (A-C) at three different probability cut-offs and the two-step approach (D).*** The net reclassification improvement for predictor set B was computed in comparison to the predictor set A and for predictor set C in comparison to predictor set B. Se: sensitivity, sp: specificity, PPV: positive predictive value, NPV: negative predictive value, NRI: net reclassification improvement, NRIe: net reclassification improvement for patients with an event, NRIne: net reclassification improvement for patients without an event. *Statistically significant.

	A. Baseline ECG-only			B. FactorECG			C. Multimodal			D. 2-step
Threshold	5%	7.5%	10%	5%	7.5%	10%	5%	7.5%	10%	
Se	80	52	45	100	92	82	95	90	78	90
Sp	36	75	83	48	63	73	62	71	78	75
PPV	7	12	14	11	14	16	13	16	18	18
NPV	97	96	96	100	99	98	100	99	98	99
NRI	Ref	Ref	Ref	32*	32*	38*	14*	6	-4	NA
NRIe	Ref	Ref	Ref	22*	44*	48*	-2	-5	-9	NA
NRIne	Ref	Ref	Ref	10*	-12*	-10*	16*	11*	5*	NA

Most important predictors in the FactorECG model were F_1 (inferolateral ST-segment and T-wave morphology, HR 0.61 [0.40 – 0.91]) and F_5 (inferolateral negative T-waves, HR 2.03 [1.34 – 3.07]), while in the multimodal model LVEF (HR 0.96 per 1% increase [95% CI 0.94 – 0.98]) and 24h PVC count (HR 1.33 per 1 log increase [95% CI 1.16 – 1.55]) were most predictive. Hazard ratios and confidence intervals for all prediction models are shown in **Table 2**. When using the three prediction models to stratify carriers in four quartiles using their predicted 5-year risk of MVA, a clear distinction in risk between the groups can be observed for the FactorECG and multimodal model (**Figure 2A-C)**. In the lowest two risk groups, almost no events are observed for these models, while the baseline ECG-only model is not able to distinguish groups without events.

Clinical applicability

Different scenarios with varying thresholds for the 5-year predicted risk of MVA to determine which patients should get an ICD implantation **(Figure 3**) were investigated. At a clinically accepted 5-year risk threshold of 5% (1% risk per year), the baseline ECG-only model performed worst with a sensitivity of 80% and specificity of only 36%. The FactorECG model outperformed the baseline model with a sensitivity of 100% and specificity of 48%, while the multimodal model had a higher specificity of 62% at the cost of a lower sensitivity of 95%. A similar trend is observed at the higher 5-year risk thresholds of 7.5% and 10%, where significantly less ICDs are implanted but with more false negatives.

Next to the implementation of the models alone, a more clinical applicable two-step approach was investigated (**Figure 4**). With this simulated approach, all patients first get an ECG, and then only in the high-risk patients as predicted by the FactorECG model, echocardiography and Holter monitoring data is needed. A threshold to determine which patients were high-risk of 7.5% was used as this provided the best trade-off of positive and negative predictive value (i.e. referring the least amount of patients without

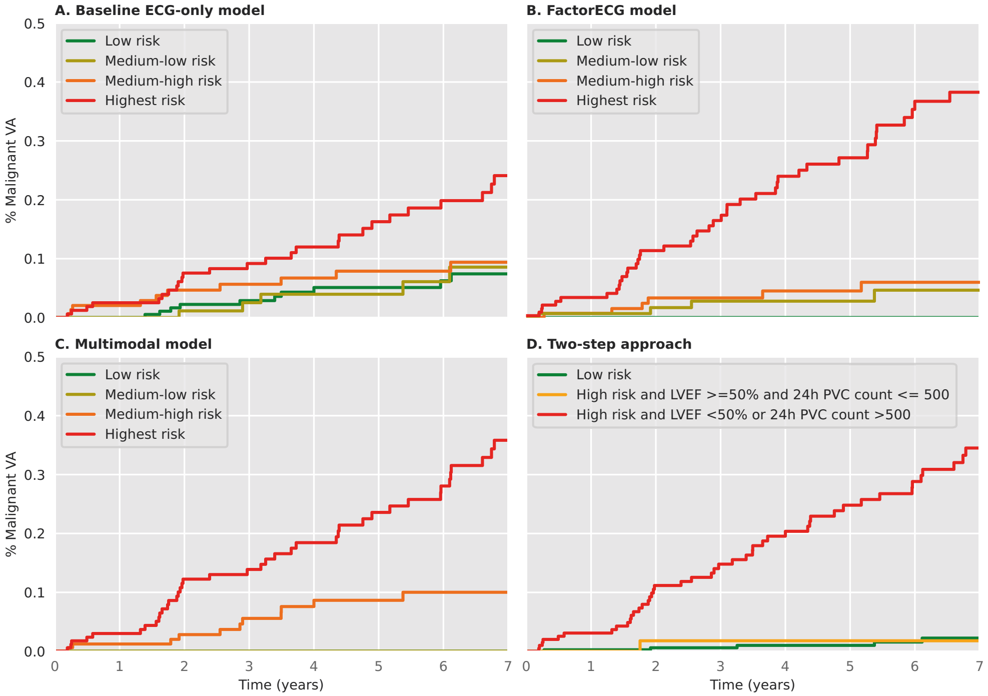


Figure 2. ***Kaplan-Meier plots for the different predictor sets (A-C) and two-step approach (D). For the prediction models (A-C), the 5-year predicted risk of MVA is split in four quartiles (risk groups).*** For the two-step approach, an approach was simulated where only patients with a high predicted risk (>7.5%) using the FactorECG model were referred for additional diagnostics. In that subgroup, we simulated that an echocardiogram and 24h Holter monitoring was performed.

missing too many patient with MVA). When applying this risk threshold, only 40% of patients need to be referred. In this referred group we simulated an ICD implantation when either the LVEF was below 50% or more than 500 PVCs/24 hours on Holter monitoring were recorded. This two-step approach outperformed all other models with a sensitivity of 90% and specificity of 75%.

Model explainability

F_1 (inferolateral ST-segment and T-wave morphology) and F_5 (inferolateral negative T-waves) were significantly associated with the risk of MVA during follow-up, with more negative values corresponding to a higher risk for F_1 and more positive values for F_5. Both these factors represent the shape of the inferolateral ST-segment and T-wave and are significantly correlated with each other in this population (Pearson r = -0.39, p < 0.001). The factor traversals of a combined change in F_1 and F_5 showed that this combination represents a change in ECG morphology from normal QRS voltage and repolarization towards lower QRS voltage and inferolateral symmetrical negative T-waves without any ST-deviation (**Figure 5**). Interestingly, the effect of this morphological change was non-linearly related with the predicted 5-year risk of MVA and the risk already exceeded 5% when the T-waves are still positive (Figure 5). Other factors were not significantly correlated (Pearson r < 0.22 for all) in this population and their factor traversals are therefore shown for each factor individually in Supplementary Figures 1-3.

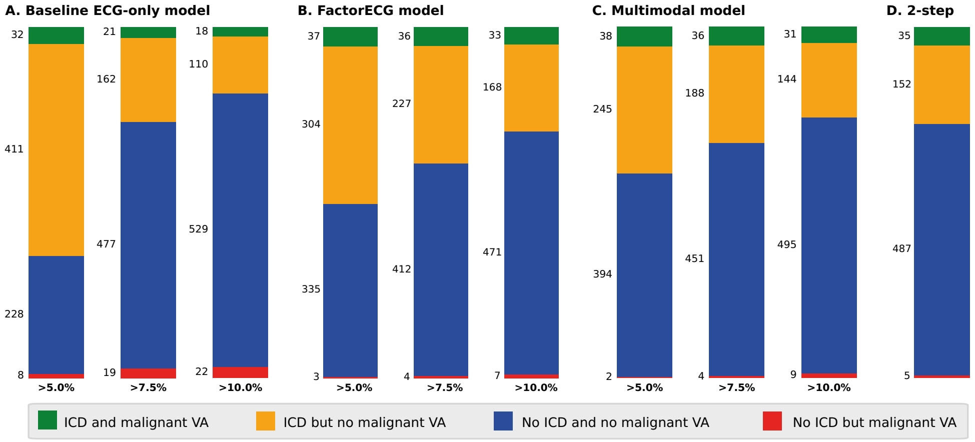


Figure 4. *Overview of the two-step approach, where an approach was simulated where all PLN variant carriers first get an ECG only.* This ECG is evaluated by the FactorECG prediction model, and only the high-risk patients are referred for additional diagnostics (echocardiography and 24h Holter monitoring). When carriers had an LVEF below 50% or more that 500 PVCs/24 hours on Holter an ICD was implanted.

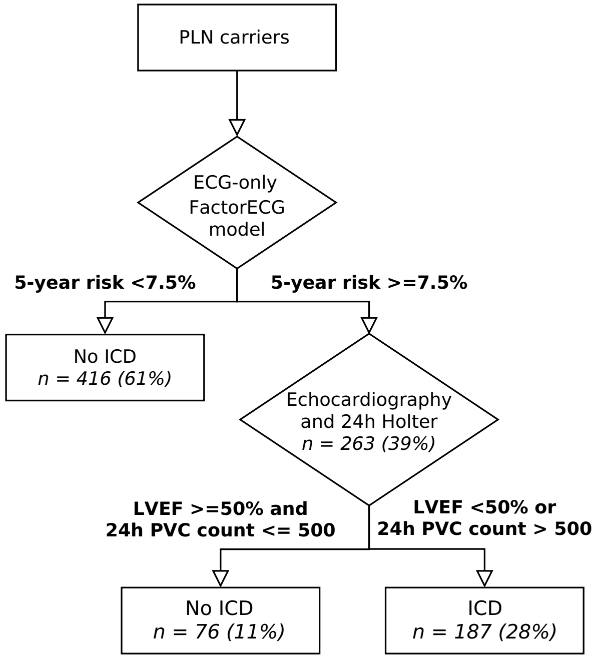


Figure 3. *Clinical utility plots for the different predictor sets (A-C) and the two-step approach (D).* The bars represent the clinical implications of using different five-year risk of malignant ventricular arrhythmia thresholds for the decision to implant an implantable cardioverter defibrillator. For the two-step approach, an approach was simulated where only patients with a high predicted risk (>7.5%) using the FactorECG model were referred for additional diagnostics. In that subgroup, we simulated that an echocardiogram and 24h Holter monitoring was performed and if a carriers had an LVEF below 50% or more that 500 PVCs/24 hours on Holter an ICD was implanted.

	HR [95% CI]	p-value
Baseline ECG-only model		
Number of leads with negative T-waves	1.12 [1.00 – 1.24]	0.03
Low QRS voltage	3.52 [2.07 – 5.97]	<0.001
FactorECG model		
Factor 1	0.61 [0.40 – 0.91]	0.01
Factor 5	2.03 [1.34 – 3.07]	<0.001
Factor 8	1.48 [0.97 – 2.25]	0.07
Factor 10	0.89 [0.62 – 1.29]	0.54
Factor 25	0.69 [0.52 – 0.93]	0.02
Factor 26	0.66 [0.44 – 1.00]	0.05
Multimodal model		
Number of leads with negative T-waves	1.10 [0.89 – 1.23]	0.10
Low QRS voltage	1.76 [0.98 – 3.20]	0.06
24h PVC count (per 1 log increase)	0.96 [0.94 – 0.98]	<0.001
LVEF (per % increase)	1.34 [1.16 – 1.54]	<0.001

Table 3. *Hazards ratios, confidence intervals and p-values for the different predictor sets evaluated in multivariable Cox proportional hazard models.*

Discussion

This study shows that an explainable deep learning-based approach using only ECG data was able to predict the risk of MVA with an optimism-corrected c-statistic 0.79 [95% CI 0.75 – 0.85] in a large cohort of PLN p.(Arg14del) carriers. Addition of echocardiographic and Holter monitoring data in the group with high predicted risk based on the FactorECG improved predictive ability further (i.e. a two-step approach), outperforming the use of the current multimodal model in all patients. Such two-step approach could allow for more efficient risk stratification of PLN p.(Arg14del) carriers, reduce the burden of monitoring visits for these carriers, and lead to a significant decrease in costs by reducing the number of visits, diagnostics and ICD implantations. Deep learning-based ECG analysis may enhance the possibilities for remote monitoring of genetic variant carriers. An online tool to convert any ECG into its FactorECG and predict prognosis in PLN patients, is available through (https://pln.ecgx.ai).

Clinical applicability and prior studies

This is the first study attempting risk stratification in carriers of the *PLN* p.(Arg14del) genetic variant using only ECG data. The current best practice in risk stratification of known *PLN* p.(Arg14del) carriers involves the use of a risk score combining structural, electrophysiological and functional parameters.[1] This multimodal algorithm has an optimism-corrected c-statistic of 0.83 [95% CI 0.79 – 0.88] in the current analysis. While an ECG-only model containing conventional ECG features of PLN cardiomyopathy (low QRS voltage and negative T-waves) was not able to reach similar predictive performance (optimism-corrected c-statistic 0.65 [95% CI 0.58 – 0.73]), the deep learning-based ECG-only model did perform comparably (optimism-corrected c-statistic 0.79 [95% CI 0.75 – 0.85]). Net reclassification analysis confirmed that the FactorECG algorithm outperformed the baseline ECG-only model at all risk thresholds, without missing patients with

MVA within 5 years compared to the multimodal algorithm (**Table 2**).

Clinically, such an ECG-only algorithm could be used in a two-step approach involving a first pass using the ECG model alone, followed by additional diagnostics in subjects deemed at-risk of MVA. If acceptable NPVs can be achieved with only ECG (possibly at home or by the general practitioner), the large burden of monitoring visits could be reduced, especially for asymptomatic carriers. Whereas the conventional ECG-only model did not reach adequate NPVs to be usable in such an approach, the FactorECG model was able to reach a NPV of 99% in 60% of the patients at a five-year risk threshold of 7.5%. As visualized in **Figure 3**, this results in more accurate risk prediction than either method alone, as well as being more accurate than using the multimodal model in all patients.

The presence of the *PLN* p.(Arg14del) genetic variant is established via genotyping of potentially affected index patients presenting with related signs and symptoms, followed by genetic cascade-screening in close family members. Both Bleijendaal and Van de Leur et al. have shown that a deep learning-based approach may aid in the diagnosis of the genetic variant in the general population as well, aiding in the identification of the aforementioned index patients.[10,11] This study builds upon their results by providing the risk stratification required for optimal management after initial diagnosis.

Explaining the AI algorithm

The term 'black box' is often used to describe models resulting from the extensive training of a machine learning algorithm.[23] These models may become too complex to be interpreted by humans using them to reach an output from a given input, which in turn may cause a level of distrust in the output.[24] Our approach provides improved explainability by allowing clinicians to visualize the influence of specific median beat ECG morphology on the predictions.[12,25] Previous studies have shown that a similar approach using the FactorECG can be used to predict risk of MVA in dilated cardiomyopathy patients and outcomes in cardiac resynchronization therapy recipients.[13,14]

Our visualizations confirm that the FactorECG prediction is mostly based on known *PLN* cardiomyopathy ECG features (e.g. reduced QRS voltage and inferolateral symmetrical negative T-waves as represented by the combinations of F_1 and F_5), as shown in Figure 5. Interestingly, it uses these features as a continuous spectrum and already predicts a risk higher than the threshold of 5% before the appearance of negative T-waves, but only with a reduced R- and T-wave height. This might explain why the model outperforms the baseline ECG-only model, as this uses binary cut-off points for QRS voltage and negative T-waves. Other ECG features shown by the visualizations are an increased PR-interval (F_8, Supplemental **Figure 1)**, rSR' in V1 with slurred S-waves inferolaterally (F_{25}, Supplemental **Figure 2**) and reduced lateral T-wave height (F_{26}, Supplemental **Figure 3**), although all borderline significant (**Table 2**). We expect that this direct input-output relationship makes using the algorithm a more attractive option to clinicians by increasing trust in the outcome. An interactive tool for explainability is available through: https://pln.ecgx.ai.

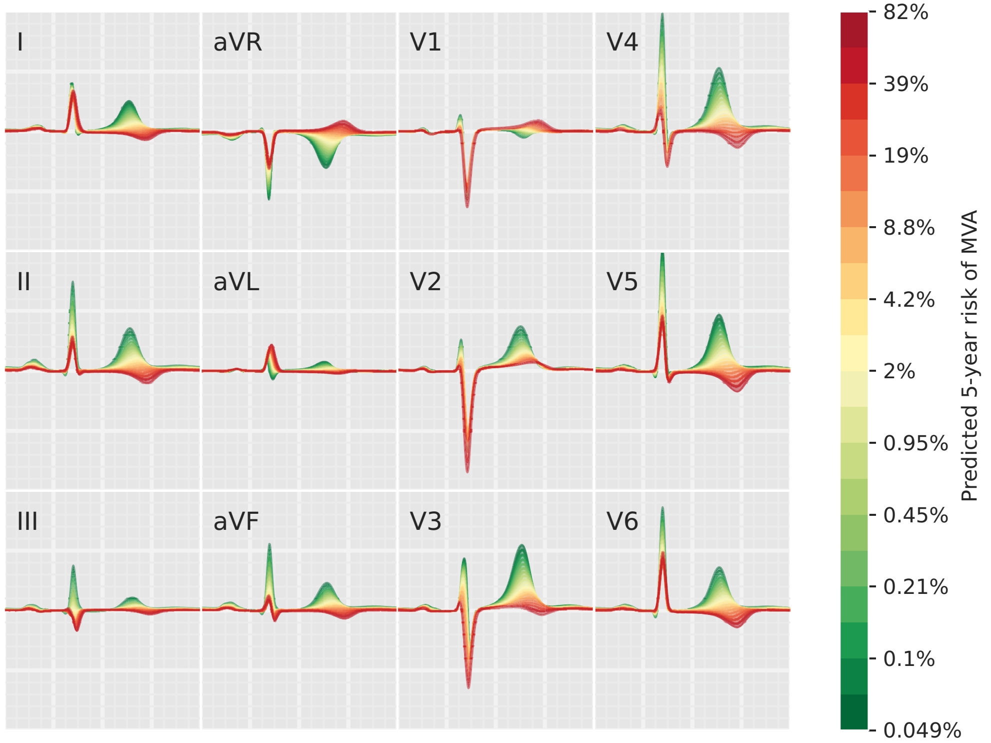


Figure 5.
Factor traversals for the two most important electrocardiogram factors to visualize ECG features that the model used to predict malignant ventricular arrhythmia (MVA). In the current plot, we varied the values for Factor 1 and Factor 5 simultaneously as these factors are strongly correlated in the current population, while keeping the other factors constant at their mean value. For each combination the five-year risk of MVA is derived using the Cox regression model and visualized. MVA: malignant ventricular arrhythmia.

Strengths and limitations

The main strength of this study is that the *PLN* registry allowed for leveraging a uniquely large cohort of deeply phenotyped *PLN* p.(Arg14del) carriers.[15] However, there are several limitations to this study. Firstly, no external validation for the prediction models or the risk thresholds in the two-step scenario analysis could be performed, as there are currently no other cohorts of *PLN* p.(Arg14del) carriers available. To minimize the risk of overoptimism, we prespecified our predictor sets before the analysis, selected a limited number of predictors in every model and performed a rigorous internal validation using a bootstrap-based resampling technique.[20] The retrospective nature of the data comes with missing values; 260 (38%) *PLN* p.(Arg14del) carriers were without a baseline raw ECG of adequate quality available for analysis. The primary outcome of MVA was defined as a composite of several endpoints, one of which was appropriate ICD intervention. Thus, appropriate ICD intervention was given the same weight as sudden cardiac death or ventricular fibrillation, similar to the current prediction model in *PLN* p.(Arg14del) variant carriers. This may result in an overestimation of the true 5-year risk of sudden cardiac death since not all appropriate ICD interventions equate cardiac arrest.

Future perspectives

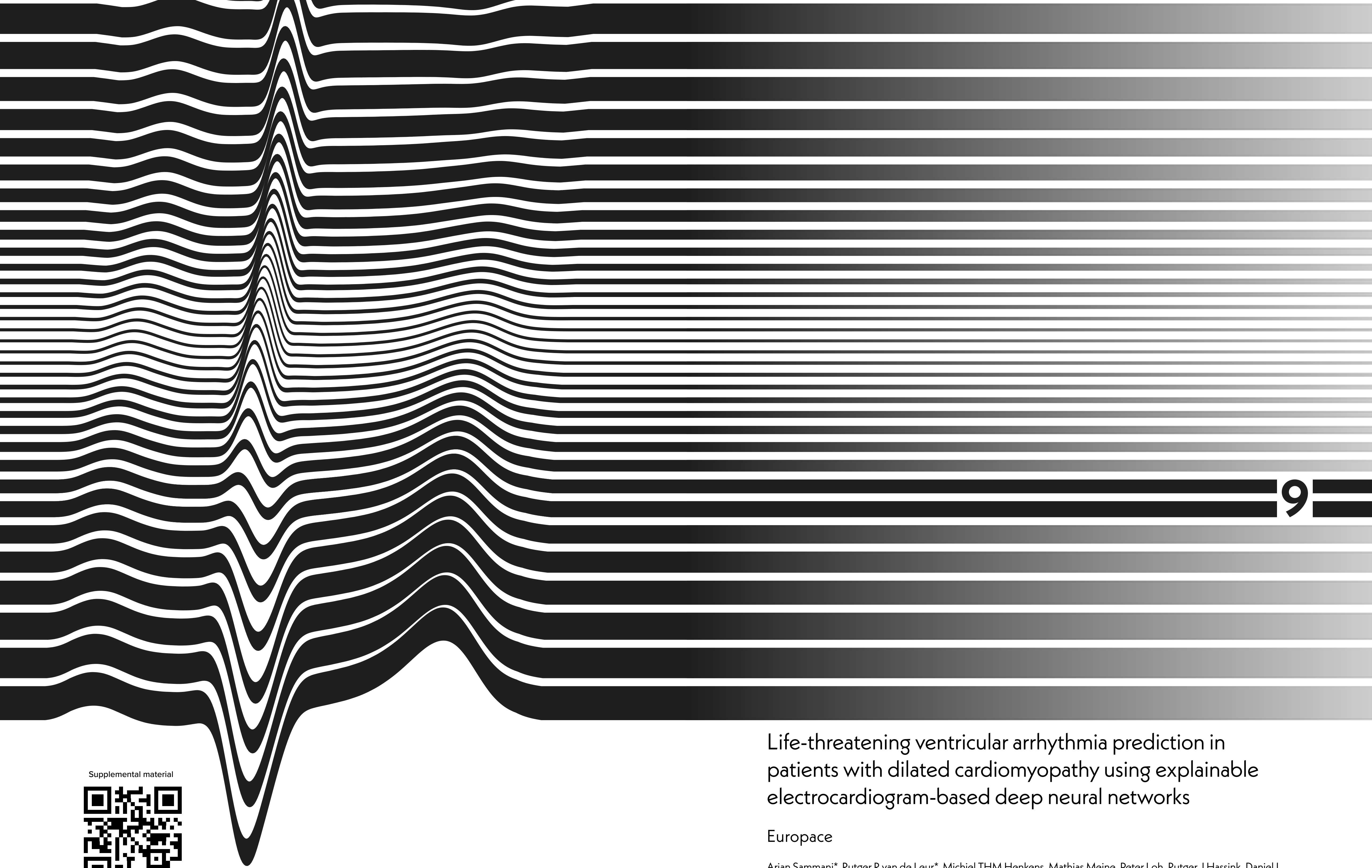
A machine learning based approach could aid in both diagnosis of the cardiomyopathy-associated variant as well as risk stratification to help clinicians more efficiently organize their healthcare system. Currently, the *PLN* p.(Arg14del) genetic variant is mainly prevalent in the Netherlands. More affected families and relatives are identified, both in the Netherlands and abroad, and the healthcare burden of diagnosis and risk assessment will rise. This is of importance due to rising healthcare costs in some nations and due to high barriers to accessing healthcare in some nations, that could manage this patient group by providing a remote solution. Moreover, the approach used in this study may also be of use for researchers studying other uncommon types of (genetic) cardiomyopathy.

Conclusion

An ECG-only explainable deep learning-based algorithm is able to predict the occurrence of MVA in *PLN* p.(Arg14del) carriers with an optimism-corrected c-statistic 0.79 [95% CI 0.75 – 0.85], which could allow for an alternative stratification relying on the ECG only, precluding additional diagnostics and follow-up. Such two-step approach could reduce the burden of monitoring visits for PLN p.(Arg14del) carriers, and lead to a significant decrease in costs by reducing the number of visits, diagnostics and ICD implantations.

REFERENCES

1. Verstraelen TE, Lint FHM van, Bosman LP, et al. Prediction of ventricular arrhythmia in phospholamban p.(Arg14del) mutation carriers–reaching the frontiers of individual risk prediction. *Eur Heart J* 2021;42:2842–2850.
2. Zwaag PA, Rijsingen IAW, Asimaki A, et al. Phospholamban R14del mutation in patients diagnosed with dilated cardiomyopathy or arrhythmogenic right ventricular cardiomyopathy: evidence supporting the concept of arrhythmogenic cardiomyopathy. *Eur J Heart Fail* 2012;14:1199–1207.
3. Rijdt WP, Tintelen JP, Vink A, et al. Phospholamban p.(Arg14del) cardiomyopathy is characterized by phospholamban aggregates, aggresomes, and autophagic degradation. *Histopathology* 2016;69:542–550.
4. Brouwer R de, Meems LMG, Verstraelen TE, et al. Sex-specific aspects of phospholamban cardiomyopathy: The importance and prognostic value of low-voltage electrocardiograms. *Heart Rhythm* 2022;19:427–434.
5. Hof IE, Heijden JF van der, Kranias EG, et al. Prevalence and cardiac phenotype of patients with a phospholamban mutation. *Neth Heart J* 2019;27:64--69.
6. Haghighi K, Gardner G, Vafiadaki E, et al. Impaired Right Ventricular Calcium Cycling Is an Early Risk Factor in R14del-Phospholamban Arrhythmias. *J Personalized Medicine* 2021;11:502.
7. Rijsingen IAWV, Zwaag PAVD, Groeneweg JA, et al. Outcome in phospholamban R14del carriers results of a large multicentre cohort study.pdf. *Circulation Cardiovasc Genetics* 2014;7:455--465.
8. Cheung CC, Healey JS, Hamilton R, et al. Phospholamban cardiomyopathy: a Canadian perspective on a unique population. *Neth Heart J* 2019;27:208--213.
9. Jiang X, Xu Y, Sun J, Wang L, Guo X, Chen Y. The phenotypic characteristic observed by cardiac magnetic resonance in a PLN-R14del family. *Sci Rep-uk* 2020;10:16478.
10. Bleijendaal H, Ramos LA, Lopes RR, et al. Computer versus cardiologist: Is a machine learning algorithm able to outperform an expert in diagnosing a phospholamban p.(Arg14del) mutation on the electrocardiogram? *Heart Rhythm* 2021;18:79–87.
11. Leur RR van de, Taha K, Bos MN, et al. Discovering and Visualizing Disease-Specific Electrocardiogram Features Using Deep Learning: Proof-of-Concept in Phospholamban Gene Mutation Carriers. *Circulation Arrhythmia Electrophysiol* 2021;14.
12. Leur RR van de, Bos MN, Taha K, et al. Improving explainability of deep neural network-based electrocardiogram interpretation using variational auto-encoders. *European Hear J - Digital Heal* 2022.
13. Wouters PC, Leur RR van de, Vessies MB, et al. ECG-based deep learning improves outcome prediction following cardiac resynchronization therapy. *European Heart Journal* 2022:in press.
14. Sammani A, Leur RR van de, Henkens MTHM, et al. Life-threatening ventricular arrhythmia prediction in patients with dilated cardiomyopathy using explainable electrocardiogram-based deep neural networks. *Ep Europace* 2022.
15. Bosman LP, Verstraelen TE, Lint FHM van, et al. The Netherlands Arrhythmogenic Cardiomyopathy Registry: design and status update. *Neth Heart J* 2019;27:480–486.
16. Healthcare G. Marquette 12SL ECG Analysis Program Physician's Guide
17. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Statist Med* 2011;30:377–399.
18. Schomaker M, Heumann C. Bootstrap inference when using multiple imputation. *Statist Med* 2018;37:2252–2266.
19. Harrell FE, Lee KL, Mark DB. Tutorial in biostatistics multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15:361--387.
20. Steyerberg EW, Harrell FE, Borsboom GJJM, et al. Internal validation of predictive models Efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 2001;54:774–781.
21. Leening MJG, Vedder MM, Witteman JCM, Pencina MJ, Steyerberg EW. Net Reclassification Improvement: Computation, Interpretation, and Controversies: A Literature Review and Clinician's Guide. *Ann Intern Med* 2014;160:122–131.
22. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med* 2015;162:55.
23. Durán JM, Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *J Med Ethics* 2021;47:329–335.
24. Petch J, Di S, Nelson W. Opening the black box: the promise and limitations of explainable machine learning in cardiology. *Can J Cardiol* 2021;38:204–213.
25. Leur RR van de, Hassink RJ, Es R van. Variational auto-encoders improve explainability over currently employed heatmap methods for deep learning-based interpretation of the electrocardiogram. *European Hear J - Digital Heal* 2022.



9

Life-threatening ventricular arrhythmia prediction in patients with dilated cardiomyopathy using explainable electrocardiogram-based deep neural networks

Europace

Arjan Sammani*, Rutger R van de Leur*, Michiel THM Henkens, Mathias Meine, Peter Loh, Rutger J Hassink, Daniel L Oberski, Stephane RB Heymans, Pieter A Doevendans, Folkert W Asselbergs, Anneline SJM te Riele and René van Es

Supplemental material

Abstract

Aims

While electrocardiogram (ECG) characteristics have been associated with life-threatening ventricular arrhythmias (LTVA) in dilated cardiomyopathy (DCM), they typically rely on human derived parameters. Deep neural networks (DNN) can discover complex ECG patterns, but interpretation is hampered by their 'black-box' characteristics. We aimed to detect DCM patients at risk of LTVA using an inherently explainable DNN.

Methods and Results

In this two-phase study we first developed a variational autoencoder DNN on more than 1 million 12-lead median beat ECGs, compressing the ECG into 21 different factors (F): factorECG. Next, we used two cohorts with a combined total of 695 DCM patients and entered these factors in a Cox regression for the composite LTVA outcome, which was defined as sudden cardiac arrest, spontaneous sustained ventricular tachycardia, or implantable cardioverter-defibrillator treated ventricular arrhythmia. Most patients were male (n=442, 64%) with a median age of 54 years [interquartile range (IQR) 44-62], and median left ventricular ejection fraction of 30% [IQR 23-39]. A total of 115 patients (16.5%) reached the study outcome. Factors F8 (prolonged PR-interval and P-wave duration, $p < 0.005$), F15 (reduced P-wave height, $p = 0.04$), F25 (increased right bundle branch delay, $p = 0.02$), F27 (P-wave axis $p < 0.005$) and F32 (reduced QRS-T voltages $p = 0.03$) were significantly associated with LTVA.

Conclusion

Inherently explainable DNNs can detect patients at risk of LTVA which is mainly driven by P-wave abnormalities.

Introduction

Patients with non-ischaemic dilated cardiomyopathy (DCM) have an estimated annual risk of life-threatening ventricular arrhythmias (LVTAs) of 4.5% and may potentially benefit from implantable cardioverter-defibrillator (ICD) implantation.[1,2] A novel risk model (DCM-SVA risk) for predicting LTVA was recently published and includes easily accessible clinical parameters, such as history of non-sustained ventricular tachycardia (VT), QRS duration and left ventricular ejection fraction (LVEF).[3] More complex electrocardiogram (ECG) characteristics such as fragmented QRS waves, heart rate variability and t-wave alternans have also been associated with LTVA, but rely on manually derived ECG parameters that remain difficult to standardize, hampering their integration into daily clinical practice.1 By using raw ECG signals and machine learning techniques, manual feature extraction is not necessary. Moreover, novel and more subtle parameters may be detected.[4]

Deep neural networks (DNN) have proven to be potent machine learning algorithms for diagnostic classification tasks using raw ECGs signals. Previous studies using DNNs on raw ECG signals in cardiomyopathies report high performance in disease classification and triaging.[5,6] However, because of the inherent lack of "explainability" of DNNs, clinical implementation remains limited.[7] Different techniques may assist in interpreting DNNs. A recently introduced pipeline for fully explainable DNNs for ECG analysis uses variational autoencoders (VAE)[8], that can compress the ECG into a lower number of explanatory and independent generative factors (factorECG), which can subsequently be used in interpretable algorithms (such as Cox regression).[9]

In this study, we aimed (i) use an inherently interpretable DNN for predicting potentially LTVA based on ECGs in patients with non-ischemic DCM, assess its added value above conventional ECG parameters and current guidelines, and (ii) interpret the model by visualizing pivotal ECG features.

Methods

Study participants

In this retrospective cohort study, we included consecutive adult patients with DCM as defined by the European Society of Cardiology (ESC) guidelines were included from the UMCU and MUMC+.[2] Only patients with a baseline non-paced 12-lead ECG acquired before Left Ventricular Assist Device (LVAD) implantation or Heart Transplantation (HTx) were eligible. Patients with a cardiac resynchronisation therapy (CRT) were excluded, as it positively affects reverse remodelling which may reduce arrhythmias.[10] This study was conducted in accordance with the principles laid out in the Declaration of Helsinki and in line with guidelines provided by ethics committees and national GDPR legislature. The participants from the UMCU cohort were included using the opt-out procedure. The UMCU cohort was exempt from the Medical Research Involving Human Subjects Act (WMO) as per judgement of the Medical Ethics Committee (18/446 and 19/222 UMCU, the Netherlands) including the requirement for informed consent. The participants of the Maastricht cohort signed informed consent at enrolment.

Data acquisition

For all subjects, the ECG closest to the date of first presentation was obtained which was considered "baseline" for the purpose of this study. The median time between diagnosis and ECG was 0 [IQR 0-28] days. All ECGs were exported from the MUSE ECG system (version 8; GE Healthcare, Chicago, IL, USA) in raw voltage format. The recordings were made using a General Electric MAC V, 5000 or 5500 device and acquired at either 250 or 500 Hz. Resampling to 500 Hz was performed via linear interpolation and transformation into 1.2-second median beats was achieved by aligning all QRS-complexes of the same shape (e.g., excluding premature ventricular complexes) and taking the median voltage to generate a representative P-QRS-T complex. Echocardiographic measurements were extracted from the electronic health record using methods described before.[11]

Pretraining and explainability of the variational autoencoder

A two-phase approach was used in this study, where a VAE was first pretrained on the complete UMCU ECG dataset, and them used in the training step to find associations with LTVA (**Figure 1**). VAEs are unsupervised deep learning encoder-decoder convolutional neural networks that are optimized to reconstruct their training data with a lower-dimensional representation (i.e., using less data) than the original training data (in this case ECGs). The current VAE network is enforced with a specific function to reach maximum disentanglement of lower-dimensional representation (i.e., to produce generative factors in the ECG that operate independently: the FactorECG).[12] Resting 12-lead 10-second ECGs of 251,473 unique patients (1,114,331 ECGs) were exported from the UMCU ECG system and used for pretraining of the VAE. In a prior study, the optimal number of dimensions was found to be 21, considering the trade-off of good reconstruction disentanglement and encoding for visible ECG abnormalities.[8]

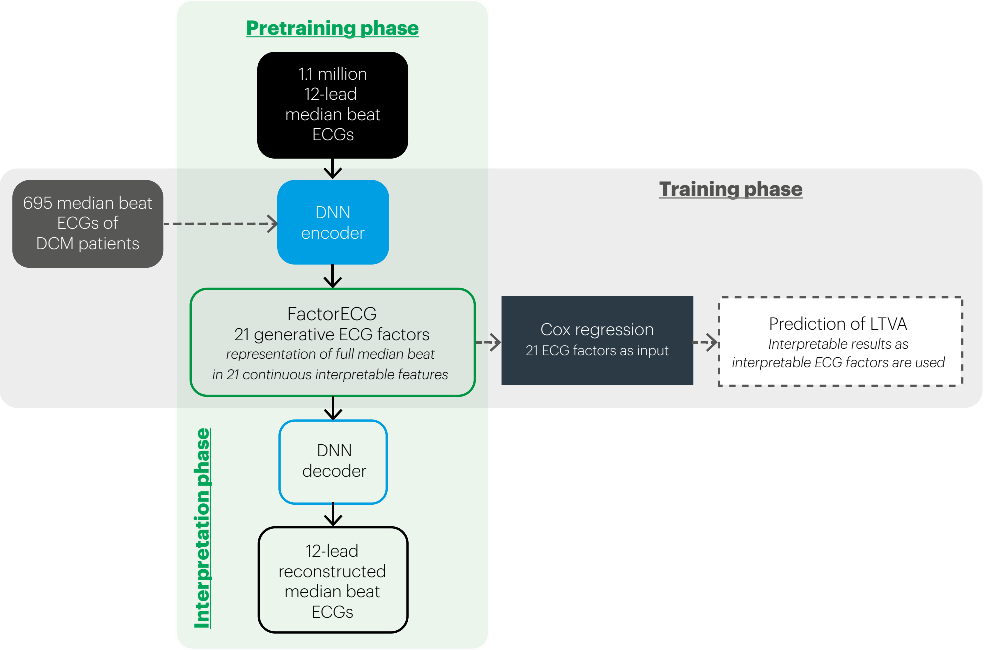

Pretraining phase
1.1 million 12-lead median beat ECGs
695 median beat ECGs of DCM patients
DNN encoder
Training phase
FactorECG
21 generative ECG factors
representation of full median beat in 21 continuous interpretable features
Cox regression
21 ECG factors as input
Prediction of LTVA
Interpretable results as interpretable ECG factors are used
DNN decoder
12-lead reconstructed median beat ECGs
Interpretation phase

Figure 1. ***Overview of the pretraining and training phases of the FactorECG algorithm.*** During the pretraining phase, 1.1 million 12-lead median beat ECGs were included for training of the variational autoencoder (VAE). The VAE was trained to compress all 12-lead median beat ECGs into 21 continuous factors of variation (the FactorECG), that can subsequently be used to reconstruct the median beat ECG. The VAE is explainable by visualizing the influence of the individual ECG factors on the ECG morphology using the decoder. In the training phase, for each of the 695 DCM patients, median beat ECGs were encoded into 21 generative factors using the pretrained encoder. These 21 generative ECG factors were used as an input in a Cox regression model to predict life-threatening ventricular arrhythmias. The importance of each ECG factor was then determined by investigating the hazard ratios of the standardized ECG factors. DCM: dilated cardiomyopathy, DNN: deep neural network, LTVA: life-threatening ventricular arrhythmia.

Explainability of the individual factors was obtained on a model level using factor traversals. Starting with a mean FactorECG for this population (ie. the mean value of the 21 ECG factors), a median beat ECG is reconstructed using the decoder. Subsequently, for each individual ECG factor, values between -4 and 4 are added and ECGs are reconstructed for every value. Meanwhile, values for the other factors are kept constant. This way we are able to visualize the effect of a single ECG factor on the median beat ECG morphology in this cohort (**Figure 2**). On the individual patient level, explainability is obtained by investigating the FactorECG values of that specific ECG. A tool to visualize the factors interactively can be found at https://dcm.ecgx.ai. The architecture and model training process were implemented using PyTorch (version 1.7.0+cu110) in Python (version 3.6.7).

Outcome definitions

The primary study outcome was LTVA, defined as the composite outcome of sustained ventricular tachycardia (VT) >100 bpm lasting >30sec or with hemodynamic compromise, ventricular fibrillation (VF), sudden cardiac death (SCD, defined as death of cardiac origin that occurred unexpectedly within one hour after the onset of new symptoms) or appropriate ICD therapy (defined as any ICD therapy delivered by the device in response to VT or VF according to stored intracardiac electrograms).[3]

Statistical analyses

For the baseline table, mean ± SD or median [interquartile range] were used where appropriate. Missingness in baseline data were not addressed. Each baseline ECG's generative factors (the FactorECGs, as computed by the VAE encoder) were included in a Cox proportional hazards model (**Figure 1**). All patients had a digitalized ECG available. The proportional hazards assumption was tested. Hazard ratios (HR) were reported, and 95% confidence intervals were computed using 2000 bootstrap samples. To rule out that the VAE model was solely considering already established ECG characteristics (ventricular rate, PR-interval, QRS-duration and Bazett corrected QT-interval), a Cox proportional hazard model was also fitted using these variables in a complete case analysis. The correlations of the significant ECG factors were plotted against the left atrial (LA) dimension and left atrial volume index (LAVI) measured on standard care clinical echocardiography using both the first (closest to baseline) and last (closest to follow-up) available measurements.[11] Additionally a Kaplan Meier curve was plotted for one of the significant VAE generative factors. All analyses were performed using Python (version 3.8.5).

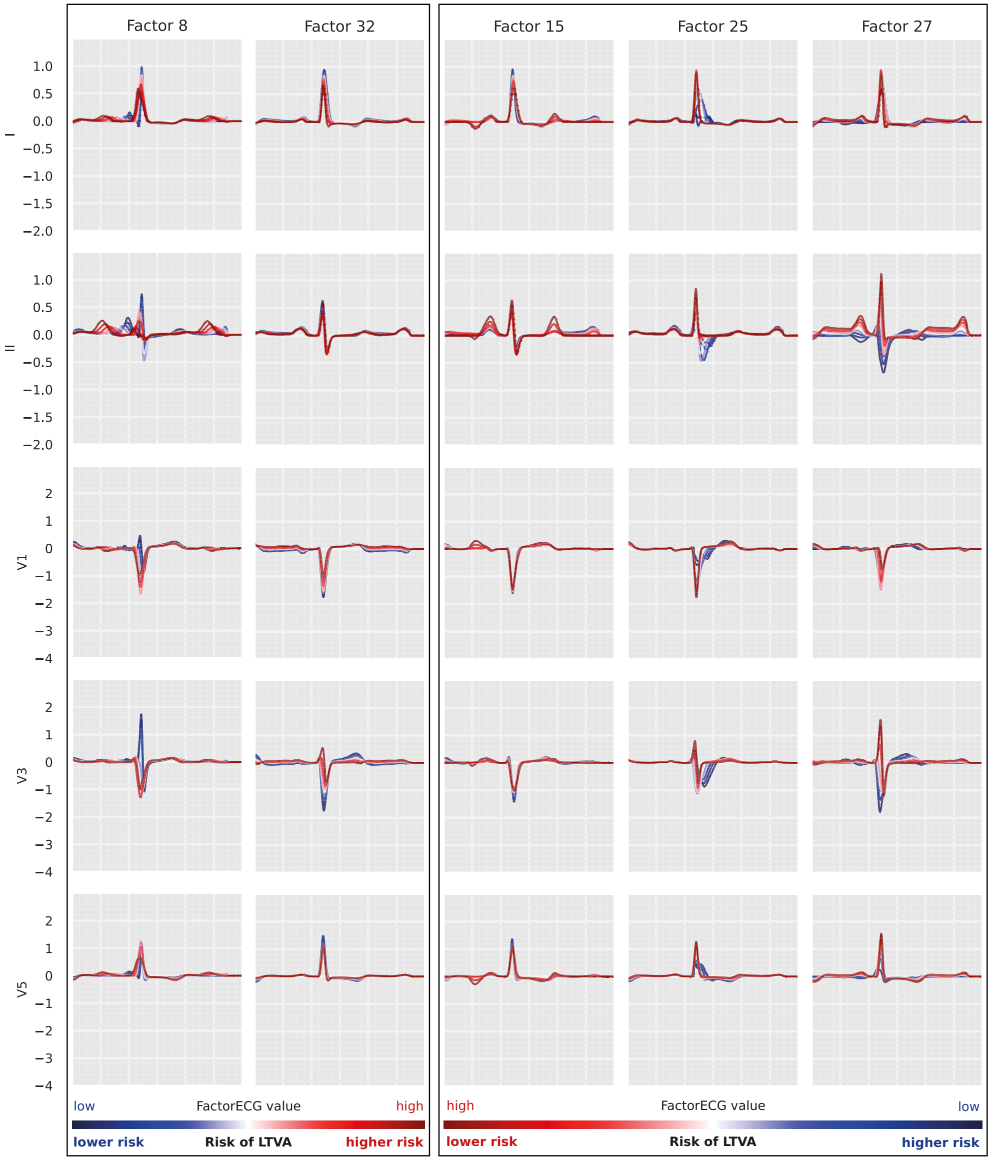


Figure 2.
Factor traversals for the ECG factors that were associated with LTVA in the DCM cohort.
We start with a mean FactorECG for this population (ie. the mean value of the 21 ECG factors) and reconstruct an ECG using the decoder (white). Subsequently, for each individual ECG factor, values between -4 (blue) and 4 (red) are added and ECGs are reconstructed for every value. Meanwhile, values for the other factors are kept constant. This way we are able to visualize the effect of a single ECG factor on the median beat ECG morphology in this cohort. For factors 8 and 32, high values of the factors were associated with a higher risk of LTVA (left). For factors 15, 25 and 27, conversely, low values of the factors were associated with a higher risk of LTVA (right). The FactorECG decoder reconstructs the full 12-lead median beat ECG, a selection of leads is shown in this figure. DCM: dilated cardiomyopathy, LTVA: life-threatening ventricular arrhythmia.

Results

Patient characteristics

Baseline characteristics stratified by centre and outcome are depicted in **Table 1**. A total of 695 patients were included from the UMCU and MUMC+, which were predominantly male (n=442, 64%) with a median age of 54 years [interquartile range (IQR) 44-62] and median LVEF of 30% [IQR 23%-39%]. A total of 115 (17%) reached the study outcome in both centres combined during a median follow-up of 4.3 years [IQR 2.0 – 7.5]. In summary, patients from the MUMC+ cohort had less severe symptoms at baseline with primarily New York Heart Association classes I and II as opposed to the UMCU cohort with primarily II and III, and a median LVEF of 33% [IQR 25-40]. A lower proportion of MUMC+ patients (25, 6%) reached the study outcome of LTVA compared to 90 (28%) UMCU patients.

Prediction of LTVA with established ECG variables

Established ECG variables (such as ventricular rate, PR-interval, QRS-duration, and QTc-time) were entered in a "baseline" Cox regression model controlled for guideline indication (complete case analysis with n = 577, excluding patients without a measurable PR-interval (n = 118, due to atrial fibrillation/flutter)). This baseline model had a C-statistic of 0.58 [95% CI 0.52 – 0.64] and no significant effects of: ventricular rate (HR 0.94 per 10 beat/s increase [95%CI 0.81-1.09], p = 0.41), QRS duration (HR 1.08 per 10 ms increase [95% CI 0.98-1.19], p=0.13) and QTc-time (HR 0.95 per 10 ms increase [95% CI 0.88-1.03], p = 0.19). The PR-interval was however significantly associated with LTVA (HR 1.06 per 10 ms increase [95%CI 1.00-1.13], p = 0.04). The results of this model are depicted in Supplementary **Table 1**.

Table 1.* *Patient characteristics at baseline (first evaluation) stratified by centre and outcome. Baseline characteristics of the included cohorts. NYHA = New York Heart Association; LTVA = Life Threatening Ventricular Arrhythmia; ICD = Implantable Cardioverter-Defibrillator; LVEF = Left Ventricular ejection fraction; * = of valid, in patients for which a NYHA class was noted in the electronic health record. MRI LGE = Magnetic Resonance Imaging Late Gadolinium Enhancement. ** = of valid, in patients with cardiac MRIs.

	UMCU all (n=317)	UMCU without LTVA (n=227)	UMCU with LTVA (n=90)	MUMC all (n=378)	MUMC without LTVA (n=353)	MUMC with LTVA (n=25)
Age (years), median [Q1-Q3]	52 [42 – 61]	51 [41 – 60]	52 [42-62]	55 [47 – 63]	56 [47-63]	54 [49-63]
Male Sex	195 (62%)	129 (57%)	66 (74%)	247 (65%)	228 (65%)	19 (76%)
NYHA-class						
I	53 (20%)*	36 (18%)*	17 (23%)*	158 (42%)	150 (43%)	8 (32%)
II	102 (39%)*	71 (36%)*	31 (41%)*	175 (46%)	163 (47%)	12 (48%)
III	79 (30%)*	56 (28%)*	23 (32%)*	37 (10%)	32 (9%)	5 (20%)
IV	27 (10%)*	36 (18%)*	4 (5%)*	8 (2%)	8 (2%)	0 (0)%
Diabetes Mellitus	42 (13%)	31 (13%)	11 (12%)	52 (14%)	50 (14%)	2 (8%)
Hypercholesterolemia	37 (13%)	26 (13%)	11 (13%)	41 (11%)	38 (11%)	3 (12%)
(Ever) smoked	203 (64%)	145 (64%)	57 (63%)	77 (20%)	72 (20%)	5 (20%)
History of LTVA	42 (13%)	17 (7%)	25 (27%)	8 (2%)	7 (2%)	1 (4%)
Family history of DCM	133 (42%)	97 (43%)	36 (40%)	47 (14%)	39 (11%)	8 (32%)
ICD implantation	233 (74%)	145 (63%)	88 (97%)	0 (0%)	0 (0%)	0 (0%)
LVEF (%), median [Q1-Q3]	25 [20-33]	25 [20-33]	25 [19-32]	33 [25-40]	28 [22-37]	33 [25-41]
MRI LGE	84 (56%**)	60 (51%*)	24 (71%**)	n/a	n/a	n/a

Prediction of LTVA with FactorECG

The VAE compressed the ECG data into 21 different ECG factors and their factor traversals are available in Supplementary **Figure 1**. In Cox regression, F_8 (HR 1.60; 95%CI [1.29-1.99], $p < 0.005$), F_{15} (HR 0.81; 95%CI

[0.66-0.99], *p = 0.04*), F_{25} (HR 0.77 95%CI [0.62 – 0.95], *p = 0.02*), F_{27} (HR 0.71, 95%CI[0.57–0.88], *p < 0.005*) and F_{32} (HR 1.26, 95%CI [1.03–1.55], *p = 0.03)* were significantly associated with the outcome after correcting for guideline indication (NYHA II/III and LVEF < 35%, *p = 0.84*). C-statistic for the model was 0.67 [95% CI 0.62 – 0.72]. A reconstruction of the significant generative factors (F_8, F_{15}, F_{25}, F_{27} and F_{32}) has been illustrated in **Figure 2**. F_8 encodes for PR-interval and P-wave morphology, where high values increase PR-interval and broaden the P-wave. F_{15} encodes for P-wave height and P/T-overlap, where low values are correlated with atrial fibrillation and third-degree AV-block. F_{25} encodes for conduction delays in the right bundle (right bundle branch block), where low values increase the block. F_{27} encodes for P- and R- axis deviation, where low values flatten out the P-wave. F_{32} encodes for QRS-T amplitudes, with low values reconstruct QRS-T microvoltages. Results of the Cox regression model and the descriptions of the generative factors are present in **Table 2** and **Supplementary Table 2.**

Table 2.* *Cox proportional hazards model of generative factors in both cohorts. Results of Cox regression and explanation of (significant) factors including their association with known electrocardiographic and echocardiographic pathologies as described in Van de Leur and Bos et al (2022)8. *significant

Factors	Factor descriptions	Hazard Ratio	95% Confidence Interval	P-value
F_1	Inferolateral ST deviation	0.91	0.72-1.14	0.39
F_5	Inferolateral T-wave height and orientation	1.17	0.92-1.48	0.19
F_6	P-wave height and/or shape	1.14	0.90-1.44	0.27
F_{8*}	PR-interval (high values associated with first degree AV-block and reduced LVEF)	1.61	1.29-1.99	<0.005
F_9	T-wave height and orientation	1.12	0.89-1.39	0.34
F_{10}	Ventricular rate	0.93	0.76-1.13	0.46
F_{11}	Subtle P- and T-wave changes	1.00	0.83-1.21	0.97
F_{12}	Onset of depolarisation	1.08	0.86-1.36	0.50
F_{13}	Anterior ST deviation	0.85	0.68-1.06	0.14
F_{15*}	P-wave height and P/T-overlap (low values associated with third degree AV-block and junctional tachycardia)	0.81	0.66-0.99	0.04
F_{16}	T-wave morphology	1.14	0.94-1.40	0.19
F_{17}	Lateral ST-deviation	1.08	0.85-1.38	0.51
F_{19}	Precordial R-wave progression and combined P-QRS-T-amplitude	1.07	0.87-1.33	0.51
F_{22}	Subtle T-wave changes	1.02	0.83-1.25	0.85
F_{23}	P-wave height and/or shape	1.13	0.93-1.37	0.21
F_{25*}	Right bundle branch delay (low values associated with ventricular tachycardia, RBBB and reduced LVEF)	0.77	0.62-0.95	0.02
F_{26}	Left bundle branch delay	1.02	0.81-1.29	0.85
F_{27*}	P- and R- axis deviation (low values associated with AF, junctional bradycardia, ventricu-lar tachycardia, and left axis deviation)	0.71	0.57-0.88	<0.005
F_{30}	QR interval	0.92	0.74-1.16	0.48
F_{31}	QRS-T amplitudes	0.86	0.71-1.05	0.15
F_{32*}	QRS-T amplitudes (high values associated with microvoltages)	1.26	1.02-1.55	0.03

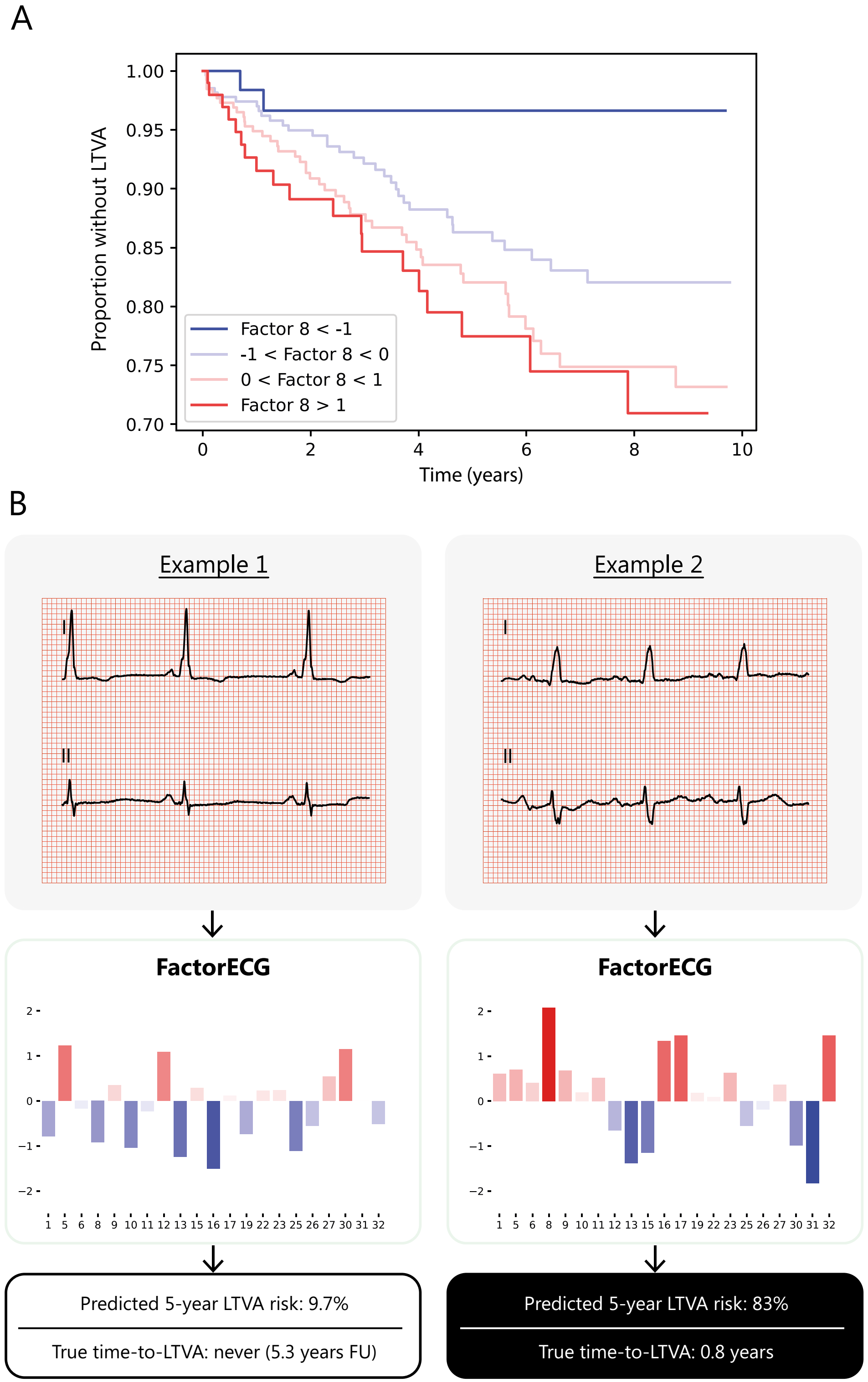


Figure 3.
Kaplan Meier curve and examples for the predictive value of factor 8.
A Kaplan Meier (A) for factor 8 and (B) two ECGs of patients with a low and high predicted five-year LTVA risk are depicted. The values for each factor are depicted below the ECG, along with the outcomes of the patient. The ECG on the left had a low value for factor 8, corresponding to a short PR and P-wave duration: this patient had a low predicted risk of LTVA and did not reach the endpoint. The ECG on the right had a high value for factor 8, corresponding to a broad P wave with a long PR interval: this patient had a high predicted risk of LTVA and reached the outcome.

The partial effects on outcome per significant factor have been plotted in Supplementary **Figure 2**. As an example, the ECGs and their corresponding values of the generative factors of two patients were plotted in **Figure 3**. A summary figure of this study was depicted in **Figure 4**. To address the effect that cardiac memory after pacing, a subgroup analysis was run excluding patients with a pacemaker (n=32) which showed similar factors to be important (**Supplementary Table 3**).

LA dimensions

To investigate the possibility that the identified factors were an effect of anatomical substrates of P-wave abnormalities, such as atrial remodelling, first and last LAVI and LA dimensions (by outcome) of complete UMCU cases (n = 219) were plotted (supplementary figures 4, 5, 6 and 7 respectively). LA's were significantly larger in the last echocardiography, compared to the first (p = 0.02, supplement figure 6). Next, the LA dimensions were plotted to these factors (supplement figure 8) which showed no association between F_8, F_{15}, F_{25}, F_{27} and F_{32} and LA dimensions.

Discussion

This is the first study to use an explainable DNNs trained with (baseline) ECGs for LTVA prediction in DCM patients on a multicentre dataset. By using an inherently explainable DNN architecture, we were able to distinguish patients at risk for LTVA whilst allowing interpretation and visualisation of pivotal ECG features.[7] The model was able to identify patients at highest risk with a predominant network focus on P-wave abnormalities. Furthermore, these identified P-wave abnormalities did not correlate to their anatomical analogues (LA dimension/LAVI), suggesting an electrophysiological substrate.

FactorECG findings in relation to prior studies

The FactorECG encompasses the median beat ECG, including most of its features, into 21 generative factors of variation (see https://dcm.ecgx.ai/). This novel strategy allows to simultaneously evaluate most characteristics that make up an ECG automatically in much smaller datasets, rather than using selected and human derived ECG features. Overall, the factors that were most predictive for LTVA primarily encoded for several P-wave characteristics, such as PR-duration, P-wave morphology, and P-wave axis (**Figure 2**). The combination of reconstructed ECGs together with the hazard ratios allow for a novel in-depth interpretation of a DNN's features. A high value in F_8 for instance, leads to PR-prolongation with a broadened P-wave, whereas a low value in F_{27} leads to removal of the P-wave, which is associated with atrial fibrillation, a known clinical risk factor for LTVA in DCM.[3] Because the baseline model using established ECG variables performed poorly, this indicated that the VAE generative factors are more complex than solely the standard ECG intervals. The combination of the 21 generative factors as well as their interpretation allow for LTVA prediction and feature detection (**Figure 3**).

The fact that atrial (i.e. P-wave) abnormalities predict ventricular events (i.e. LTVA) may be considered remarkable. However, this association has been described before, and has been thought to be due to shared mechanistic pathologies between atria and ventricles, such as ion-channel abnormalities, or atrioventricular fibrosis due to atrial remodelling.[1,3,13] In a recently published population study of 13580 participants, abnormal P-wave indices were independently associated with LTVA, after adjustment for age, sex, race and study centre.[14] As it is likely that these P-wave indices are caused by atrial remodelling, we investigated the association of anatomical LA characteristics and our identified ECG factors. As expected, LA dimensions increased significantly over time, indicating disease progression. However, we did not find any association to the significant ECG factors, suggesting an exclusive electrophysiological substrate. This is in line with other reports, in which individual ECG P-wave changes were not reliable predictors of anatomic atrial enlargement.[15,16]

Myocardial fibrosis is often seen in patients with DCM and may cause zones of slow conduction in the myocardium, resulting in zigzag pathways that are prone to causing ventricular tachycardias.[17] These cellular mechanisms may be visible on the ECG as increased QRS duration and bundle branch block or low voltages. In this study, F_{25} reflects increased QRS duration in case of right bundle branch blocks and was associated with LTVA as well. Left bundle branch block (F_{26}) however, was not associated with LTVA. As these patients generally are CRT recipients which were excluded from our analyses, this may have caused an underestimation of the effect of left bundle branch blocks in our model. Lower voltages seem to be reflected in the FactorECG in F_{32}, which is also associated with LTVA. More complex electrocardiographic markers, such as QRS fractionation, T-wave alternans and QRS-T angle, have also been proposed.[1] QRS fractionation and T-wave abnormalities, however, did not appear as an explanatory ECG variable in our model. T-wave alternans (defined as changing T-wave morphology, occurring in each alternant beat) has been repeatedly associated with LTVA in DCM, but cannot be measured in the single ECG median beat that is used in the current research. All these ECG markers are limited by standardization difficulties, which may be decreased by (automatic) interpretation using

DNNs.[1,4,18,19]

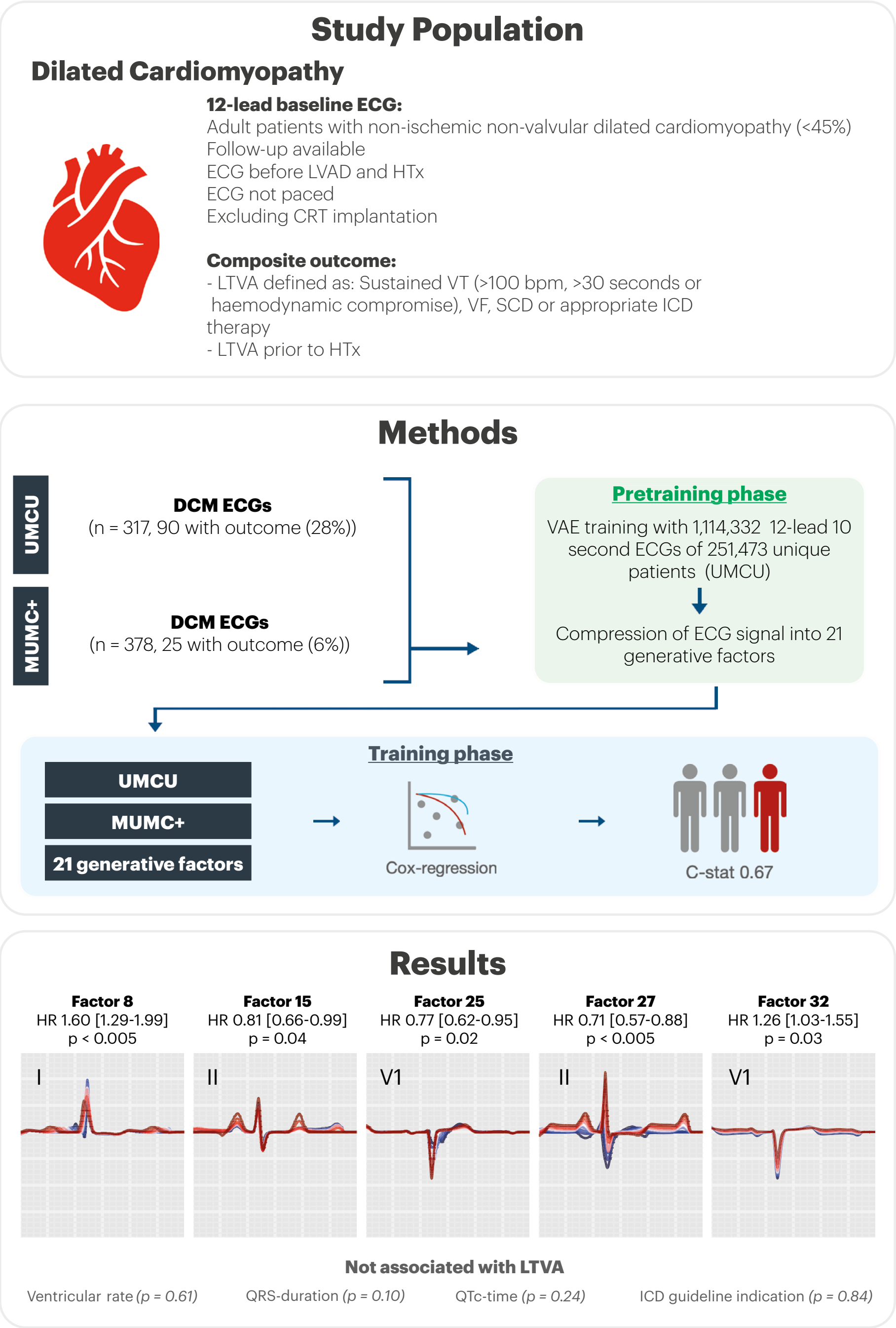


Figure 4. ***Study summary figure, including the methods and results.*** The study population were patients with dilated cardiomyopathy, in which an explainable pre-trained deep neural network (FactorECG) was trained for the outcome of life-threatening ventricular arrhythmias. This network encoded the median beat ECG into 21 factors to generate an ECG using only these factors, allowing to evaluate most characteristics that make up an ECG automatically, in a relatively small dataset. LVAD = Left Ventricular Assist Device. HTx = Heart Transplantation, CRT = Cardiac Resynchronisation Therapy. ECG = electrocardiogram. VT = Ventricular Tachycardia. VF = Ventricular Fibrillation. SCD = Sudden Cardiac Death. ICD = Implantable Cardioverter-Defibrillator. HR = Hazard Ratio. UMCU = University Medical Centre Utrecht. MUMC+ = Maastricht University Medical Centre.

As these networks are generally “black-box” algorithms that need very large datasets for training, a strategy of reducing the ECG into its generative factors was used. These interpretable factors were then used in a common statistical model (Cox regression), that allowed for pivotal ECG features to be visualized.

Future studies are warranted to prospectively validate the identified ECG abnormalities and their electrophysiological substrate for LTVA prediction in DCM, including a comparison with accepted risk factors for LTVA. Since longer PR-interval and wide QRS duration were associated with LTVA, assessment of the value of hemiblocks may also be considered. Importantly, the addition of prolonged measurements (such as exercise tests or Holter for T-wave alternans) in DNNs remains to be investigated.

Genotype-phenotype associations

DCM has a genetic basis in 30-50% of cases and specific genotype-phenotype associations are known to lead to arrhythmogenic phenotypes. One study analysed over 75.000 ECGs from the UK Biobank and established several genetic ECG signatures. A polygenic effect on PR-interval for instance, was identified, as well as genetic variants related to the Q-wave in DCM. The strongest Q-wave locus was discovered in *BAG3*: a gene in which pathogenic variants have been described for DCM with high penetrance and a high risk of progressive heart failure.[20] As our VAE model assessed the entire ECG, an interesting significant factor included QRS-T voltages (F_{32}), with high values in this factor associated with microvoltages. These microvoltages are an established ECG characteristic for phospholamban cardiomyopathy, which can lead to both a highly arrhythmogenic DCM phenotype and arrhythmogenic cardiomyopathy.[2] Integrating genome and phenome provides unique opportunities to study ECG biology in relation to genetic risk which can be explored by future studies using DNNs.[20–22] Furthermore, these studies may pave the way for using artificial intelligence models for risk prediction in DCM patients to estimate an individual’s lifetime (genetic) risk of developing a specific arrhythmogenic DCM phenotype.

Limitations

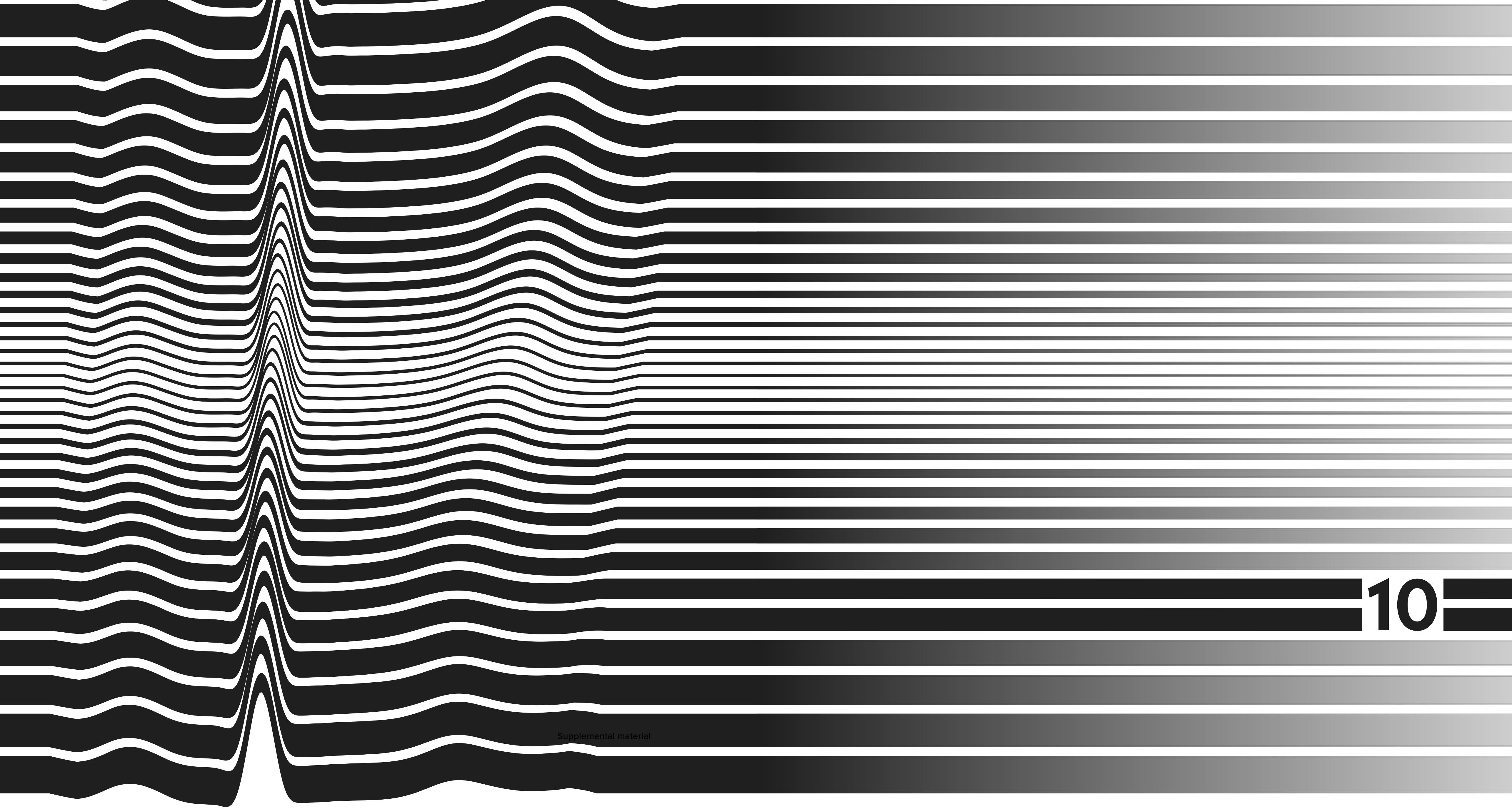
This study has several limitations to address. Given the nature of retrospective cohorts, data may contain missingness not at random and bias may be present requiring prospective evaluation of the findings. As the UMCU is a heart transplantation centre, this may have caused a selection bias. To account for this, an external cohort was added from the MUMC+ (non-heart transplantation centre) of which the patients logically presented with less severe phenotypes (**Table 1**). Unfortunately, the characteristics of the implanted ICDs in this population were not available, which may have biased our findings. More importantly, since ICD shocks are not a true surrogate for sudden cardiac death in patients with DCM, the results need confirmation in a study population with fewer ICD carriers or considering only fast events (i.e., >250/min).[23] Because DCM is relatively rare, the results may be due to sample size and require confirmation in larger (prospective) studies.

Conclusion

To the best of our knowledge, this study is the first to use interpretable DNNs trained with ECGs for LTVA prediction in DCM patients. We observed that the VAE network combined with an interpretable Cox regression can distinguish patients at risk of LTVA. The use of this inherently explainable DNN pipeline allowed interpretation and visualisation of pivotal ECG features.[7] While the VAE network encompasses the complete ECG, predictions were mainly driven by P-wave abnormalities that did not correlate with LA dimensions, suggesting an electrophysiological substrate. Future studies are warranted to validate these findings and elucidate their electrophysiological substrate for LTVA prediction in DCM.

REFERENCES

1. Sammani A, Kayvanpour E, Bosman LP, et al. Predicting sustained ventricular arrhythmias in dilated cardiomyopathy: a meta- analysis and systematic review. *Esc Hear Fail* 2020;7:1430–1441.
2. McDonagh TA, Metra M, Adamo M, el al. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failureDeveloped by the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) With the special contribution of the Heart Failure Association (HFA) of the ESC. *Eur Heart J* 2021;42:ehab368.
3. Kayvanpour E, Sammani A, Sedaghat-Hamedani F, et al. A novel risk model for predicting potentially life-threatening arrhythmias in non-ischemic dilated cardiomyopathy (DCM-SVA risk). *Int J Cardiol* 2021;339:75–82.
4. Leur RR van de, Boonstra MJ, Bagheri A, et al. Big Data and Artificial Intelligence: Opportunities and Threats in Electrophysiology. *Arrhythmia Electrophysiol Rev* 2020;9:146–154.
5. Ko W-Y, Siontis KC, Attia ZI, et al. Detection of Hypertrophic Cardiomyopathy Using a Convolutional Neural Network-Enabled Electrocardiogram. *J Am Coll Cardiol* 2020;75:722–733.
6. Leur RR van de, Blom LJ, Gavves E, et al. Automatic Triage of 12-Lead ECGs Using Deep Convolutional Neural Networks. *J Am Heart Assoc* 2020;9:e015138.
7. 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.
8. Leur RR van de, Bos MN, Taha K, et al. Explainable deep neural network for enhanced interpretation of 12-lead electrocardiograms. *Submitted* 2021.
9. Leur R van de, Taha K, Bos MN, et al. Discovering and Visualizing Disease-specific Electrocardiogram Features Using Deep Learning: Proof-of-concept in Phospholamban Gene Mutation Carriers. *Circulation Arrhythmia Electrophysiol* 2021;14:CIRCEP.120.009056.
10. Sapp JL, Parkash R, Wells GA, et al. Cardiac resynchronization therapy reduces ventricular arrhythmias in primary but not secondary prophylactic implantable cardioverter defibrillator patients: Insight from the resynchronization in ambulatory heart failure trial. *Circulation Arrhythmia Electrophysiol* 2017;10:e004875.
11. Sammani A, Jansen M, Linschoten M, et al. UNRAVEL: big data analytics research data platform to improve care of patients with cardiomyopathies using routine electronic health records and standardised biobanking. *Neth Heart J* 2019;27:426–434.
12. Higgins I, Matthey L, Pal A, et al. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *Conference Track Proceedins*. 5th International Conference on Learning Representations; 2018.
13. Spezzacatene A, Sinagra G, Merlo M, et al. Arrhythmogenic Phenotype in Dilated Cardiomyopathy: Natural History and Predictors of Life- Threatening Arrhythmias. *J Am Heart Assoc* 2015;4:e002149.
14. Maheshwari A, Norby FL, Soliman EZ, et al. Association of P-Wave Abnormalities With Sudden Cardiac and Cardiovascular Death: The ARIC Study. *Circulation Arrhythmia Electrophysiol* 2021;14:e009314.
15. Tsao CW, Josephson ME, Hauser TH, et al. Accuracy of electrocardiographic criteria for atrial enlargement: validation with cardiovascular magnetic resonance. *J Cardiov Magn Reson* 2008;10:7–7.
16. Truong QA, Charipar EM, Ptaszek LM, et al. Usefulness of electrocardiographic parameters as compared with computed tomography measures of left atrial volume enlargement: from the ROMICAT trial. *J Electrocardiol* 2011;44:257–264.
17. Bakker JM de, Capelle FJ van, Janse MJ, et al. Slow conduction in the infarcted human heart. "Zigzag" course of activation. *Circulation* 2018;88:915–926.
18. Pei J, Li N, Gao Y, et al. The J wave and fragmented QRS complexes in inferior leads associated with sudden cardiac death in patients with chronic heart failure. *Ep Europace* 2012;14:1180–1187.
19. Vandenberk B, Robyns T, Goovaerts G, et al. Inter- and intra-observer variability of visual fragmented QRS scoring in ischemic and non-ischemic cardiomyopathy. *J Electrocardiol* 2018;51:549–554.
20. Verweij N, Benjamins J-W, Morley MP, et al. The Genetic Makeup of the Electrocardiogram. *Cell Syst* 2020;11:229-238.e5.
21. Meder B, Rühle F, Weis T, et al. A genome-wide association study identifies 6p21 as novel risk locus for dilated cardiomyopathy. *Eur Heart J* 2014;35:1069–1077.
22. Lewis CM, Vassos E. Polygenic risk scores: From research tools to clinical instruments. *Genome Med* 2020;12:44.
23. Ellenbogen KA, Levine JH, Berger RD, et al. Are implantable cardioverter defibrillator shocks a surrogate for sudden cardiac death in patients with nonischemic cardiomyopathy? *Circulation* 2006;113:776–782.



10

Electrocardiogram-based deep learning improves outcome prediction following cardiac resynchronization therapy

European Heart Journal

Philippe C Wouters*, Rutger R van de Leur*, Melle B Vessies, Antonius M W van Stipdonk, Mohammed A Ghossein, Rutger J Hassink, Pieter A Doevendans, Pim van der Harst, Alexander H Maass, Frits W Prinzen, Kevin Vernooy and René van Es

Supplemental material

Abstract

Aims

This study aims to identify and visualize ECG features using an explainable deep learning-based algorithm to predict cardiac resynchronization therapy (CRT) outcome. Its performance is compared to current guideline ECG criteria and QRSAREA.

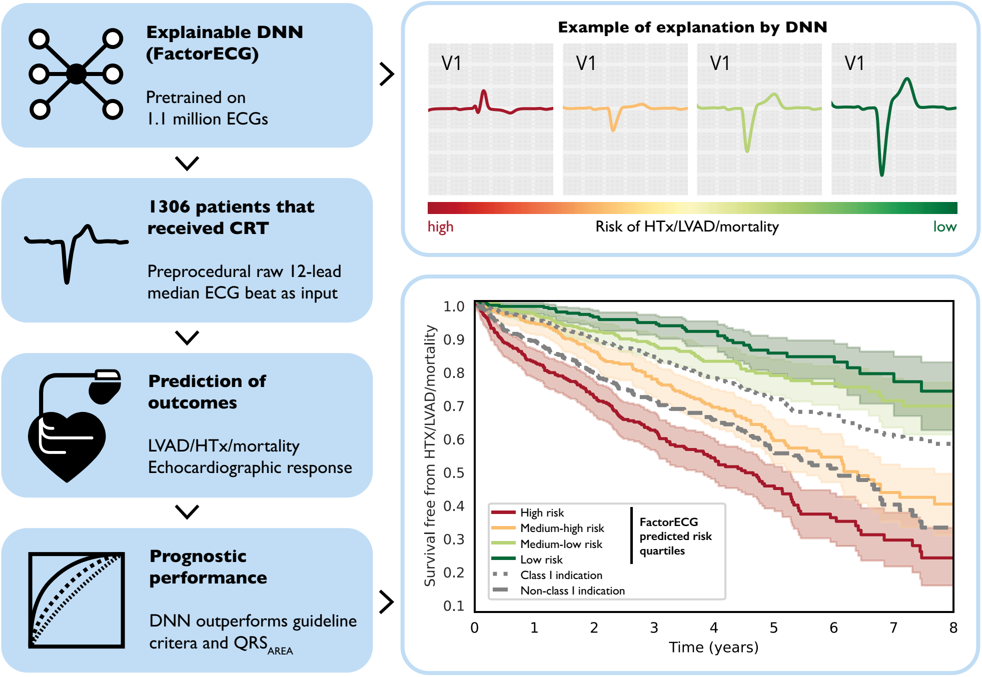
Methods and results

A deep learning algorithm, trained on 1.1 million ECGs from 251,473 patients, was used to compress the median beat ECG, thereby summarizing most ECG features in only 21 explainable factors (FactorECG). Pre-implantation ECGs of 1306 CRT patients from three academic centers were converted into their respective FactorECG. FactorECG predicted the combined clinical endpoint of death, left ventricular assist device, or heart transplantation (c-statistic 0.69 [95% CI 0.66 – 0.72]), significantly outperforming QRSAREA and guideline ECG criteria (c-statistic 0.61 [95% CI 0.58 – 0.64] and 0.57 [95% CI 0.54 – 0.60], $p < 0.001$ for both). Addition of 13 clinical variables was of limited added value for the FactorECG model when compared to QRSAREA (Δ c-statistic 0.03 versus 0.10). FactorECG identified inferolateral T-wave inversion, smaller right precordial S-wave and T-wave amplitude, ventricular rate, and increased PR-interval and P-wave duration to be important predictors for poor outcome. An online visualisation tool was created to provide interactive visualizations (https://crt.ecgx.ai).

Conclusion

Requiring only a standard 12-lead ECG, FactorECG held superior discriminative ability for the prediction of clinical outcome as compared to guideline criteria and QRSAREA, without requiring additional clinical variables. End-to-end automated visualisation of ECG features allows for an explainable algorithm, which may facilitate rapid uptake of this personalised decision-making tool in CRT.

Graphical abstract


Explainable DNN (FactorECG)
Pretrained on 1.1 million ECGs
1306 patients that received CRT
Preprocedural raw 12-lead median ECG beat as input
Prediction of outcomes
LVAD/HTx/mortality
Echocardiographic response
Prognostic performance
DNN outperforms guideline critera and QRS$_{AREA}$
Example of explanation by DNN
V1
V1
V1
V1
high
Risk of HTx/LVAD/mortality
low
Survival free from HTX/LVAD/mortality
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
7
8
Time (years)
High risk
Medium-high risk
Medium-low risk
Low risk
Class I indication
Non-class I indication
FactorECG predicted risk quartiles

First, an artificial intelligence algorithm (variational auto-encoder) was pretrained on 1.1 million ECGs to learn the underlying continuous factors that generate the ECG (i.e., the FactorECG). In this process, the VAE learns to reconstruct ECGs as accurate as possible using only 21 continuous factors without any human input. In the training phase, preprocedural median beat ECGs of 1306 CRT patients were each converted into their FactorECG. These 21 factors were subsequently used as input in a Cox model to predict the primary composite endpoint of LVAD implantation, heart transplantation and all-cause death, and the secondary endpoint of echocardiographic response. FactorECG significantly improved outcome prediction following CRT as compared to the current guidelines and QRSAREA. The algorithm is explainable by using the decoder to visualize the effect of the ECG factors that significantly predicted outcome on the median beat ECG morphology. Here, for example, the influence of factor 9 (F9) is visualized, where higher values represent a more left bundle branch block-like ECG morphology and lower values a more right bundle branch block-like morphology. Legend: CRT; cardiac resynchronization therapy, DNN; deep neural network, ECG; electrocardiogram, HTx; heart transplantation, LVAD; left ventricular assist device.

Introduction

In patients with dyssynchronous heart failure (HF), cardiac resynchronization therapy (CRT) can effectively restore left ventricular (LV) electrical activation and mechanical function, thereby improving clinical outcome.[1] However, for CRT to be beneficial, sufficient LV electrical conduction delay must be present.[2] Currently, patients are selected based on various requirements set out by different guidelines. However, despite indicating the highest level of recommendation, by itself, a class I indication does not necessarily ensure a sustained response after CRT.[3] Conversely, effectiveness of CRT is variable and doubted in patients without left bundle branch block (LBBB) morphology or intermediate QRS-duration.[2,4,5] Although a substantial portion of these patients will benefit regardless, they are at increased risk of not being considered for treatment.[5,6] Accurate and objective identification of the underlying electrical substrate is therefore crucial to optimize patient selection and ensure optimal treatment.

Currently, electrical characteristics derived from the electrocardiogram (ECG), such as LBBB-morphology and QRS-duration, are used to determine eligibility for CRT.[2] Multiple ECG criteria for LBBB have been defined[7], and inter-observer variability is high.[8] Moreover, a variety of LV electrical activation patterns are concealed in the ECG, further complicating clinical decision making.[9] Recently, QRS_{AREA} has emerged as a new and objective computerized measure.[3,10] QRS_{AREA} is independently associated with survival and echocardiographic response, outperforming LBBB-morphology and QRS-duration.[3,10] As such, QRS_{AREA} partly overcomes the challenges of subjective ECG interpretation, but (subtle) ECG characteristics, also besides the QRS-complex, are still not considered.

Machine learning has gained interest as a means of integrating large amounts of variables, thereby producing advanced clinical decision models. The SEMMELWEIS-CRT score, for example, outperforms many already existing risk scores, but relies on 33 clinical variables.[11] Besides being laborious to use, such models also rely on human interpretation of input variables such as left ventricular ejection fraction (LVEF), New York Heart Association (NYHA), LBBB-morphology and QRS-duration, which are all subjectively assessed. Hence, although such models may predict response to CRT, large amounts of clinical variables will still need to be acquired, extracted and entered in such models.[11–14]

A recent development in the field of machine learning, called deep learning, can learn features from the raw ECG signal without the necessity for any human interpretation.[15] Deep learning algorithms may therefore be used to automatically detect, identify and classify ECG abnormalities that are associated with non-response or poor outcome after CRT. Although the need for very large datasets and the lack of interpretability were deemed common drawbacks of deep learning, a novel technique that uses a variational auto-encoder (the FactorECG) was recently introduced, [16] This approach enables physicians to better understand and verify the learned ECG features of deep learning algorithms, and make the technique available to much smaller datasets.

The present study seeks to compare contemporary guideline ECG criteria for CRT implantation and QRS_{AREA} with the FactorECG for the prediction of a combined clinical endpoint and echocardiographic response. In addition, we aim to identify and visualise ECG features associated with these outcome measures.

Methods

Study design

All data were acquired for routine patient care and handled anonymously, and were collected as part of the multicentre Maastricht-Utrecht-Groningen (MUG) registry[10]. Under these circumstances, informed consent was waived by the Institutional Review Board at the time of the study. All study procedures were performed in compliance with the Declaration of Helsinki.

Only patients who received a de novo CRT-device with a transvenous LV-lead were considered for the present study **(Figure 1)**. A baseline ECG (within 3 months before implantation) was required for the primary endpoint analysis, whereas paired echocardiographic examination at baseline and follow-up (6 to 12 months) was required for the secondary endpoint. Echocardiographic exams from various vendors were used to determine LV end-systolic volume (LVESV), and LVEF was calculated using the Simpson's modified biplane method (IntelliSpace Cardiovascular, Philips, Eindhoven).

The primary endpoint was a combined clinical endpoint consisting of left ventricular assist device (LVAD) implantation, heart transplantation (HTx), and all-cause mortality. The secondary endpoint was echocardiographic non-response, defined as relative decrease in LVESV of less than 15%.[17] In addition, three tertiary endpoints were investigated: 1) a composite of HF hospitalization and the primary endpoint, 2) HF hospitalization alone, and 3) ≥ 1 point NYHA functional class improvement.

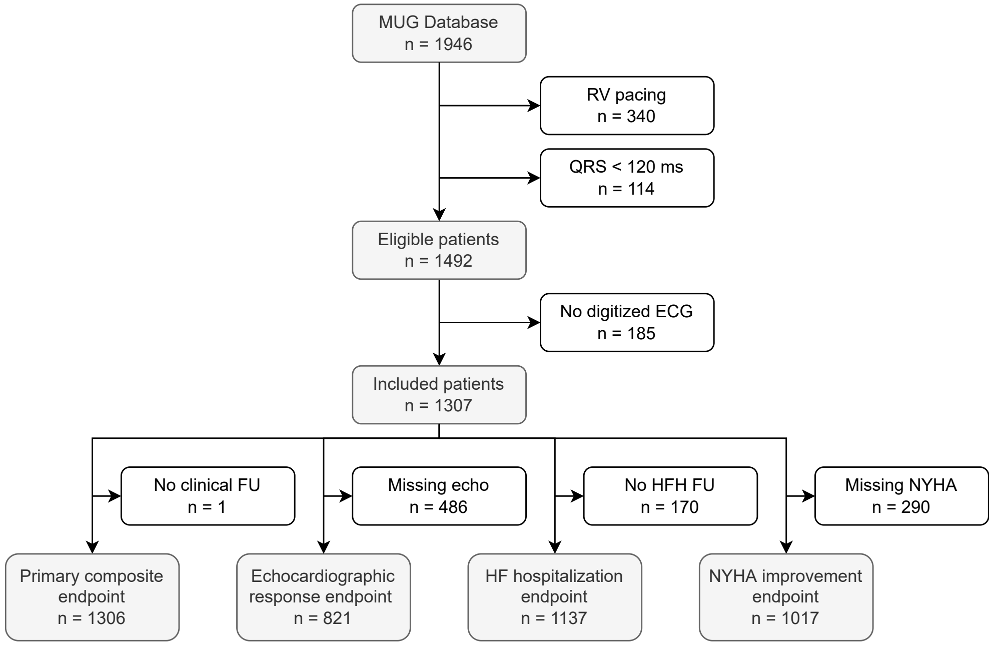


Figure 1. ***Flowchart for the inclusion of patients in this study.*** Legend: ECG, electrocardiogram; FU, follow up; HF, heart failure; HFH, heart failure hospitalization; NYHA, New York Heart Association; RV, right ventricular.

Electrocardiographic data

For all patients, standard 12-lead ECGs were exported and converted into median heart beats using the MUSE ECG system (MUSE version 8, GE Healthcare, Chicago, IL). The median beat data were constructed by aligning all QRS-complexes in the 10 second ECG of the same shape (e.g., excluding premature ventricular complexes), and generating a representative QRS-complex by taking the median voltage.[18] Automated ECG readings were used to derive QRS-duration and other typical ECG parameters. LBBB-morphology was defined according to the 2013 ESC and 2013 AHA criteria at the time **(Supplemental Table 1)**, as previously reported.[7] Using these morphological definitions, indications for CRT implantation were determined according to the current ESC 2021 guidelines.[2] Strauss criteria provide similar risk-stratification as compared to the ESC 2013 criteria[7], and were therefore not evaluated. Without exception, all digitally available ECGs were selected for analysis.

To calculate QRS_{AREA}, first all ECGs were semi-automatically recoded into vectorcardiograms, consisting of three orthogonal leads (X, Y, and Z). To this end, the Kors conversion matrix was used in custom Matlab software (MathWorks Inc).[19] The three orthogonal leads from the vectorcardiogram form a 3D-vector loop, from which QRS_{AREA} was calculated as the sum of the area under the QRS-complex as .

Deep learning approach

A recently developed approach to use deep neural networks in an explainable method, referred to as the FactorECG, was used. Here, the complete median beat ECG is analysed using a variational auto-encoder (VAE), which divided the ECG into morphological features without any assumptions (e.g., an agnostic approach). For this approach, the VAE was pretrained to learn these morphological features (or underlying generative factors) of the ECG, using a dataset of 1,144,331 ECGs from 251,473 consecutive patients that underwent ECG recording in the University Medical Centre Utrecht between July 1991 and August 2020.[16] Overlap in the pretraining cohort and patients included in this study was negligible at 0.04%, and could not influence the results since the VAE was trained unsupervised (i.e., without any knowledge of CRT outcome).

The VAE is a generative artificial intelligence algorithm that consists of three parts: 1) an encoder neural network, 2) the FactorECG (a compressed version of the ECG in only 32 disentangled continuous factors) and 3) the decoder neural network (**Figure 2A**). The goal of the VAE is to learn to 'compress' the ECG, without human interference, into a reduced number of continuous and independent variables that are presumably related to the underlying (patho)physiological generative processes of the ECG. Pretraining of the VAE was performed unsupervised by entering the median beat ECGs into the algorithm and reconstructing the same ECG, while calculating the difference between the original and reconstructed ECG to optimize the network. After training, the first part of the VAE (encoder) can be used to convert any median beat ECG into its FactorECG, the distinctive set of 32 factors that represent that ECG. Importantly, it has been shown before that only 21 of the 32 factors encode significant information.[16] Hence, only these 21 factors were used in subsequent models. In the training step of the current analysis, the 21 continuous FactorECG values for every ECG, as calculated by the encoder, are used in Cox and logistic regression models to perform prediction of the different endpoints (**Figure 2B**).

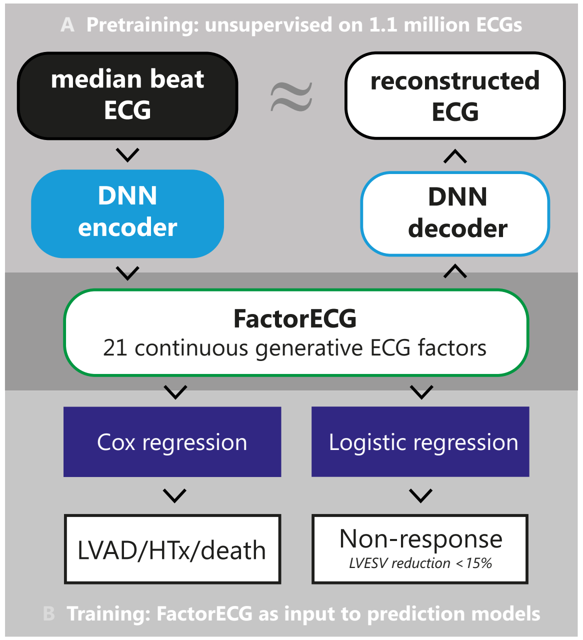


***Figure 2.* Schematic representation of the series of algorithms and processes: a variational auto-encoder, the FactorECG and reconstructions.** A: in the pretraining phase, the variational auto-encoder is trained on a dataset of 1.1 million median beat ECGs from the University Medical Center Utrecht to learn the underlying factors that generate the ECG. In this process, the VAE learns to reconstruct ECGs as accurate as possible using only the FactorECG continuous factors. B: in the training phase, the 21 significant ECG factors for every median beat ECG in the CRT population are obtained using the encoder. These factors are used as input in Cox and logistic regressions models to predict outcome (composite of LVAD implantation, heart transplantation and death) or non-response (LVESV reduction less than 15% after CRT implantation). Legend: DNN; deep neural network, ECG; electrocardiogram, HTx; heart transplantation, LVAD; left ventricular assist device, LVESV; left ventricular end-systolic volume, VAE; variational auto-encoder.

Explainability of the individual ECG factors was achieved by visualizing their influence on the median beat ECG morphology. This was done on the model-level by varying the values of the individual ECG factors between -3 and 3, while reconstructing the ECG using the decoder. The other factors were kept constant, which allows for visualization of the distinct median beat ECG morphology that every factor entails. Moreover, patient-level explanations can be obtained by investigating the FactorECG values of that specific ECG, in combination with the coefficients of the model. This way, we can determine which factors were important in a specific patient to make the prediction. Interactive visualizations of the model are available on https://crt.ecgx.ai. The architecture and training procedures for the FactorECG have been described in detail before.[16]

Statistical analysis

Baseline characteristics were expressed as mean ± standard deviation (SD), or median with interquartile range (IQR), where applicable. Depending on normality of data, differences in continuous variables were assessed using the Student t test or Mann-Whitney U test. Conversely, categorical variables were tested using the χ2 test or Fisher exact.

Models using different guideline criteria, QRS_{AREA} and the per-patient 21 significant standardized FactorECG values, were compared. For the primary endpoint, multivariable Cox proportional hazard models were fitted to take time-to-event into account. For the secondary endpoint, a similar approach was applied, with multivariable logistic regression to predict the binary endpoint of LVESV non-response < 15%. In a second step, the added value of the models to a combination of 13 standard clinical parameters was assessed. Clinical parameters known to be associated with CRT outcome were entered in the multivariable models (i.e. Cox regression for the primary endpoint and logistic regression for secondary endpoint): sex, age, etiology (i.e. ischemic cardiomyopathy [ICM] or non-ICM), weight, height, baseline NYHA class, rhythm (sinus rhythm or atrial fibrillation), baseline LVEF, baseline end-diastolic volume, baseline interventricular mechanical delay (IVMD), hemoglobin, creatinine levels, and presence of diabetes. As there were missing values of some parameters, multivariate imputation using chained equations was performed using only these clinical parameters as input.

For all models, non-linear relationships were investigated using natural cubic splines, and for the Cox models the proportional hazards assumption was verified. Hazard ratios (HR) and odds ratios (OR) were reported to investigate the importance of individual predictors, such as the standardized FactorECG values. Model fit for all models was assessed using Akaike's Information Criterion (AIC), discrimination using Harell's C-statistic, and calibration using the calibration slope. The apparent C-statistic and calibration slope were obtained by applying the model on the original data. Internal validation was performed by using a bootstrap-based optimism estimation technique, where all model development steps are repeated on the 500 bootstrap samples and the model is tested on the original data.[20] The "optimism", which is the mean difference between the performance measure in the original and bootstrapped dataset, was subtracted from the apparent performance measures. These optimism-corrected measures have shown to be an unbiased estimate of the generalizability of the model, without losing any data for training.[21] Confidence intervals (CI) around the performance measures were obtained using 2000 bootstrap samples. All statistical analyses were performed using Python version 3.8. The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis Statement for the reporting of diagnostic models was followed, where applicable.[22]

Results

Baseline characteristics

A real-world CRT population was gathered from three Dutch academic hospitals (n = 1946), of which 1492 were eligible after exclusion for RV-pacing and QRS-duration < 120ms. Of the 1492 patients, 1307 had a digital ECG available in the 90 days before implantation and 1306 were included in the analysis for the primary endpoint (**Figure 1, Table 1**). Median time between ECG and implantation was 1 day [IQR 1-6 days]. The pretrained VAE performed well in the current population, with a Pearson correlation between the original and reconstructed ECG of 0.86. ESC guideline CRT indication, using the ESC 2013 criteria for LBBB, were as follows: class I 737 (56%), class IIa 401 (31%), class IIb and class III 168 (13%) (**Table 1**). When applying the AHA criteria for LBBB, indications were as follows: class I 134 (10%), class IIa 786 (60%), class IIb and class III 385 (30%).

Primary endpoint: combined clinical outcome

A total of 385 patients (30%) reached the primary endpoint of LVAD implantation (n = 11), HTx (n = 4), or all-cause mortality (n = 370). The median follow-up time was 3.5 years [IQR 2.1 – 5.2 years]. Optimism corrected C-statistics were derived for the different predictor sets in predicting the occurrence of the primary endpoint (**Table 2, Supplemental Tables 2 and 4-8**). According to current guideline criteria for CRT implantation, a class I indication was significantly associated with freedom of the primary endpoint, when compared to a non-class I indication. However, this association was only seen when using the ESC (c-statistic 0.57 [95% CI 0.54 – 0.60]), but not with the AHA definition (c-statistic 0.50 [95% CI 0.47 – 0.53]) of LBBB morphology (**Table 2**). A stronger association with outcome was seen using the FactorECG (c-statistic 0.69 [95% CI 0.66 – 0.72], $p < 0.001$ for both AHA and ESC definitions). Moreover, FactorECG had a significantly stronger association with outcome than QRS_{AREA} (c-statistic 0.61 [95% CI 0.58 – 0.64], $p < 0.001$).

	VARIABLE	MISSING – N (%)	OVERALL (N = 1306)
PATIENT DEMOGRAPHICS	**Age (years)**	0 (0)	68.3 [60.0-74.7]
	Male sex – n (%)	0 (0)	919 (70.4)
	Length (cm) – mean (SD)	66 (5.1)	174 (8.9)
	Weight (kg) – mean (SD)	60 (4.6)	81.9 (16.1)
	ICM – n (%)	0 (0)	649 (49.7)
	DM – n (%)	2 (0.1)	328 (25.2)
	Preprocedural NYHA – n (%)	29 (2.2)	
	I		513 (40.2)
	II		672 (52.6)
	III		64 (5.0)
	IV		1226 (93.9)
	ICD – n (%)	0 (0)	
LABORATORY MEASURMENTS	**NT-proBNP (pmol/L) – median [IQR]**	605 (46.3)	1379 [587-2845]
	Hemoglobin (mmol/L) – median [IQR]	436 (33.3)	8.5 [7.8-9.1]
	Creatinine (mmol/L) – median [IQR]	48 (3.7)	102 [83-130]
ELECTROCARDIOGRAPHY	**Sinus rhythm – n (%)**	9 (0.7)	1096 (84.5)
	PR duration (ms) – median [IQR]	216 (16.5)	184.0 [164.0-213.5]
	QRS duration (ms) – median [IQR]	0 (0)	158.0 [146.0-172.0]
	QTc duration (ms) – median [IQR]	0 (0)	486.0 [463.0-510.0]
	LBBB (ESC 2013) – n (%)	0 (0)	1028 (78.7)
	LBBB (AHA) – n (%)	0 (0)	173 (13.3)
	QRS_{AREA} (µVs) – median [IQR]	0 (0)	108.2 [76.0-151.0]
ECOCARDIOGRAPHY	**LVEDV (ml) – median [IQR]**	355 (27.2)	205.0 [157.1-271.0]
	LVESV (ml) – median [IQR]	349 (26.7)	151.0 [113.0-209.0]
	LVEF (%) – median [IQR]	321 (24.5)	24.0 [18.9-30.0]
	IVMD (ms) - median [IQR]	522 (40.0)	45.0 [22.0-64.0]
PROCEDURE CHARACTERISTICS	**CRT-P – n (%)**	0 (0)	80 (6.1)
	LV lead position – n (%)	38 (2.9)	
	Anterior		135 (10.6)
	Lateral		466 (36.8)
	Posterior		667 (52.6)
OUTCOMES	**Duration of follow-up (years) – median [IQR]**	0 (0)	3.48 [2.08-5.24]
	Primary endpoint (LVAD, HTx or death) – n (%)	0 (0)	385 (30)

OUTCOMES

VARIABLE	MISSING – N (%)	OVERALL (N = 1306)
LVESV reduction (%) - median [IQR]	485 (37.1)	20.9 [0.5-41.4]
LVESV non-responder endpoint – n (%)	485 (37.1)	355 (43)
Composite of primary endpoint and heart failure hospitalization – n (%)	169 (12.9)	406 (35.7)
Heart failure hospitalization endpoint – n (%)	169 (12.9)	133 (11.7)
Postprocedural NYHA – n (%)	249 (19.1)	
I		178 (16.8)
II		650 (61.5)
III		216 (20.4)
IV		13 (1.2)
NYHA improvement endpoint	289 (22.1)	509 (50)

Table 1.
Baseline characteristics.
CRT-P, cardiac resynchronization therapy pacemaker; DM, diabetes mellitus; ICD, implantable cardioverter defibrillator; ICM, ischemic cardiomyopathy; IVMD, Interventricular mechanical delay; IQR, interquartile range; LBBB, left bundle branch block; LV, left ventricular; LVEDV, left ventricular end diastolic volume; LVEF, left ventricular ejection fraction; LVESV, left ventricular end systolic volume; NT-proBNP, N-terminal pro-B-type natriuretic peptide; NYHA, New York Heart Association; SD, standard deviation.

Predictors	Outcome		Response	
	C-statistic	95% CI	C-statistic	95% CI
AHA 2013 criteria	0.50	[0.47-0.53]	0.56	[0.53–0.60]
ESC 2013 criteria	0.57	[0.54-0.60]	0.61	[0.57–0.64]
QRS_{AREA}	0.61	[0.58-0.64]	0.70	[0.67–0.74]
FactorECG	0.69	[0.66-0.72]	0.69	[0.65–0.72]
Clinical	0.69	[0.67-0.72]	0.67	[0.64–0.71]
QRS_{AREA} / Clinical	0.71	[0.68-0.74]	0.72	[0.68-0.75]
FactorECG / Clinical	0.72	[0.69-0.75]	0.70	[0.67–0.74]

Table 2.
Optimism corrected C-statistic for outcome and response.
AHA, American Heart Association; ESC, European Society of Cardiology; ICM, ischemic cardiomyopathy; LBBB, left bundle branch block.

When subdividing QRS_{AREA} and FactorECG in four quartiles, better discriminative performance for the occurrence of the primary endpoint was achieved using FactorECG (**Figure 3**). A significantly higher event free survival at three years was seen in the lowest risk FactorECG group as compared to QRS_{AREA} ≥ 150 µVs (94% versus 89%; log rank p = 0.01). Additionally, three-year event free survival for the highest risk FactorECG quartile was significantly worse than in patients with QRS_{AREA} < 75 µVs (63% versus 73%; *log rank p < 0.005*).

Secondary endpoint: echocardiographic non-response

Pre- and postprocedural echocardiograms were available in 821 patients. Long-term echocardiographic non-response was observed in 355 patients (43%). All evaluated models were significantly associated with echocardiographic non-response (**Table 2, Supplemental Tables 3 and 9-13**). However, guideline classifications performed the worst, using either the ESC (c-statistic 0.61 [95% CI 0.57 – 0.64]) or AHA definition (c-statistic

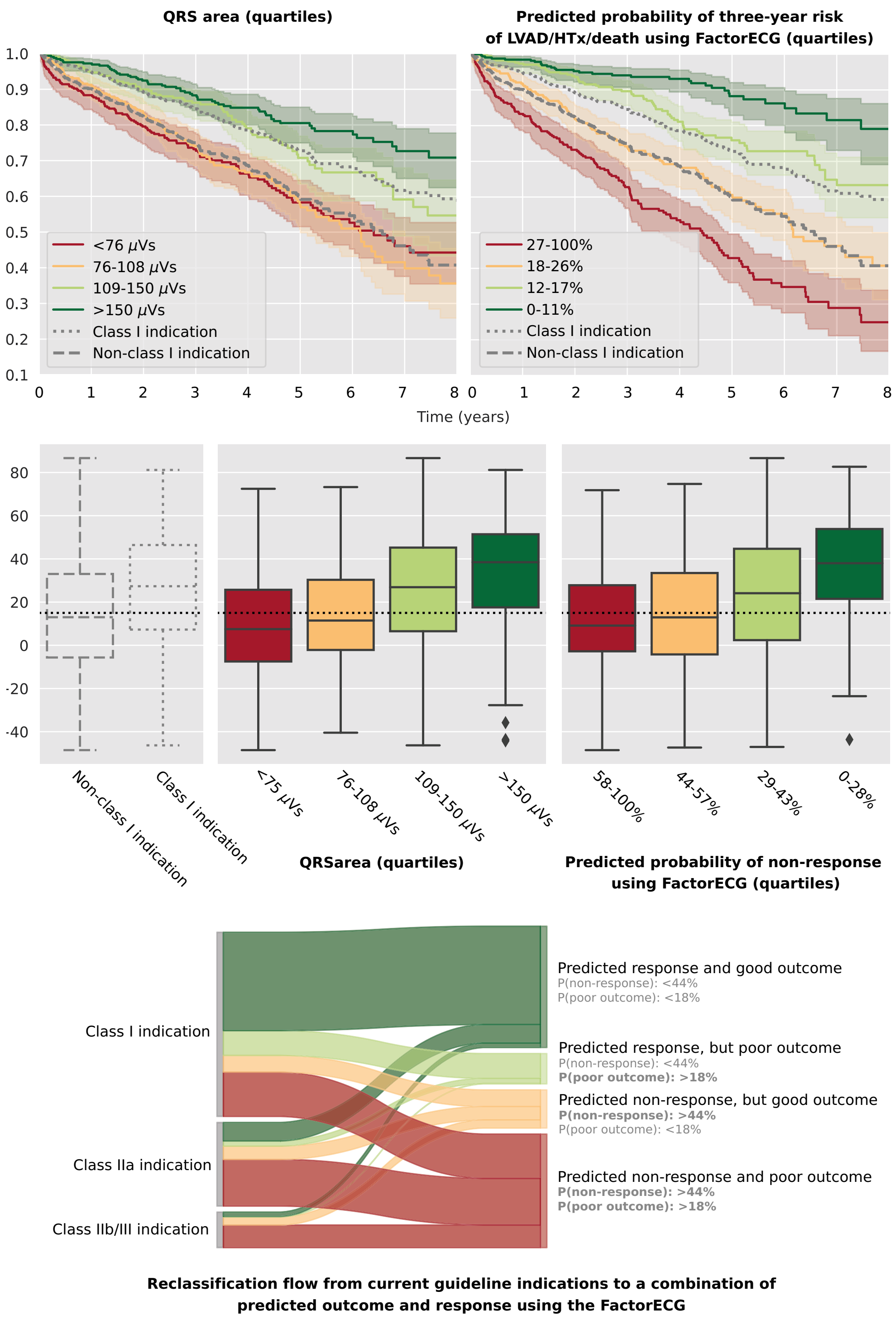


Figure 3.
Clinical utility of FactorECG and QRS$_{AREA}$ in CRT. QRS$_{AREA}$ and FactorECG predicted probabilities were divided into four quartiles of equal size. Quartiles of FactorECG better differentiate clinical outcome as compared to QRS$_{AREA}$ and guidelines using the ESC criteria of LBBB (panel A). Similar associations with echocardiographic response were seen when compared to QRS$_{AREA}$, while still outperforming guideline criteria (panel B). Reclassification flow form the guidelines to the FactorECG predictions is shown in panel C. Here, a combination of predicted clinical outcome and response is assessed by setting the probability cut-off at 50% of the data. Probability cut-offs in panel C therefore correspond to the upper two and lower two quartiles in panels A and B combined. Legend: ECG; electrocardiogram, HTx; heart transplantation, LVAD; left ventricular assist device, LVESV; left ventricular end-systolic volume.

0.56 [95% CI 0.53 – 0.60]) of LBBB-morphology. FactorECG (c-statistic 0.69 [95% CI 0.65 – 0.72]) and QRS$_{AREA}$ (c-statistic 0.70 [95% CI 0.67 – 0.74]) had similar associations with non-response (p = 0.12) but were both significantly stronger associated with response than either guideline recommendation ($p < 0.001$, **Figure 3**). Differences in the extent of reverse remodelling, stratified according to four groups of FactorECG and QRS$_{AREA}$, was similar (**Figure 3**).

Tertiary endpoints

Availability of tertiary endpoints are summarised in **Figure 1** and **Table 1.** FactorECG was significantly associated with the composite of the primary endpoint combined with HF hospitalization (c-statistic = 0.67 [95% CI 0.65 – 0.70], and HF hospitalization alone (c-statistic = 0.70 [95% CI 0.66 – 0.74]), outperforming QRS$_{AREA}$ and the guideline criteria ($p < 0.001$ for all comparisons (**Supplemental Tables 14-24**). None of the models showed additional predictive value for prediction ≥ 1 point NYHA improvement as compared to a baseline model that only consisted of preprocedural NYHA class (**Supplemental Tables 14, 25-29**).

Subgroup analysis

Performance of FactorECG and QRS$_{AREA}$ were compared, stratified by known subgroups associated with clinical outcome (**Table 3**). The strongest association of FactorECG was observed in patients with non-ICM (c-statistic 0.77 [95%CI 0.73 – 0.81]), which was significantly higher as compared to QRS$_{AREA}$ (c-statistic 0.62 [95%CI 0.57 – 0.67]). Using the ESC definition of LBBB-morphology, FactorECG outperformed QRS$_{AREA}$ in patients with LBBB (c-statistic 0.71 [95%CI 0.68 – 0.74] versus c-statistic 0.61 [95% CI 0.58 – 0.65]), and non-LBBB (c-statistic 0.66 [95% CI 0.60 – 0.71] versus c-statistic 0.52 [95% CI 0.46 – 0.58]). The same observation was made when evaluating patients with an intermediate QRS-duration, below 150 ms, and patients with ICM. Importantly, FactorECG and QRS$_{AREA}$ demonstrated comparable associations with echocardiographic response, regardless of the subgroup analysed (**Table 3**).

Subgroup	Outcome (C-statistic [95% CI])		Response (C-statistic [95% CI])	
	QRS$_{AREA}$	FactorECG	QRS$_{AREA}$	FactorECG
Male	0.60 [0.57-0.63]	0.67 [0.64-0.70]	0.69 [0.65-0.73]	0.70 [0.66-0.74]
Female	0.61 [0.53-0.69]	0.77 [0.71-0.83]	0.70 [0.63-0.77]	0.73 [0.66-0.79]
ICM	0.58 [0.54-0.62]	0.63 [0.60-0.67]	0.65 [0.59-0.70]	0.67 [0.61-0.72]
Non-ICM	0.62 [0.57-0.67]	0.77 [0.73-0.81]	0.72 [0.67-0.77]	0.74 [0.70-0.79]
LBBB*	0.61 [0.58-0.65]	0.71 [0.68-0.74]	0.71 [0.67-0.75]	0.73 [0.69-0.76]
Non-LBBB*	0.52 [0.46-0.58]	0.66 [0.60-0.71]	0.53 [0.43-0.63]	0.55 [0.46-0.65]
QRS ≥150ms	0.62 [0.58-0.66]	0.70 [0.66-0.73]	0.71 [0.67-0.75]	0.73 [0.69-0.77]
QRS <150ms	0.58 [0.53-0.63]	0.72 [0.67-0.76]	0.62 [0.55-0.70]	0.67 [0.60-0.73]

Table 3. *Optimism corrected C-statistic in various subgroups. ICM, ischemic cardiomyopathy; LBBB, left bundle branch block. ** Morphology evaluated according to ESC 2013 criteria

Additional value of clinical model

Readily available patient characteristics, known to be associated with CRT outcome, were entered into a clinical model [11]. The clinical model was significantly associated with outcome (c-statistic 0.69 [95% CI 0.67 –0.72]) and response (c-statistic 0.60 [95%CI 0.56 – 0.64]) (**Table 2, Supplemental Table 4-13**). However, for both endpoints, the ECG-only FactorECG model demonstrated similar associations as compared to the clinical model (p = 0.48 and p = 0.10, respectively). For outcome, the addition of a 13-variable clinical model significantly improved upon QRS$_{AREA}$ (Δ c-statistic 0.10, p < 0.001), whereas its addition to FactorECG was of limited added value (Δ c-statistic 0.03, p < 0.001). By contrast, concerning echocardiographic non-response, the added value of the clinical model was negligible (Δ c-statistic 0.01, p = 0.002).

Explainable deep learning through factor visualisation

ECG factors that were significantly associated with outcome and non-response are summarised in **Figure 4**. Exact hazard ratios for outcome and odds ratios for non-response are summarised in **Supplemental Table 5 and 10**, respectively. Visualisations of the most important ECG factors, using factor traversals, are shown in **Figure 5**, whereas **Supplemental Figure 1** displays complete 12-lead visualisation of all factors. Factors associated with '*both*' non-response and poor outcome were interpreted as follows: F_1 (absent QRS-notching and ST-deviation, but lateral T-wave inversion), F_9 (transition from LBBB-morphology to more right bundle branch block-morphology with smaller right precordial S-wave amplitudes), F_{10} (increased ventricular rate), and F_{19} (decreased anterior QS-amplitude and lateral notched R). Importantly, F_8 and F_{15} (increased PR-interval and P-wave duration) were only associated with worse outcome, whereas F_5 (decreased QRS duration and JTc-interval) and F_{26} (decreased QRS-duration and amplitude of the LBBB-morphology)

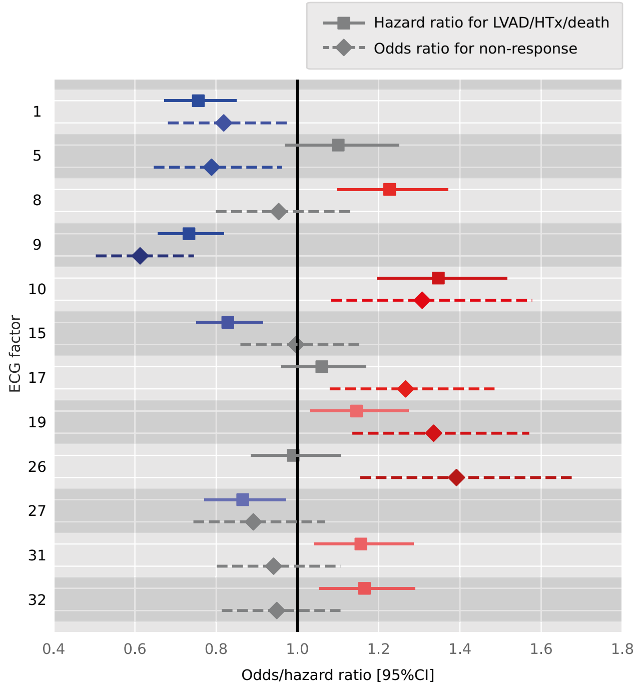


Figure 4. *Hazard and odds ratios for the models predicting either the clinical endpoint or echocardiographic non-response (LVESV reduction < 15%) using the ECG factors as the only input for the model.* Colors (red and green) correspond with factor traversal reconstructions in Figure 5. All ECG factors were standardized and hazard and odds ratios can be interpreted as importance scores. Legend: ECG; electrocardiogram, HTx; heart transplantation, LVAD; left ventricular assist device, LVESV: left ventricular end-systolic volume.

was only associated with non-response. Similar factors (F_1, F_9 and F_{19}), mostly representing reduced QRS and T-wave voltage with increased QT duration, were found to be predictive for HF hospitalization when compared to the model for the primary outcome alone (**Supplementary Table 21**). However, F_{25}, which represents reduced QRS duration, was also predictive for HF hospitalization.

Clinical applicability using risk groups

Using a combination of predictions of the FactorECG algorithm for both echocardiographic non-response and 3-year clinical outcome, four distinct groups could be identified to assist patient selection (**Supplemental Table 30**). Here, QRS_{AREA} could not differentiate between good and poor outcome in echocardiographic responders (median QRS_{AREA} 151 versus 152 µVs, respectively) or non-responders (median QRS_{AREA} 84 versus 83 µVs, respectively).

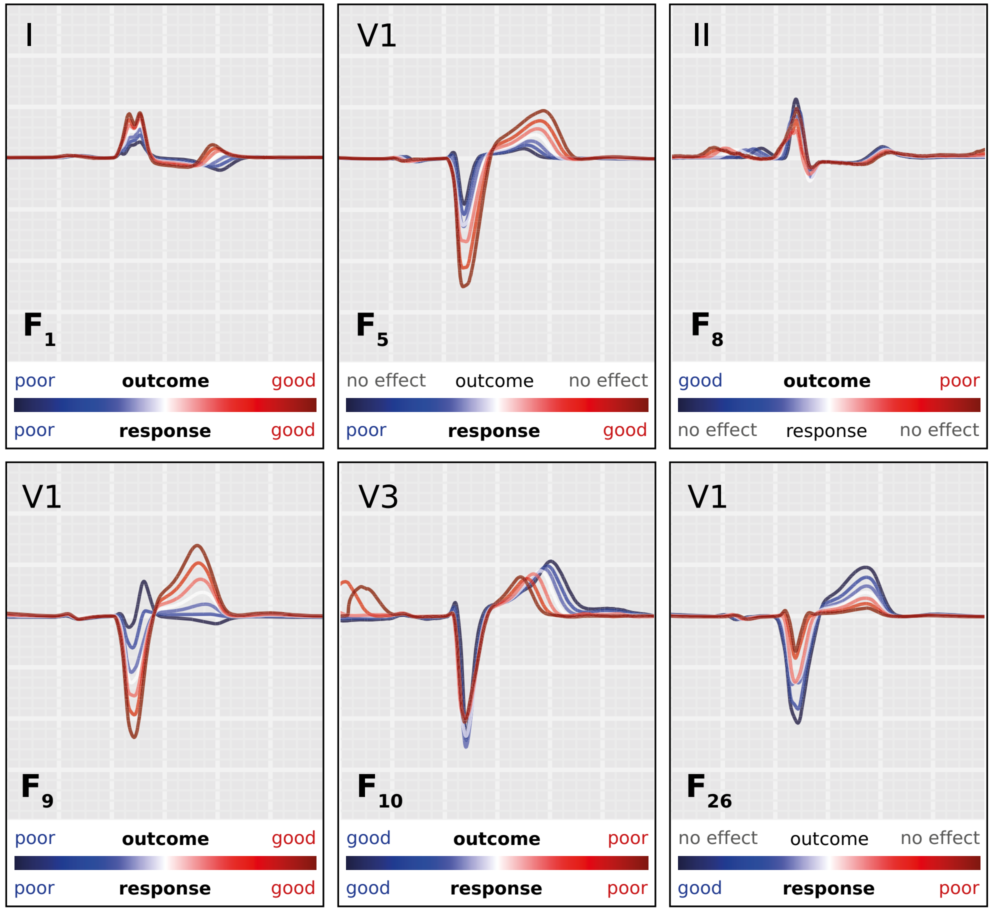


Figure 5.
Factor traversals of a subset of the ECG factors associated with both, clinical outcome (composite endpoint of LVAD/HTx/death) and echocardiographic response (LVESV reduction > 15%).
In each graph the corresponding factor is varied from -3 (blue) to 3 (red) standard deviations from the mean of 0 (white line), which represents a mean ECG in the CRT population.
For each factor, the lead showing the most easily interpretable effect is shown in the upper left corner. Complete 12-lead ECG of all factors can be found in Supplemental Figure 1. Legend: ECG; electrocardiogram, HTx; heart transplantation, LVAD; left ventricular assist device, LVESV; left ventricular end-systolic volume.

In the first group, with both predicted response and good outcome (n = 338), 76% of the patients were responders, and only 14% experienced the primary endpoint during follow-up. In the second group of poor 3-year outcome despite an echocardiographic response (n = 72), patients were more frequently male, had ICM, higher NT-proBNP, high QRS-duration, and the worst ESV and LVEF. Conversely, in the third group, CRT non-responders with good clinical outcome regardless (n = 96) were predominantly characterised by shorter QRS-duration, lowest LVESV, and highest LVEF. In the fourth group of patients, with both poor outcome and non-response, significant more ICM, NYHA III, and non-LBBB was observed as compared to the other subgroups. In this worst performing subgroup (n = 314) the primary endpoint occurred in 46% of the patients during follow-up, and response occurred in only 36% of patients.

In contrast, when using the current ESC guidelines for selection of patients eligible for CRT, in class I patients (n = 499) response occurred in 65% and the primary outcome endpoint in 26% during follow-up. In patients with class IIa (n = 226) or IIb/III (n = 96) indications, response occurred in 50% and 33%, and the primary outcome endpoint in 35% and 37%, respectively. A comparison of the classification in the four FactorECG groups and the guideline-based groups can be found in **Figure 2C**.

Discussion

In this large, multicentre, real-world dataset, an explainable deep learning-based algorithm (the FactorECG) was predictive for long-term clinical outcome, HF hospitalization, and echocardiographic non-response after CRT implantation. FactorECG outperformed contemporary guideline criteria and vectorcardiographic QRS$_{AREA}$ for clinical outcome, and HF hospitalization. Importantly, only a readily available 12-lead ECG is required since little added value was obtained using additional clinical input variables. The user-independent analysis and automated visualization of key ECG features allows for patient-specific interpretation of the algorithm (**Figure 6**), which may facilitate its adoption into clinic practice as a valuable alternative for the selection of CRT candidates. Lastly, an online visualisation tool was created to provide interactive visualizations (https://crt.ecgx.ai).

Deep learning-based prediction of outcome

For the first time, deep learning has been used to predict clinical outcome after CRT using only the raw preprocedural ECG (c-statistic 0.69 [95% CI 0.66-0.72]). In contrast, previous studies aimed to predict CRT outcome using machine learning to unify a vast number of clinical variables in a single model. The SEMMELWEIS-CRT score combined 33 clinical variables for the prediction of all-cause mortality, reporting a mean internally calculated c-statistic of 0.69, derived from 1510 patients in a single centre.[11] Similarly, three other studies combined a plethora of pre-implantation characteristics, including ECG and complex echocardiography data, totalling 19, 45 or even 77 variables.[12–14] Another study compared an unsupervised principal component analysis model with QRS$_{AREA}$.[23] Here, similar results for QRS$_{AREA}$ (HR = 0.46 [95% CI 0.39–0.55]) and their model (HR = 0.45 [95% CI 0.38–0.53]) were seen for the composite endpoints of death, LVAD, or HTx.

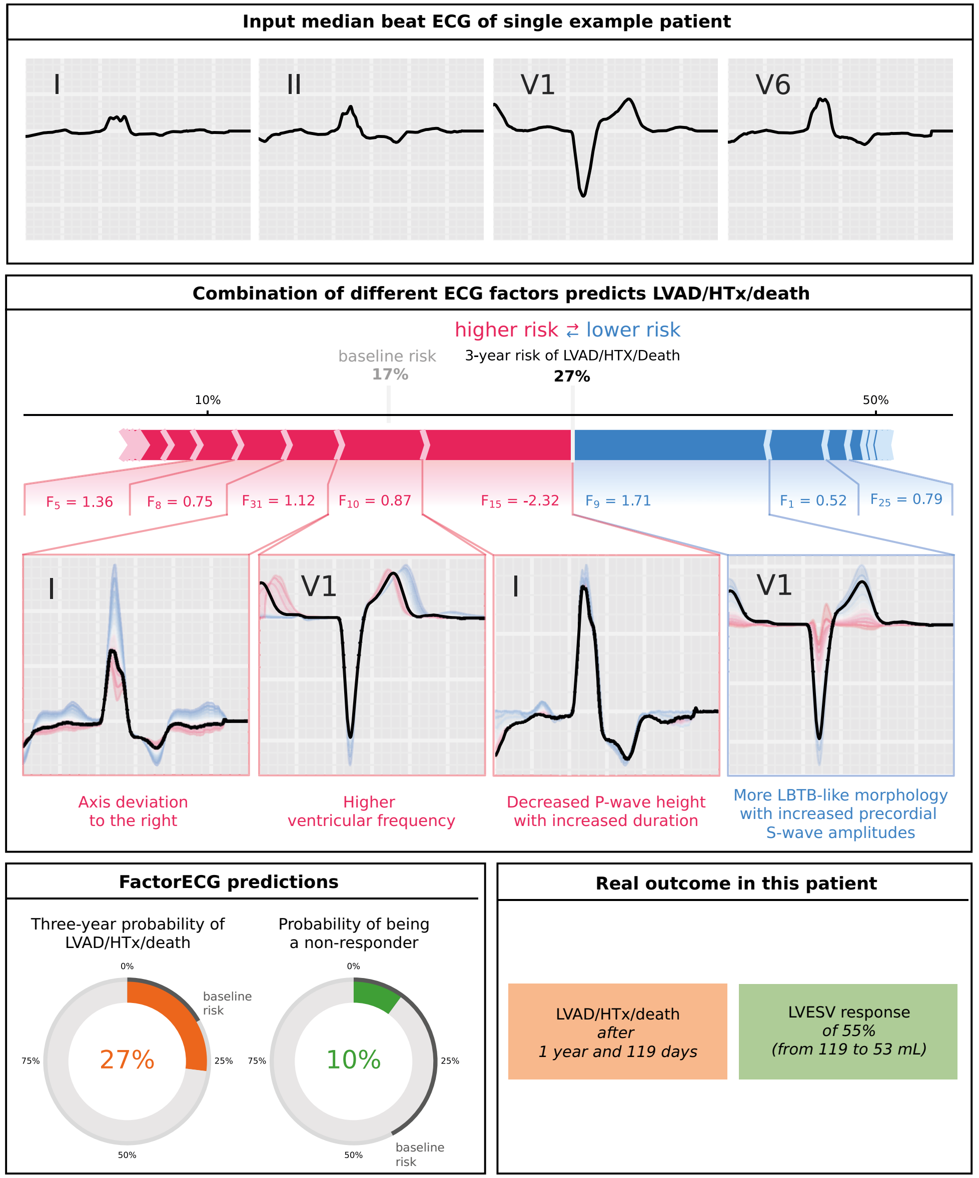


***Figure 6.* Patient-level example of a prediction with the FactorECG explanation.** A standard 12-lead ECG is entered into a deep learning model, which automatically translates this ECG into its FactorECG containing all distinct features. These factors are entered into the Cox and logistic regression models and predicted probabilities for both LVAD/HTx/death and non-response are shown to the user. This patient responded well to CRT, but died within three years regardless. Despite presence of a 'typical' LBBB morphology (F9), FactorECG demonstrates that this prediction of high probability of poor outcome is driven by increased ventricular frequency (F10), long PR-interval with broad P-wave (F15), and axis deviation to the right (F31). Legend: ECG; electrocardiogram, HTx; heart transplantation, LVAD; left ventricular assist device, LVESV, left ventricular end-systolic volume.

Differences in primary clinical endpoints in the aforementioned studies complicate a direct comparison with the present study. However, similar or better performance was observed with respect to predicting clinical outcomes, without relying on complex *'statistical'* models.[11,13] Moreover, our approach outperformed QRS$_{AREA}$ with respect to clinical outcome, whereas unsupervised machine learning of baseline QRS-waveforms previously failed to do so.[23] Most importantly, all previously proposed models require collection and calculation of many clinical variables, which are highly operator dependent, cumbersome, and likely to dissuade clinicians to rapidly adopt such an approach.[11–14] Although significant added benefit was obtained upon addition of a clinical model to QRS_{AREA}, the increase in model performance was 3-fold smaller for FactorECG. Rather, our proposed approach requires only a standard 12-lead ECG, without heavily depending on additional clinical input variables, or manual selection of the QRS-complex. It is therefore conceivable that the clinical practicality of our ECG-only approach outweighs the limited benefit of increasing the c-statistic by 0.03 using 13 clinical variables. For research purposes, an online tool has been developed where the ECG can be uploaded, and predictions for CRT outcome can be made (https://crt.ecgx.ai and https://encoder.ecgx.ai).

Echocardiographic and functional response

The proportion of 43% non-responders is in accordance with previous literature and highlights the need for better patient selection.[3,17] In our study, a head-to-head comparison of FactorECG and QRS_{AREA} provided similar results for the prediction of echocardiographic non-response (c-statistic 0.69 [95% CI 0.65 – 0.72] and 0.70 [95% CI 0.67 – 0.74], p = 0.12). However, next to identifying the electrical substrate on the ECG, characterisation of the extent of mechanical impairment is of importance as well, especially in patients with ICM. In fact, adding strain-based parameters of mechanical dyssynchrony to QRS_{AREA} improves prediction of 6-month response (c-statistic 0.76), and is therefore also likely of added value to FactorECG.[3] Simple multivariate logistic regression models, consisting of only four variables, have also shown to be associated with sustained echocardiographic response (c-statistic 0.774), a surrogate marker of stable disease remission.[3] None of the described models provided added value to predict NYHA improvement, likely because NYHA is non-specific, and its assessment is subjective and prone to bias.[24]

Identifying ECG features beyond the QRS-complex

FactorECG improves upon heatmap-based attempts to make deep learning explainable, as such approaches merely highlight *'where'* on the ECG significant features are detected but provide no information on which morphological change explains the prediction.[16] Rather, FactorECG allows for *'quantifiable'* identification of specific ECG features, rendering physicians able to evaluate and confirm the clinical rationale of said features. This is reflected by our results that confirm the known importance of LBBB-morphology and QRS_{AREA} for the prediction of echocardiographic response *[3,10]*. Using FactorECG, all types of LV conduction delay, as reflected in the QRS-complex, can be represented by combining ECG Factors 5, 9, 19, and 27. Interestingly, although QRS_{AREA} was associated with outcome, ECG factors that incorporate QRS-duration were not associated with outcome (**Figure 5**). This may be because, in the presence of sufficient electrical substrate, a subset of patients with moderate QRS-prolongation are still likely to respond.[2,3,5,25] This is also underscored by our results, since FactorECG also predicted outcome in patients with QRS-duration < 150 ms (c-statistic 0.72 [0.67-0.76]). Likewise, when corrected for various other ECG features, no significant association with QRS-duration and outcome remains, as also reported previously.[26]

Visualisation of ECG factors also identified various other ECG characteristics known to be to be associated with outcome and/or response, including the PR-interval and P-wave duration (F_8 and F_{15}). The fact that correction of atrioventricular dromotropathy increases LV filling and LV pump function may explain the increased risk of poor outcome in the present study.[27] Similarly, prolonged P-wave duration > 120 ms, indicating interatrial myopathy, has been linked to supraventricular arrhythmias, stroke, and mortality.[28] In addition, the QRS-T angle[29], JTc-interval[30], and T-wave area[31] have been raised as potentially important determinants of response or outcome. However, various other subtle markers of ischemia, dyssynchrony, or risk of arrhythmia may be represented by FactorECG.

Indeed, when evaluated by itself, a large number of other factors can be identified from the ECG.[7] Unfortunately, accurately identifying these factors, and interpreting their interrelated meaning, is highly complex. In the first place because there is lack of consensus[7] and inter-observer disagreement[8] as to what truly defines LBBB-morphology. Matters are further complicated when septal and LV activation patterns are concealed, or wrongly mimicked.[9] Lastly, various unknown ECG-criteria may have remained undetected. Interpretation of the LBBB ECG is therefore complex and misleading. In this regard, FactorECG allows for a unified and agnostic approach, is user-independent, and inherently explainable**.**

Clinical implications

The FactorECG algorithm can be used in every patient that is considered for CRT. When provided with the baseline ECG, the patient-specific ECG-factors that are associated with response and outcome are identified and combined into an individual risk score, and a patient-specific visualisation of these factors is given (**Figure 6**). Hence, assessment of the electrical substrate as a "continuum", rather than the current binary classification of LBBB morphology, is achieved. While similar in size, the CRT non-response *and* poor outcome subgroup, as predicted by the FactorECG, performed worse than patients without a class I indication for CRT according to the ESC guideline criteria. Importantly, 39% of patients in this worst performing subgroup had a class I indication (**Figure 2C**). FactorECG therefore enables better classification of patient eligibility, without compromising the total proportion of patients deemed suitable for CRT implantation.

Future perspectives

Our self-contained ECG-based model was especially effective in females and the non-ICM population (c-statistic = 0.77 for both), but additional clinical variables are required to improve performance in patients with ICM. A future study will address the importance of adding strain-based mechanical dyssynchrony to FactorECG.[3] In addition, optimal placement of the LV lead is of importance to enhance response in CRT patients. This is particularly important in patients with scar, but also in patients with heterogenous LV electrical activation.[9] In the future, FactorECG may use ECG-derived data to identify the site of latest electrical activation, thereby guiding LV lead implantation.[9] Moreover, the results need to be validated in a patient group that received a CRT-P device, as recent reports have shown similar survival between patients with a CRT-P and a CRT-D.[32]. Lastly, prospective studies with FactorECG are warranted to acquire CE certification, allowing its use as a medical device.

Strengths and limitations

Our data was derived from a large multicentre database, and thereby represents a real-world population. Internal validation by means of bootstrapping was performed, which allows for unbiased validation in the complete dataset, and is therefore considered the recommended approach for internal validation of any prediction model.[20,21,32] As a result, performance was not assessed in a single train-test split, because this approach only validates an example model in an arbitrarily chosen and small data subset and produces a poorer model by default.[33] We acknowledge that external validation in datasets with a different patient population remains important to investigate the generalizability of our results. However, by using regular prediction models (i.e., logistic regression and Cox regression) with a limited number of predicting variables as input (only the 21 factors), the risk of overfitting is low. Although ECG data was derived from a single vendor, previous studies have shown that ECG-based deep learning results generalize well to other cohorts with different ECG manufacturers.[34,35] Despite QRS_{AREA} being calculated manually, performance is identical relative to automated calculation.[36] Although measurement of LVESV is user-dependent, excellent intra- and inter-observer reliability was previously demonstrated in a subpopulation of this study.[3]

Many clinicians regard deep learning as a 'black box', which limits trust in such algorithms.[16] However, our approach to make the model inherently explainable may abate this concern, and increase willingness to facilitate clinical adoption of the FactorECG. Although an overall c-statistic of 0.69 leaves room for improvement, our approach is unique in its clinical practicality, with better risk-stratification than $QRS_{AREA.}$ Addition of a few important clinical values might further increase the predictive value of FactorECG. Especially use of strain-parameters has shown to be highly predictive, also in addition to QRS_{AREA}[3], or when used in machine-learning models.[12] As a result, no direct comparison with pre-existing scores could be performed.[11] Conversely, our approach only requires a standard 12 lead ECG, and no advanced and highly user dependent measurements are needed. Lastly, ethnicity and cause of death were not systemically gathered, and our results cannot be generalized to patients receiving upgrade to CRT.

Conclusion

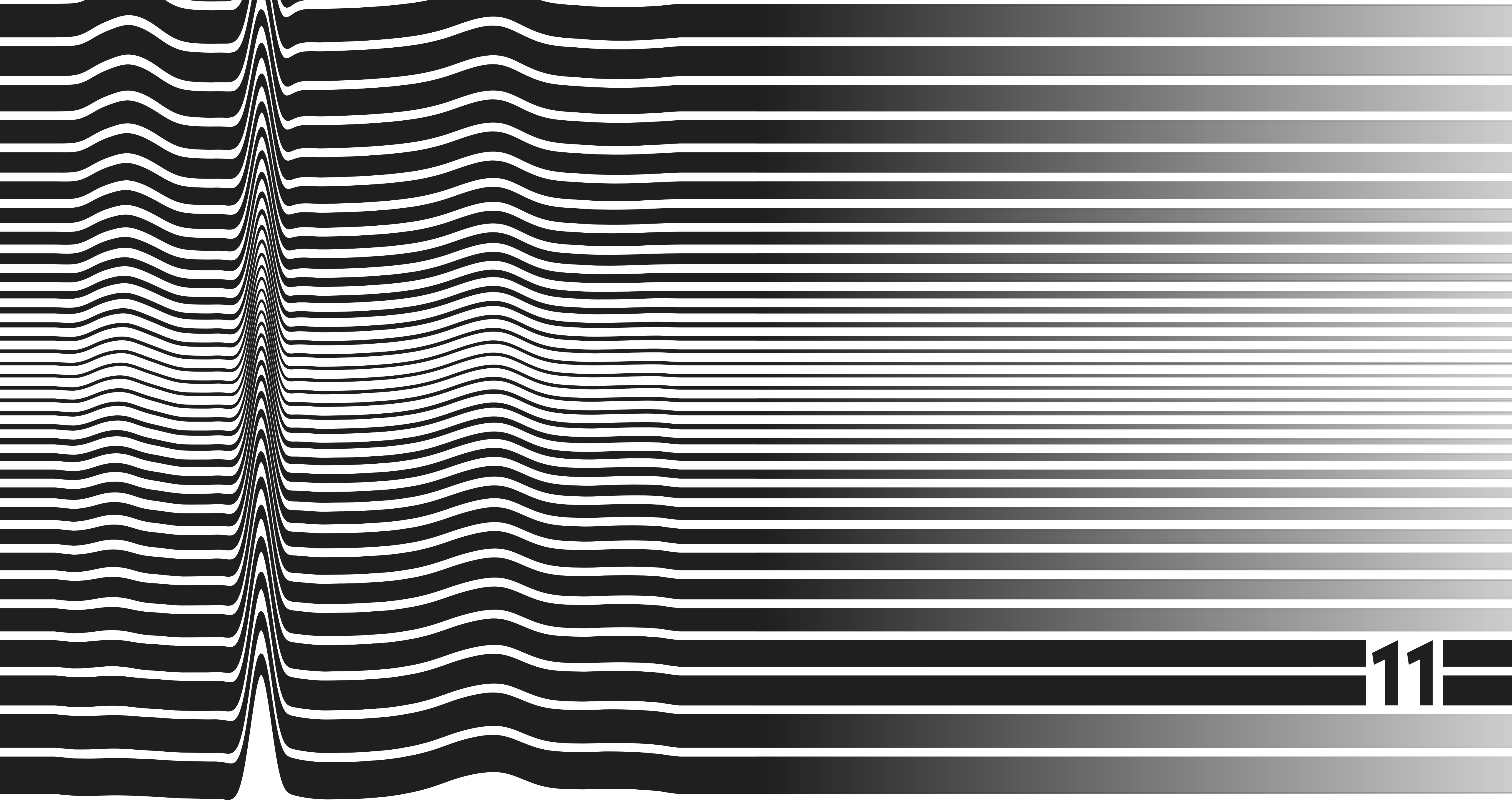
The FactorECG, an inherently explainable and end-to-end automated deep learning model, can accurately predict long-term clinical outcome, HF hospitalization, and echocardiographic non-response in patients eligible for CRT. Moreover, it outperformed contemporary guideline ECG-criteria and QRS_{AREA} with superior discriminative ability. This approach is based solely on a standard 12-lead ECG, without heavily relying on additional clinical parameters, and visualises patient-specific key features associated with outcome and response. Besides QRS-morphology, T-wave amplitude and inversion, ventricular rate, and PR-interval and P-wave duration were identified as important ECG factors. The FactorECG thereby facilitates personalised decision making in CRT, while being easy-to-use, allowing rapid uptake for everyday clinical practice.

REFERENCES

1. Vernooy K, van Deursen CJM, Strik M, Prinzen FW. Strategies to improve cardiac resynchronization therapy. *Nat Rev Cardiol* 2014;11:481–493.
2. Glikson M, Nielsen JC, Kronborg MB, et al. 2021 ESC Guidelines on cardiac pacing and cardiac resynchronization therapy: Developed by the Task Force on cardiac pacing and cardiac resynchronization therapy of the European Society of Cardiology (ESC) With the special contribution of the European Hear. *Eur Heart J* 2021.
3. Wouters PC, van Everdingen WM, Vernooy K, et al. Does mechanical dyssynchrony in addition to QRS area ensure sustained response to cardiac resynchronization therapy? *Eur Hear journal Cardiovasc Imaging* 2021.
4. Sipahi I, Chou JC, Hyden M, et al. Effect of QRS morphology on clinical event reduction with cardiac resynchronization therapy: meta-analysis of randomized controlled trials. *Am Heart J* 2012;163:260-7.e3.
5. Sipahi I, Carrigan TP, Rowland DY, et al. Impact of QRS Duration on Clinical Event Reduction With Cardiac Resynchronization Therapy: Meta-analysis of Randomized Controlled Trials. *Arch Intern Med* 2011;171:1454–1462.
6. Salden OAE, Vernooy K, van Stipdonk AMW, et al. Strategies to Improve Selection of Patients Without Typical Left Bundle Branch Block for Cardiac Resynchronization Therapy. *JACC Clin Electrophysiol* 2020;6:129 LP – 142.
7. Van Stipdonk AMW, Hoogland R, Horst I ter, et al. Evaluating Electrocardiography-Based Identification of Cardiac Resynchronization Therapy Responders Beyond Current Left Bundle Branch Block Definitions. *JACC Clin Electrophysiol* 2020;6:193–203.
8. Van Stipdonk AMW, Vanbelle S, Horst IAH Ter, et al. Large variability in clinical judgement and definitions of left bundle branch block to identify candidates for cardiac resynchronisation therapy. *Int J Cardiol* 2019;286:61–65.
9. Wouters PC, Vernooy K, Cramer MJ, et al. Optimizing lead placement for pacing in dyssynchronous heart failure: The patient in the lead. *Hear Rhythm* 2021;18:1024–1032.
10. Ghossein MA, van Stipdonk AMW, Plesinger F, et al. Reduction in the QRS area after cardiac resynchronization therapy is associated with survival and echocardiographic response. *J Cardiovasc Electrophysiol* 2021;32:813–822.
11. Tokodi M, Schwertner WR, Kovács A, et al. Machine learning-based mortality prediction of patients undergoing cardiac resynchronization therapy: the SEMMELWEIS-CRT score. *Eur Heart J* 2020;41:1747–1756.
12. Cikes M, Sanchez-Martinez S, Claggett B, et al. Machine learning-based phenogrouping in heart failure to identify responders to cardiac resynchronization therapy. *Eur J Heart Fail* 2019;21:74–85.
13. Kalscheur MM, Kipp RT, Tattersall MC, et al. Machine Learning Algorithm Predicts Cardiac Resynchronization Therapy Outcomes. *Circ Arrhythmia Electrophysiol* 2018;11:e005499.
14. Liang Y, Ding R, Wang J, et al. Prediction of response after cardiac resynchronization therapy with machine learning. *Int J Cardiol* 2021;344:120–126.
15. Van de Leur RR, Boonstra MJ, Bagheri A, et al. Big Data and Artificial Intelligence: Opportunities and Threats in Electrophysiology. *Arrhythmia Electrophysiol Rev* 2020;9:146–154.
16. Van de Leur RR, Bos MN, Taha K, et al. Improving explainability of deep neural network-based electrocardiogram interpretation using variational auto-encoders. *Eur Hear*

J - Digit Heal 2022:ztac038.

17. Foley PWX, Chalil S, Khadjooi K, et al. Left ventricular reverse remodelling, long-term clinical outcome, and mode of death after cardiac resynchronization therapy. *Eur J Heart Fail* 2011;13:43–51.
18. GE Healthcare. Marquette 12SL ECG Analysis Program Physician's Guide. 2012. Chicago, United States: Accessed April 30, 2020. https://www.gehealthcare.com/products/diagnostic-cardiology/marquette-12slhttps://www.gehealthcare.com/products/diagnostic-cardiology/marquette-12sl (30 April 2020)
19. Kors JA, van Herpen G, Sittig AC, van Bemmel JH. Reconstruction of the Frank vectorcardiogram from standard electrocardiographic leads: diagnostic comparison of different methods. *Eur Heart J* 1990;11:1083–1092.
20. Harrell FEJ, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15:361–387.
21. Steyerberg EW, Harrell FEJ, Borsboom GJ, et al. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 2001;54:774–781.
22. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD Statement. *Br J Surg* 2015;102:148–158.
23. Feeny AK, Rickard J, Trulock KM, et al. Machine Learning of 12-Lead QRS Waveforms to Identify Cardiac Resynchronization Therapy Patients With Differential Outcomes. *Circ Arrhythm Electrophysiol* 2020;13:e008210.
24. Raphael C, Briscoe C, Davies J, et al. Limitations of the New York Heart Association functional classification system and self-reported walking distances in chronic heart failure. *Heart* 2007;93:476–482.
25. Van Stipdonk AMW, Horst I Ter, Kloosterman M, et al. QRS Area Is a Strong Determinant of Outcome in Cardiac Resynchronization Therapy. *Circ Arrhythm Electrophysiol* 2018;11:e006497.
26. Khidir MJH, Delgado V, Ajmone Marsan N, Schalij MJ, Bax JJ. QRS duration versus morphology and survival after cardiac resynchronization therapy. *ESC Hear Fail* 2017;4:23–30.
27. Salden FCWM, Huntjens PR, Schreurs R, et al. Pacing therapy for atrioventricular dromotropathy: a combined computational-experimental-clinical study. *Europace* 2021.
28. Martínez-Sellés M, Elosua R, Ibarrola M, et al. Advanced interatrial block and P-wave duration are associated with atrial fibrillation and stroke in older adults with heart disease: the BAYES registry. *Europace* 2020;22:1001–1008.
29. Sweda R, Sabti Z, Strebel I, et al. Diagnostic and prognostic values of the QRS-T angle in patients with suspected acute decompensated heart failure. *ESC Hear Fail* 2020;7:1817–1829.
30. Maass AH, Vernooy K, Wijers SC, et al. Refining success of cardiac resynchronization therapy using a simple score predicting the amount of reverse ventricular remodelling: results from the Markers and Response to CRT (MARC) study. *Europace*.
31. Engels EB, Végh EM, van Deursen CJM, et al. T-Wave Area Predicts Response to Cardiac Resynchronization Therapy in Patients with Left Bundle Branch Block. *J Cardiovasc Electrophysiol* 2015;26:176–183.
32. Hadwiger M, Dagres N, Haug J, et al. Survival of patients undergoing cardiac resynchronization therapy with or without defibrillator: the RESET-CRT project. *Eur Heart J* 2022;43:2591–2599.
33. Steyerberg EW, Harrell FEJ. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* 2016;69:245–247.
34. Attia IZ, Tseng AS, Benavente ED, et al. External validation of a deep learning electrocardiogram algorithm to detect ventricular dysfunction. *Int J Cardiol* 2021;329:130–135.
35. Van de Leur RR, Bleijendaal H, Taha K, et al. Electrocardiogram-based mortality prediction in patients with COVID-19 using machine learning. *Netherlands Hear J* 2022;30:312–318.
36. Plesinger F, van Stipdonk AMW, Smisek R, et al. Fully automated QRS area measurement for predicting response to cardiac resynchronization therapy. *J Electrocardiol* 2020;63:159–163.



11

General discussion and future perspectives

The electrocardiogram (ECG) was introduced by Willem Einthoven over 100 years ago and it is still a central tool in clinical medicine. Einthoven already predicted in 1912 that there was much more to gain from the technology when he said that the method of electrocardiography is “still a young plant that will continue to bear good fruit”.[1] Interpretation of the ECG is a complex task, requires extensive training and physicians at all levels of training have deficiencies in ECG interpretation.[2] Cardiologists have the highest accuracy in interpreting ECGs, but still with high inter- and intra-rater variability.[3,4] Given these difficulties, efforts to computerize ECG interpretation started around 1960, but the algorithms have not reached physician-level accuracy yet. Therefore, all computer-based reports should be systematically overread, which places a heavy logistical burden on clinical practice.[4,5]

With the discovery that deep neural networks (DNNs), a type of artificial intelligence (AI), can learn to interpret ECGs when trained on big datasets, we are now beginning to unleash the full potential of the ECG.[6] Deep neural networks are computer algorithms based on the structure and function of the human brain, and consists of many layers of neurons that can learn non-linear relationships between patterns in the raw ECG by themselves.[7] In a sense this is similar to how humans interpret the ECG, by recognizing different patterns and combining the findings.[8] Humans usually use some sort of rule-based pathway with different (binary) thresholds to determine whether a feature is present or not. To determine whether a typical left bundle branch block is present, for example, one needs a QRS duration of more than 120ms, QS or rS in V1 and broad and slurred R-waves without Q waves laterally.[9] DNNs are not limited to such pathways or (binary) thresholds and might recognize more complex, more subtle and non-linearly interrelated patterns in an ECG. Studies have shown that DNNs applied to ECGs can detect a person’s sex, the presence of left ventricular dysfunction or arrhythmias not present at the time of recording.[10–13] When utilized for such tasks, DNNs can probably combine (subtle) electrical information beyond which a human can typically comprehend.

In this thesis, we aimed to bridge the gap between technical AI research and clinical implementation in ECG-AI to bring the algorithms closer to making a real change in clinical practice. In **Chapter 1**, we described several prerequisites for successful development and implementation of ECG-AI in clinical practice: ask the right clinical questions, perform rigorous quality control, prevent overfitting, investigate possible bias using explainability techniques, have measures of uncertainty, and perform proper implementation studies. Here, we discuss this thesis by highlighting the most important four prerequisites (**Figure 1**) and four clinical opportunities (Figure 2) for ECG-AI.

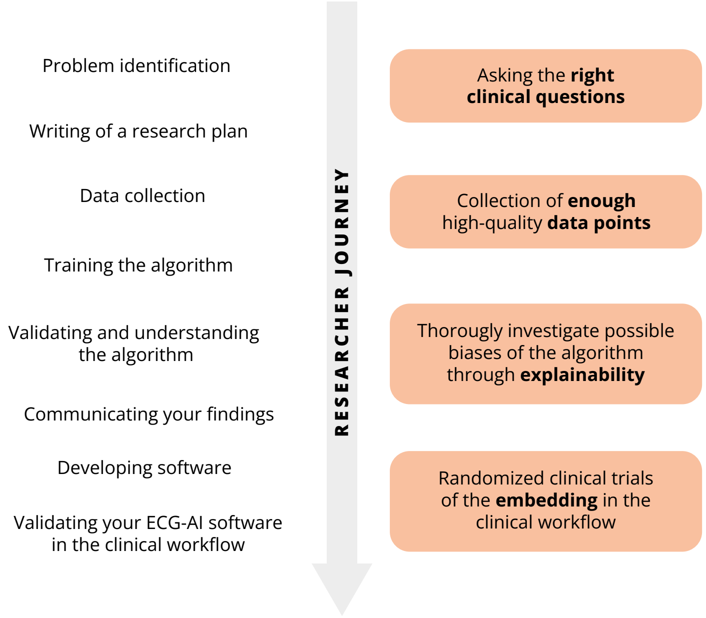

Problem identification
Writing of a research plan
Data collection
Training the algorithm
Validating and understanding the algorithm
Communicating your findings
Developing software
Validating your ECG-AI software in the clinical workflow
RESEARCHER JOURNEY
Asking the right clinical questions
Collection of enough high-quality data points
Thorougly investigate possible biases of the algorithm through explainability
Randomized clinical trials of the embedding in the clinical workflow

Figure 1. *Four of the most important prerequisites for successful implementation of artificial intelligence for the ECG as identified in this thesis.*

Prerequisite one: asking the right clinical questions

Over 1500 ECG-AI algorithms have been proposed in the last few years, but over 90% of the developed ECG-AI algorithms never made it into clinical practice. This is partly because most ECG-AI algorithms developed today are based on retrospectively collected data and research questions are often formulated based on the data available instead of the relevant medical problem. The most important prerequisite for successful implementation of any diagnostic or predictive algorithm, either based on traditional statistics or AI, is to make sure it could be solving a relevant medical problem.

To develop algorithms that could really improve patient care, it is essential to first clearly identify where the algorithm needs to be implemented in the clinical workflow.[14] Will it be an add-on to improve the current workflow or replace a current prediction model? In the latter case, it is of the utmost importance to thoroughly investigate the added benefit of the novel algorithm. When only taking studies with a low risk of bias into account, one systematic review found no evidence of superior performance of machine learning over more classical statistical methods, such as logistic regression, for example.[15]

Secondly, we need to identify by whom the algorithm will be used in the workflow. Where expert knowledge is not available, or by the experts themselves? If used by experts on the topic, the algorithm should have added benefit over their interpretations. A recent study subjected a state-of-the-art AI algorithm to the Fellowship of the Royal College of Radiologists (FRCR) examination, and found that the AI only passed 2 of 10 mock exams, while young radiologists passed 4 out of 10.[16]

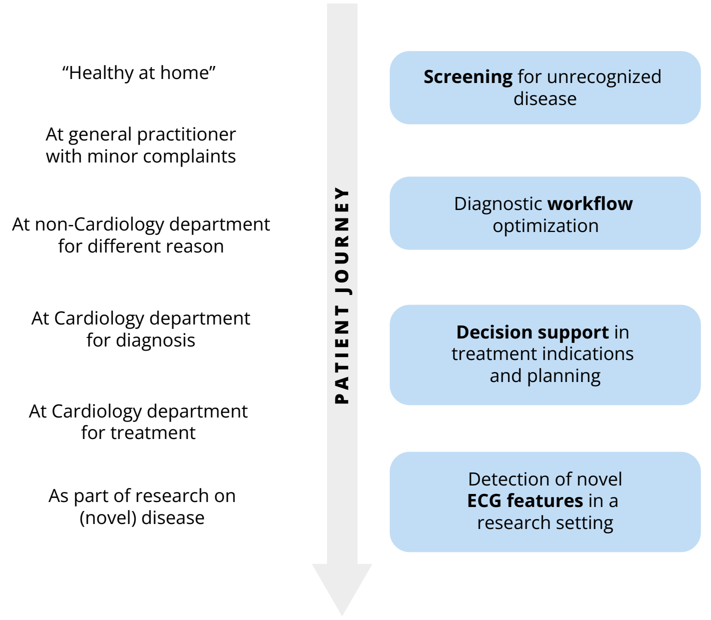

"Healthy at home"
At general practitioner with minor complaints
At non-Cardiology department for different reason
At Cardiology department for diagnosis
At Cardiology department for treatment
As part of research on (novel) disease
PATIENT JOURNEY
Screening for unrecognized disease
Diagnostic **workflow** optimization
Decision support in treatment indications and planning
Detection of novel **ECG features** in a research setting

Figure 2.
Four of the most promising applications of artificial intelligence in a typical patient journey from being asymptomatic at home, to first contact with the general practitioner or in the hospital for a non-cardiological reason and finally at the Cardiology department for possible treatment and research.

Thirdly, even when an algorithm solves a clear clinical question, has a well-defined position in the clinical workflow, and outperforms the currently used pathway, one should consider practical barriers for implementation from the start. Can the algorithm be fitted into the current software landscape? Is the data accessible? Will implementation save money? Who will pay for the algorithm? Does implementation lead to an in- or decrease of workload for the personnel that needs to apply it?[17]

In this thesis, we identified the following four most promising opportunities for the first successful implementation of ECG-AI in practice: diagnostic workflow optimization, screening for unrecognized or rare diseases, decision support in treatment indications and discovery of novel ECG features (**Figure 2**).

Opportunity one: diagnostic workflow optimization

There are several important challenges in healthcare today, including the rising costs in many developed countries, the limited access to care in many developing countries and the shortage of qualified medical personnel. In the Netherlands, more and more news outlets are reporting a 'healthcare infarction' with reduced access to care for many citizens due to an imbalance in the number of patients and number of healthcare personnel. It has been predicted that in 2040 one in every four job positions would be in healthcare to maintain the current healthcare model in the Netherlands.[18] Naturally, one of the first 'right clinical questions' for AI in ECG analysis is to optimize the diagnostic workflow and keep healthcare systems viable.

Especially in non-cardiology departments and pre-hospital care, expert knowledge to interpret ECGs might not always be readily available.[2] Given the life-threatening nature of a suspected acute coronary syndrome and ventricular arrhythmias, timely ECG interpretation places a heavy logistic burden on clinical practice for cardiologists and cardiology residents. In the UMC Utrecht, for example, around 30.000 ECGs are made yearly at non-cardiology departments, of which 50% is normal and requires no follow-up (unpublished data). Country-wide this would accumulate to 600.000 ECGs in hospitals that need to be overread or lead to a referral, of which 300.000 are normal. In prehospital care in the

Netherlands every year over 200.000 chest pain patients are referred from the general practitioner to the emergency room, of which only 21% had a final diagnosis of acute coronary syndrome.[19]

As we assumed that detailed ECG interpretation will most likely remain the task of a cardiologist, we proposed a novel algorithm in **Chapter 2** that only prioritizes which ECGs need referral to a cardiologist.[20] The DNN algorithm was trained on 300.000 ECGs annotated by physicians as part of the regular clinical workflow, and classified ECGs as normal (no referral needed), abnormal not acute (overreading within 24 hours) or abnormal (sub)acute (fast consultation) with excellent discriminatory performance (c-statistic 0.93 [95% confidence interval (CI) 0.92 – 0.95]) in an expert annotated test set. In **Chapter 3**, we prospectively validated the algorithm in a hospital setting and performed a background implementation study to see whether implementation is safe and efficacious.[21] In this study we took the full clinical pathway of every patient into account and show that no important diagnoses were missed and the number of ECGs predicted as acute that did not require follow-up was very limited. Future studies are currently being designed to show whether implementation of such an algorithm will lead to reduced workload for cardiologists and improved time-to-treatment.

One very important aspect to address for algorithms that are being used by non-experts (i.e. users that will not always know when the algorithm is wrong), is to make sure the algorithm will inform the user when it doesn't know for sure. DNNs always provide a diagnosis or prediction, even if the input is too noisy to interpret or contains an abnormality the algorithm has never seen before. In a real-world setting, clinicians acknowledge when they are uncertain and request additional tests or consult colleagues or literature. In **Chapter 4**, we sought to give DNNs a similar opportunity to express their uncertainty by training multiple versions of the same algorithm and showing when they disagree.[22] While pressure testing this new uncertainty measure in a clinical simulation, we showed that by thresholding the uncertainty estimates and thereby rejecting uncertain ECGs we could improve accuracy in the remaining data. Furthermore, we found a strong correlation between estimated uncertainty and disagreement between cardiologists.

Next to more generic algorithms, such as a triage algorithm trained on physicians ECG annotations, more specific models could be useful for specific diagnostic pathways. In the general practice, for example, ECGs are mostly acquired for the diagnosis of heart failure, next to detecting rhythm disorders and coronary artery disease. Current guidelines advice an ECG, alongside history taking, physical examination and B-type natriuretic peptide (BNP) measurements, for determining which patients need follow-up and echocardiography.[23] Unfortunately, this approach leads to high numbers of both underdiagnosis (patients with unrecognized heart failure, around 15-20% in an elderly general population with shortness of breath) and overdiagnosis (patients referred for echocardiography but without heart failure, probably between 40 and 80%).[24–28] One possible reason could be that only the subjective finding of an 'abnormal ECG' is used in most diagnostic decision rules, and that performance of general practitioners in interpreting ECGs for heart failure is suboptimal, with a mean sensitivity of 53-63% and specificity of 63-73%.[24,26,28]

In **Chapter 5**, we described an algorithm that is able to predict left ventricular systolic dysfunction (LVSD, here defined as ejection fraction <40%) with a c-statistic of 0.89 [95% CI 0.89 – 0.91], sensitivity of 89% and specificity of 70%.[13] Such an algorithm could be of great use in a decision rule to determine which patients to refer for echocardiography, but many research steps remain before implementation is possible. Other reasons that warrant referral for echocardiography include heart failure with preserved ejection fraction and valvular disorders, and although studies have shown value of ECG-AI in detecting those, a composite model that is truly able to rule-out any relevant echocardiogram abnormality should be developed.[29–31] Moreover, the value of such models in combination with other predictors, such as BNP measurements, should be investigated first.[32]

Future studies should elucidate whether broad models, such as the triage algorithm, or more specific models, such as an algorithm specifically trained to detect left ventricular systolic dysfunction, aortic stenosis or occluding myocardial infarction, have most value in optimizing and reorganizing clinical workflows. Currently, several groups around the world are investigating this, and multiple clinical trials are underway to show whether ECG-AI could really lead to a more efficient use of our healthcare resources.[33]

Opportunity two: screening for unrecognized and rare disease

In addition to optimizing the current clinical diagnostic workflow, screening for unrecognized and rare disease in possibly asymptomatic patients is a second major opportunity for ECG-AI. Other groups have shown feasibility for detecting silent atrial fibrillation, dilated and hypertrophic cardiomyopathy, valvular disease and long QT syndrome using ECG-AI.[12,30,31,34–36] In this thesis, next to the algorithm for reduced left ventricular ejection fraction, we have developed one other algorithm that could be useful for screening purposes.

In **Chapter 7**, we proposed a DNN developed on median beat ECGs of 86 phospholamban (PLN) p.Arg14del variant carriers to distinguish them from age and sex-matched controls.[37] This algorithm performed well with a c-statistic of 0.95 [95% CI 0.91-0.99] and sensitivity and specificity of 0.82 and 0.93, respectively. The PLN p.Arg14del genetic variant originates from an ancient Dutch founder and its prevalence in the Netherlands is estimated to be around 1:500-1000. It has also been identified in several other countries including Spain, Greece, Vietnam, China, Japan, Canada, and the United States.[38–40] As the typical ECG characteristics of PLN p.Arg14del variant carriers are largely unknown in many parts of the world, such a screening algorithm could be of great use to determine which patients need genetic testing. The current analysis is limited by its case-control design, however, and future studies validating the model in relevant populations with a more realistic prevalence are necessary.

While the algorithms proposed for optimizing the diagnostic workflow above should lead to reduced workload for the healthcare system, algorithms proposed for screening of unrecognized disease could lead to a dramatic increase. Especially screening for asymptomatic LVSD or silent atrial fibrillation could lead to a short-term burden on healthcare, and future studies should focus on the implications of implementation of such algorithms, including health technology assessments. One randomized controlled trial was performed where an ECG-AI algorithm for detection of LVSD was available in primary care. In this study, the proportion of patients that underwent echocardiography (19%) did not increase in the intervention group, which is a first reassuring finding.[41] In other trials, however, where screening for LVSD using Apple Watch ECGs or screening for silent AF was performed, implementation of ECG-AI would lead to referral in 50% to 65% of patients.[42,43]

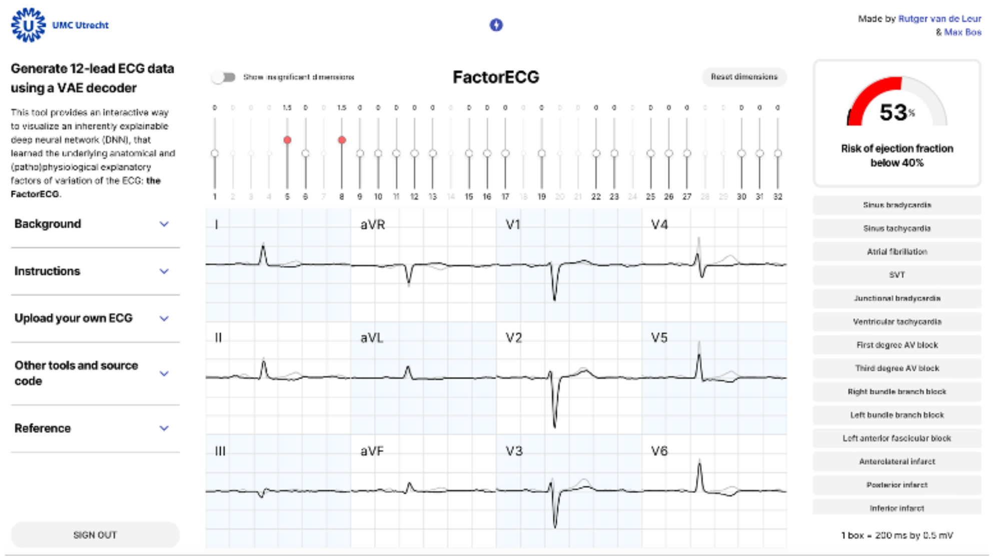
Prerequisites two and three: explainability and applicability to small data

As shown above, DNNs excel when applied to very large datasets of raw unstructured data, such as manually annotated ECGs or ECGs linked to echocardiograms. The algorithms are considered ‘black boxes’, however. They cannot provide meaningful information about the logic behind their decisions, which is warranted by the European Unions General Data Protection Regulation laws.[44] It has therefore been argued that all AI systems should be explainable to their clinical users.[45] On the other hand, if the accuracy of an algorithm is shown in robust validation and implementation studies across relevant (marginalized) subgroups, the need for explainability remains questionable for most use cases of ECG-AI.[46] The current medical system is already well adapted to ‘black boxes’ and many drugs and devices function as such. The mechanism of action of paracetamol, for example, is only partially understood, but we know it is a safe and effective pain medication due to numerous randomized controlled trials (RCTs).[47] RCTs have been the standard way to evaluate medical interventions, and this should not be different for AI algorithms.[48] Although explainable AI techniques are thus probably not a prerequisite when using ECG-AI systems in clinical practice, they are a prerequisite in the development of such algorithms for detection of bias novel ECG features, possibly improving our understanding of pathophysiology and increasing trust in ECG-AI.[6,46] In dermatology, for example, heatmaps were used to discover that an AI erroneously focused on surgical skin markings to detect skin cancer.[49]

Most research on explainable AI for the ECG uses heatmap-based techniques to detect what parts of the ECG the algorithm bases its decision on. As we argue in **Chapter 6**, however, currently used heatmap-based methods are uninformative, unreliable, and prone to confirmation bias.[50] These methods highlight a broad area on the individual ECG to show where the algorithm focuses on, but that does not give any insight into what the model is doing with that area. Moreover, you can pick many different examples and heatmap methods, with each a different explanation. This poses the risk of confirmation bias, as Cynthia Rudin explains: “You could have many explanations for what a complex model is doing. Do you just pick the one you ‘want’ to be correct?”[51] Moreover, it remains debatable if such explainability techniques will ever provide accurate explanations, as it is contradictory to first train a highly complex algorithm, and then try to explain this using a very simple heatmap.

Given these shortcomings of current explainability techniques, we sought to develop a novel approach for explainable ECG-AI in **Chapter 5**. Here, we decouple feature discovery from classification in DNNs by using a variational auto-encoder (VAE) to first decompose the ECG into its generative principal components (the FactorECG).[13,52] The VAE was trained on over 1 million ECGs to learn these 32 ECG factors by itself. Using extensive visualizations and association analyses, we show that many factors represent known ECG morphology, while others do not correlate to any human-named feature and possibly represent novel features (https://decoder.ecgx.ai, **Figure 3**). As the ECG factors are subsequently used in interpretable prediction models, such as logistic regression, we can directly link changes in ECG morphology to decisions of the algorithm. In current analysis, for example, we showed that inferolateral ST elevation is associated with LVSD, which illustrates that the algorithm might also pick up myocardial dysfunction due to acute ischemia. This could severely hamper the generalizability of the model for screening purposes in the general population and is one of the reasons why explainable models are imperative. Next to the improved explainability, the VAE also performs enormous dimensionality reduction to only 32 factors (from 7200 data points in the 12-lead median beat ECG), which broadens the usability of DNNs to much smaller labeled datasets than before.

While the FactorECG provides an important advance towards using AI for ECG feature discovery in smaller datasets, there are still several improvements warranted. Firstly, the ability of the VAE to compress the ECG into its 32 factors is not perfect, and the correlation between input and reconstructed ECG is 0.86. This means subtle or high-frequency components, such as pacing spikes or the R-wave height, and rare ECG variations, such as ventricular tachycardia, are sometimes poorly encoded and further technological advances are needed to provide more lossless reconstructions. A novel approach (the FactorECG 2.0) is currently under active development.[53] Furthermore, additional insight into the (patho)physiological mechanism behind the different ECG factors is needed by associating them with other structural parameters (such as from echocardiography) or genetics. Finally, we hope to encourage other groups to also apply the FactorECG to their ECG data and independent validate our results or expand the applicability to new datasets (an online tool to convert a batch of ECGs is available through https://encoder.ecgx.ai).



This tool provides an interactive way to visualize an inherently explainable deep neural network (DNN), that learned the underlying anatomical and (patho)physiological explanatory factors of variation of the ECG: **the FactorECG**.

Figure 3. *Example of explanation by the FactorECG for prediction of an ejection fraction below 40%.* By adjusting the different factors, we can visualize which ECG morphology is associated with the prediction. In this example, a high value in factor 5 (inferolateral negative T-waves) and in factor 8 (increased PR-interval and P-wave duration) leads to a very high probability of left ventricular systolic dysfunction. This tool is accessible via https://decoder.ecgx.ai and also allows for upload of ECGs in a research setting.

Opportunity three: decision support in treatment indications and planning

As the FactorECG can use what it learned on over a million ECGs in much smaller datasets, one new field where ECG-AI may provide benefit in current clinical practice could be in supporting decision making for treatment indications. The ECG plays or could play an important role for decision-making in many treatments in Cardiology, such as the decision to implant an implantable cardioverter defibrillator (ICD) or cardiac resynchronization therapy (CRT) device and to perform a cardiac ablation procedure.[6] It could also provide improved localization of premature ventricular complexes or the location of the accessory pathway in Wolff-Parkinson-White patients, which could help in optimizing treatment planning.[54–56] For most of these treatments, the indications or procedure planning depends on manually derived ECG parameters that are too simplistic, such as the number of negative T-waves, or remain difficult to standardize, such as the definitions of typical left bundle branch block.[57–59] In this thesis, we developed three algorithms to assist in treatment decision support by using the power of the FactorECG. This is a very timely subject, as these devices and procedures are expensive and better selection could lead to reduced healthcare expenditure.

In **Chapter 8**, we used the pretrained FactorECG model from **Chapter 5** to predict the risk of malignant ventricular arrhythmia (MVA) in PLN p.Arg14del variant carriers with an optimism-corrected c-statistic 0.79 [95% CI 0.75 – 0.85] using only 12-lead ECG data. Addition of echocardiographic and Holter monitoring data in the group with high predicted risk based on the ECG improved predictive ability further, resulting in a positive predictive value (PPV) of 18% and negative predictive value (NPV) of 99%, outperforming the use of the currently used multimodal model. We compared the model to an approach using conventional ECG parameters, such as the number of negative T-waves and the presence of low QRS voltage and show that it greatly outperforms such model. Clinically, such an ECG-only model can be used in a two-step approach involving a first pass using the ECG model alone, followed by additional diagnostics in subjects deemed at-risk of MVA to determine which patients should receive and ICD implantation. As acceptable NPVs can be achieved with only ECG at home or via the general practitioner in a large subgroup, the health care burden of PLN monitoring visits could be reduced, lowering the burden on asymptomatic carriers significantly as well.

In **Chapter 9,** we utilized the same approach to predict MVA in a broader group of patients with dilated cardiomyopathy using only ECG data.[60] In contrast to the PLN population, in DCM patients there are no clear ECG criteria for deciding on ICD implantation and only the LVEF cut-off of 35% is used in this group. Although this cut-off showed no predictive value for the risk of MVA in this population, the deep learning-based approach reached a c-statistic of 0.67 [95% CI 0.62 – 0.72] in a Cox regression model. This model performed slightly worse than the model in PLN p.Arg14del variant carriers, most likely due to the fact that it is a much more diverse group of patients with more diverse and subtle ECG abnormalities.

In **Chapter 10,** we developed and validated the FactorECG methodology again to predict echocardiographic response and clinical outcomes after CRT implantation.[61] For the first time, deep learning has been used to predict clinical outcome after CRT using only the raw preprocedural ECG (c-statistic 0.69 [95% CI 0.66-0.72]). The algorithm outperformed contemporary guideline criteria and vectorcardiographic QRS_{AREA} for stratifying which patients would benefit from CRT implantation, without compromising the total proportion of patients deemed suitable for CRT implantation. One of the major advantages is the reproducible, user-independent and consisted analysis of the 12-lead ECG morphology, in contrast to manually defining which patients have a typical left bundle branch block (which suffers from high inter-observer disagreement and lack of consensus)[62]

In the future, the FactorECG can be applied to many more disease groups and treatments. Currently, our group is developing algorithms to determine the need for ICD therapy in other disease groups, such as hypertrophic cardiomyopathy and arrhythmogenic cardiomyopathy.

Opportunity four: discovery of novel ECG features

Finally, one opportunity of explainable AI is to use DNNs as a ‘feature detector’ for diseases where the ECG features might not be known. If one wants to discover which ECG features are associated with a specific, possible novel or rare disease, manual interpretation of large numbers of ECGs using prespecified criteria is needed. This might not always be feasible, and restricts the possible features found to the prespecified ones. We assumed that DNNs could be used to learn those features in an agnostic way and might provide insight into which ECG morphology is diagnostic or predictive.

For DNNs to be used as a ‘feature detector’, some kind of explainability was needed. In **Chapters 2 and 7**, we employed heatmap-based methods, such as Guided Grad-CAM to determine which segments of the median beat ECG were of interest to the algorithm. These methods only work on individual ECGs, and in **Chapter 2** we were therefore only able to validate this methodology by looking at some examples.[20] In **Chapter 7**, we expanded on this method by providing an overall heatmap of the whole dataset, by taking the mean of all time-aligned individual heatmaps.[37] Using such analysis, we were able to show that the DNN looked at similar ECG segment to detect PLN p.Arg14del variant carriers as a cardiologist would, but was able to bring a new focus on which detailed features were the most distinctive (R and T-wave attenuation in V2 and V3 and increased PR-duration). The limitations of heatmap-based methods, as described above and in **Chapters 5 and 6,** emerged during the analysis, we used the FactorECG in **Chapters 8, 9** and **10** as the 'feature detector'.

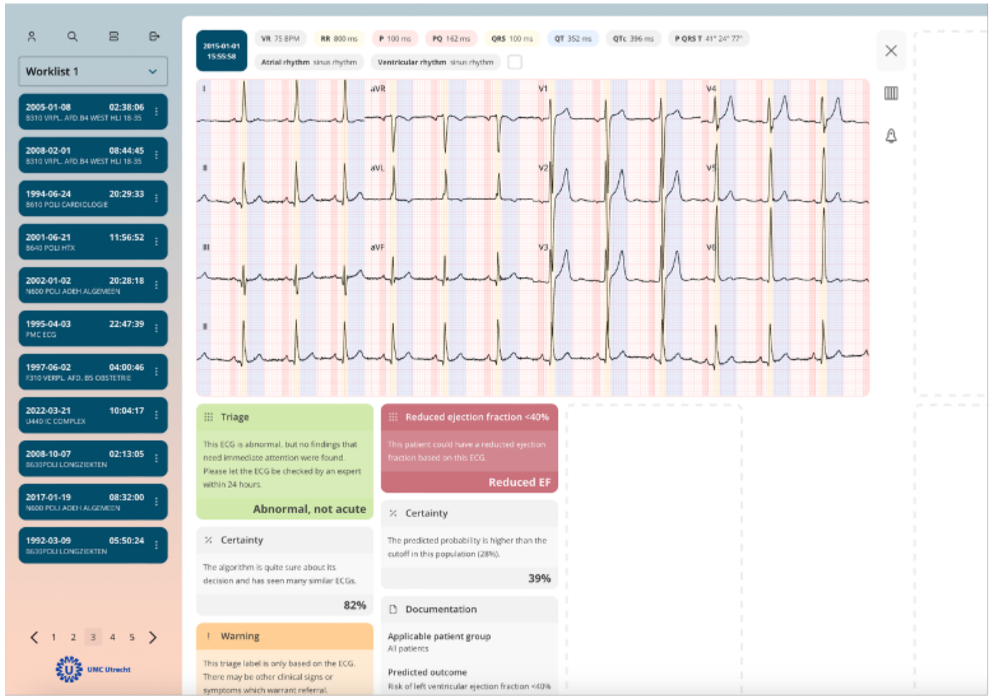
In **Chapter 8**, where the FactorECG was used to predict MVA in PLN p.Arg14del variant carriers, we found that ECG factors 1 and 5 predicted MVA and represented known ECG features (reduced QRS voltage and inferolateral symmetrical negative T-waves). Interestingly, it used these features as a continuous spectrum and already predicts a high risk before the appearance of negative T-waves, but only with a reduced R- and T-wave height. This might explain why the model outperforms manual ECG interpretation, as this uses binary cut-off points for QRS voltage and negative T-waves. In **Chapter 9**, we predict MVA in the broader group of DCM patients, where we did the surprising finding that factors associated to the P-wave are most predictive (factor 8 and 27). While factor 8 is also predictive in PLN p.Arg14del variant carriers to a lesser extent, it is probably a better marker in a heterogeneous group of DCM patients.[63] In **Chapter 10**, we used the FactorECG to predict response to CRT and found that many factors outside of the QRS complex are associated. We are currently investigating what these factors represent physiologically. Future studies could apply the FactorECG approach to detect novel ECG features in other diseases groups, such as other cardiomyopathies and idiopathic ventricular fibrillation.

Prerequisite four: real validation using randomized trials of the embedding in the clinical workflow

As described, research from our and other groups has shown that AI applied to the ECG could be very useful in clinical practice, when asking the right questions to the algorithm. Currently, the performance of ECG-AI is mostly supported by retrospective or preliminary implementation studies.[6,33,48] We are at the forefront for real-world testing of the clinical benefits of ECG-AI, and the coming years should focus on large-scale randomized trials and other implementation studies. Such studies should especially focus on performance in subgroups and other ECG devices. Before this is possible, however, we need a software environment that seamlessly embeds ECG-AI in the workflow of the clinician, from general practitioner to electrophysiologist. Such a platform should be developed together with AI specialists, hybrid physicians (i.e. with technical and medical background), other clinicians and the IT department of the hospital. Ideally, it would be vendor-neutral (i.e. working with every ECG data source, from single-lead Apple Watch ECGs to 12-lead 10 second and 30 day ECGs) and modular (i.e. easily expandable with new algorithms). Its interface should provide results in a safe and intuitive way for the clinician who has no experience with AI. This could, for example, be performed by providing a 'model facts' label.[64] An example of such a platform, as currently being developed in the UMC Utrecht, is shown in **Figure 4**.

Conclusion

In conclusion, during this project much progress is made in ticking the boxes of prerequisites for the successful implementation of ECG-AI in clinical practice. We show promising results of ECG-AI in screening for unrecognized disease, diagnostic workflow optimization, decision support in treatment planning and detection of novel ECG features. Major challenges remain improvements to the FactorECG algorithm, the need for large-scale randomized controlled trials and implementation studies and the development of software platforms to integrate algorithms in the clinical workflow. Importantly, future studies should elucidate whether implementation of ECG-AI leads to improved patient outcomes and a reduced financial and logistical burden on healthcare.



This ECG is abnormal, but no findings that need immediate attention were found. Please let the ECG be checked by an expert within 24 hours.

Abnormal, not acute

This patient could have a reduced ejection fraction based on this ECG.

Reduced EF

The algorithm is quite sure about its decision and has seen many similar ECGs.

82%

The predicted probability is higher than the cutoff in this population (28%).

39%

Applicable patient group: All patients

Predicted outcome: Risk of left ventricular ejection fraction <40%

This triage label is only based on the ECG. There may be other clinical signs or symptoms which warrant referral.

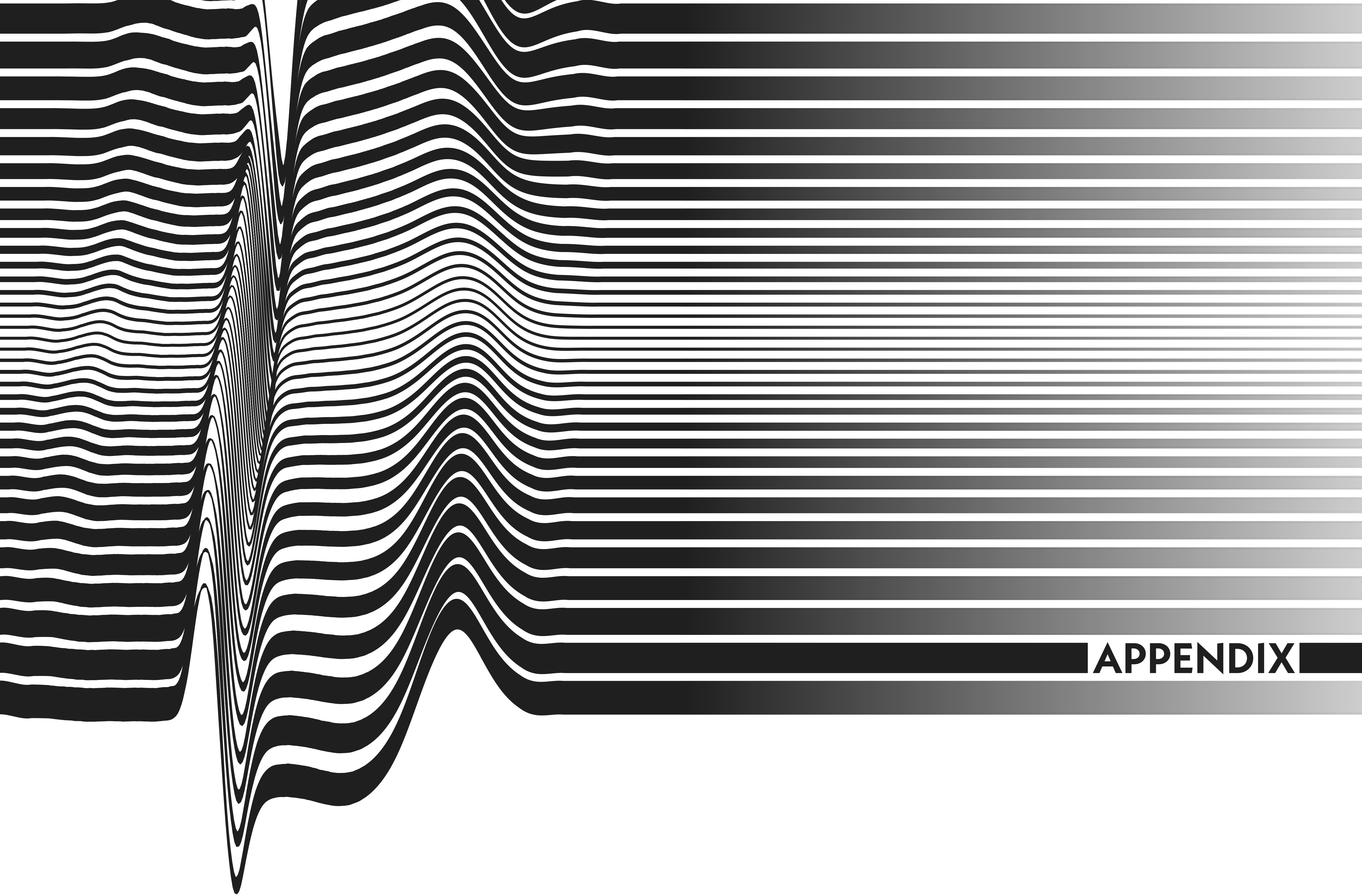
Figure 4. *Dashboard and platform for implementation of artificial intelligence algorithms of 12-lead 10 second electrocardiograms in the clinical hospital workflow that is currently under development in the UMC Utrecht.* This is a screenshot from a demo patient, where the triage algorithm predicts an abnormal ECG, but without findings that warrant immediate attention. The reduced ejection fraction (EF) algorithm, however, detects a high chance for this patient having reduced EF.

REFERENCES

1. Einthoven W. The Different Forms of the Human Electrocardiogram and their signification. Lancet 1912; 179: 853–861.
2. Cook DA, Oh S-Y, Pusic MV. Accuracy of Physicians' Electrocardiogram Interpretations. Jama Intern Med 2020; 180: 1461.
3. Holmvang L, Hasbak P, Clemmensen P, et al. Differences between local investigator and core laboratory interpretation of the admission electrocardiogram in patients with unstable angina pectoris or non-Q-wave myocardial infarction (a thrombin inhibition in myocardial ischemia [TRIM] substudy). Am J Cardiol 1998; 82: 54--60.
4. Salerno SM, Alguire PC, Waxman HS. Competency in Interpretation of 12-Lead Electrocardiograms: A Summary and Appraisal of Published Evidence. Ann Intern Med 2003; 138: 751.
5. 5. Schläpfer J, Wellens HJ. Computer-Interpreted Electrocardiograms Benefits and Limitations. J Am Coll Cardiol 2017; 70: 1183–1192.
6. 6. Leur RR van de, Boonstra MJ, Bagheri A, et al. Big Data and Artificial Intelligence: Opportunities and Threats in Electrophysiology. Arrhythmia Electrophysiol Rev 2020; 9: 146–154.
7. Goodfellow I, Bengio Y, Courville A. Deep Learning. The MIT Press, 2016.
8. Wood G, Batt J, Appelboam A, et al. Exploring the Impact of Expertise, Clinical History, and Visual Search on Electrocardiogram Interpretation. Med Decis Making 2013; 34: 75–83.
9. Members AF, Brignole M, Auricchio A, et al. 2013 ESC Guidelines on cardiac pacing and cardiac resynchronization therapyThe Task Force on cardiac pacing and resynchronization therapy of the European Society of Cardiology (ESC). Developed in collaboration with the European Heart Rhythm Association (EHRA). Ep Europace 2013; 15: 1070–1118.
10. Siegersma KR, Leur RR van de, Onland-Moret NC, et al. Deep neural networks reveal novel sex-specific electrocardiographic features relevant for mortality risk. European Hear J - Digital Heal. Epub ahead of print 2022. DOI: 10.1093/ehjdh/ztac010.
11. Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nat Med 2019; 25: 70--74.
12. Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. 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; 6736: 1--7.
13. Leur RR van de, Bos MN, Taha K, et al. Improving explainability of deep neural network-based electrocardiogram interpretation using variational auto-encoders. European Hear J - Digital Heal. Epub ahead of print 2022. DOI: 10.1093/ehjdh/ztac038.
14. Smeden M van, Heinze G, Calster BV, et al. Critical appraisal of artificial intelligence-based prediction models for cardiovascular disease. Eur Heart J 2022; 43: 2921–2930.
15. Christodoulou E, Ma J, Collins GS, et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019; 110: 12–22.
16. Shelmerdine SC, Martin H, Shirodkar K, et al. Can artificial intelligence pass the Fellowship of the Royal College of Radiologists examination? Multi-reader diagnostic accuracy study. Bmj 2022; 379: e072826.
17. Watson J, Hutyra CA, Clancy SM, et al. Overcoming barriers to the adoption and implementation of predictive modeling and machine learning in clinical care: what can we learn from US academic medical centers? Jamia Open 2020; 3: 167–172.
18. Sociaal Economische Raad. Zorg voor de toekomst: over de toekomstbestendigheid van de zorg, https://www.ser.nl/-/media/ser/downloads/adviezen/2020/zorg-voor-de-toekomst.pdf (June 1, 2020).
19. Hoorweg BBN, Willemsen RTA, Cleef LE, et al. Frequency of chest pain in primary care, diagnostic tests performed and final diagnoses. Heart 2017; 103: 1727--1732.
20. Leur RR van de, Blom LJ, Gavves E, et al. Automatic Triage of 12-Lead Electrocardiograms Using Deep Convolutional Neural Networks. J Am Heart Assoc; 9. Epub ahead of print 2020. DOI: 10.1161/jaha.119.015138.
21. Leur RR van de, Sleuwen MTMG van, Zwetsloot PP, et al. Automatic Triage of 12-lead ECGs using Deep Convolutional Neural Networks: A First Implementation Study. submitted.
22. Vranken JF, Leur RR van de, Gupta DK, et al. Uncertainty estimation for deep learning-based automated analysis of 12-lead electrocardiograms. European Hear J - Digital Heal. Epub ahead of print 2021. DOI: 10.1093/ehjdh/ztab045.
23. McDonagh TA, Metra M, Adamo M, et al. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failureDeveloped by the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) With the special contribution of the Heart Failure Association (HFA) of the ESC. Eur Heart J 2021; 42: ehab368.
24. Mant J, Doust J, Roalfe A, et al. Systematic review and individual patient data meta-analysis of diagnosis of heart failure, with modelling of implications of different diagnostic strategies in primary care. Health Technol Asses 2009; 13: 1–232.
25. Rutten FH, Cramer M-JM, Grobbee DE, et al. Unrecognized heart failure in elderly patients with stable chronic obstructive pulmonary disease. Eur Heart J 2005; 26: 1887–1894.
26. 26. Rutten FH, Moons KGM, Cramer M-JM, et al. Recognising heart failure in elderly patients with stable chronic obstructive pulmonary disease in primary care: cross sectional diagnostic study. Bmj 2005; 331: 1379.
27. Valk MJ, Mosterd A, Broekhuizen BD, et al. Overdiagnosis of heart failure in primary care: a cross-sectional study. Br J Gen Pract 2016; 66: e587–e592.
28. Lim TK, Collinson PO, Celik E, et al. Value of primary care electrocardiography for the prediction of left ventricular systolic dysfunction in patients with suspected heart failure. Int J Cardiol 2007; 115: 73–74.
29. Kwon J, Kim K-H, Eisen HJ, et al. Artificial intelligence assessment for early detection of heart failure with preserved ejection fraction based on electrocardiographic features. European Hear J - Digital Heal. Epub ahead of print 2020. DOI: 10.1093/ehjdh/ztaa015.
30. Elias P, Poterucha TJ, Rajaram V, et al. Deep Learning Electrocardiographic Analysis for Detection of Left-Sided Valvular Heart Disease. J Am Coll Cardiol 2022; 80: 613–626.
31. Ulloa-Cerna AE, Jing L, Pfeifer JM, et al. rECHOmmend: An ECG-Based Machine Learning Approach for Identifying Patients at Increased Risk of Undiagnosed Structural Heart Disease Detectable by Echocardiography. Circulation 2022; 146: 36–47.
32. Adedinsewo D, Carter RE, Attia Z, et al. Artificial Intelligence-Enabled ECG Algorithm to Identify Patients with Left Ventricular Systolic Dysfunction Presenting to the Emergency Department with Dyspnea. Circulation Arrhythmia Electrophysiol. Epub ahead of print 2020. DOI: 10.1161/circep.120.008437.
33. Siontis KC, Noseworthy PA, Attia ZI, et al. Artificial intelligence-enhanced electrocardiog-

raphy in cardiovascular disease management. Nature Reviews Cardiology. Epub ahead of print 2021. DOI: 10.1038/s41569-020-00503-2.

34. Shrivastava S, Cohen-Shelly M, Attia ZI, et al. Artificial Intelligence-Enabled Electrocardiography to Screen Patients with Dilated Cardiomyopathy. Am J Cardiol 2021; 155: 121–127.
35. Ko W-Y, Siontis KC, Attia ZI, et al. Detection of Hypertrophic Cardiomyopathy Using a Convolutional Neural Network-Enabled Electrocardiogram. J Am Coll Cardiol 2020; 75: 722--733.
36. Bos JM, Attia ZI, Albert DE, et al. Use of Artificial Intelligence and Deep Neural Networks in Evaluation of Patients With Electrocardiographically Concealed Long QT Syndrome From the Surface 12-Lead Electrocardiogram. Jama Cardiol 2021; 6: 532–538.
37. Leur RR van de, Taha K, Bos MN, et al. Discovering and Visualizing Disease-Specific Electrocardiogram Features Using Deep Learning: Proof-of-Concept in Phospholamban Gene Mutation Carriers. Circulation Arrhythmia Electrophysiol; 14. Epub ahead of print 2021. DOI: 10.1161/circep.120.009056.
38. Cheung CC, Healey JS, Hamilton R, et al. Phospholamban cardiomyopathy: a Canadian perspective on a unique population. Neth Heart J 2019; 27: 208--213.
39. Jiang X, Xu Y, Sun J, et al. The phenotypic characteristic observed by cardiac magnetic resonance in a PLN-R14del family. Sci Rep-uk 2020; 10: 16478.
40. Hof IE, Heijden JF van der, Kranias EG, et al. Prevalence and cardiac phenotype of patients with a phospholamban mutation. Neth Heart J 2019; 27: 64--69.
41. Yao X, Rushlow DR, Inselman JW, et al. Artificial intelligence–enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. Nat Med 2021; 1–5.
42. Noseworthy PA, Attia ZI, Behnken EM, et al. Artificial intelligence-guided screening for atrial fibrillation using electrocardiogram during sinus rhythm: a prospective non-randomised interventional trial. Lancet 2022; 400: 1206–1212.
43. Attia ZI, Harmon DM, Dugan J, et al. Prospective evaluation of smartwatch-enabled detection of left ventricular dysfunction. Nat Med 2022; 1–7.
44. Goodman B, Flaxman S. European Union Regulations on Algorithmic Decision-Making and a “Right to Explanation.” Ai Mag 2017; 38: 50–57.
45. Kundu S. AI in medicine must be explainable. Nat Med 2021; 1–1.
46. 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.
47. Kirkpatrick P. New clues in the acetaminophen mystery. Nat Rev Drug Discov 2005; 4: 883–883.
48. Siontis GCM, Sweda R, Noseworthy PA, et al. Development and validation pathways of artificial intelligence tools evaluated in randomised clinical trials. Bmj Heal Care Informatics 2021; 28: e100466.
49. Winkler JK, Fink C, Toberer F, et al. 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.
50. Leur RR van de, Hassink RJ, Es R van. Variational auto-encoders improve explainability over currently employed heatmap methods for deep learning-based interpretation of the electrocardiogram. European Hear J - Digital Heal. Epub ahead of print 2022. DOI: 10.1093/ehjdh/ztac063.
51. 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.
52. Attia ZI, Friedman PA. Explainable AI for ECG-based prediction of cardiac resynchronization therapy outcomes: learning from machine learning? Eur Heart J. Epub ahead of print 2022. DOI: 10.1093/eurheartj/ehac733.
53. Esser P, Rombach R, Ommer B. A Disentangling Invertible Interpretation Network for Explaining Latent Representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
54. Nishimori M, Kiuchi K, Nishimura K, et al. Accessory pathway analysis using a multimodal deep learning model. Sci Rep-uk 2021; 11: 8045.
55. Kabra R, Israni S, Vijay B, et al. Emerging role of artificial intelligence in cardiac electrophysiology. Cardiovasc Digital Heal J; 3: 263–275.
56. Zheng J, Fu G, Abudayyeh I, et al. A High-Precision Machine Learning Algorithm to Classify Left and Right Outflow Tract Ventricular Tachycardia. Front Physiol 2021; 12: 641066.
57. Calle S, Timmermans F, Pooter JD. Defining left bundle branch block according to the new 2021 European Society of Cardiology criteria. Neth Heart J 2022; 30: 495–498.
58. Stipdonk AMW van, Horst I ter, Kloosterman M, et al. QRS Area Is a Strong Determinant of Outcome in Cardiac Resynchronization Therapy. Circulation Arrhythmia Electrophysiol 2018; 11: e006497.
59. Verstraelen TE, Lint FHM van, Bosman LP, et al. Prediction of ventricular arrhythmia in phospholamban p.Arg14del mutation carriers–reaching the frontiers of individual risk prediction. Eur Heart J 2021; 42: 2842–2850.
60. Sammani A, Leur RR van de, Henkens MTHM, et al. Life-threatening ventricular arrhythmia prediction in patients with dilated cardiomyopathy using explainable electrocardiogram-based deep neural networks. Ep Europace. Epub ahead of print 2022. DOI: 10.1093/europace/euac054.
61. Wouters PC, Leur RR van de, Vessies MB, et al. ECG-based deep learning improves outcome prediction following cardiac resynchronization therapy. European Heart Journal 2022; in press.
62. Stipdonk AMW van, Vanbelle S, Horst IAH ter, et al. Large variability in clinical judgement and definitions of left bundle branch block to identify candidates for cardiac resynchronisation therapy. Int J Cardiol 2019; 286: 61–65.
63. Henkens MTHM, Martínez HL, Weerts J, et al. Interatrial Block Predicts Life- Threatening Arrhythmias in Dilated Cardiomyopathy. J Am Heart Assoc 2022; 11: e025473.
64. Sendak MP, Gao M, Brajer N, et al. Presenting machine learning model information to clinical end users with model facts labels. Npj Digital Medicine 2020; 3: 41.



APPENDIX

English summary

The 12-lead electrocardiogram (ECG) is a widely used diagnostic tool in clinical practice, playing a pivotal role in identifying a range of cardiac abnormalities. However, accurately interpreting the ECG remains a complex task with significant inter- and intra-rater variability, and access to expert knowledge is not universally available. These challenges have instigated the introduction of algorithms for the computerized interpretation of the ECG (CIE). Current versions have, however, not been able to reach physician level accuracy. A substantial improvement of CIE is forthcoming with the discovery that one specific type of artificial intelligence (AI) algorithm, called deep neural network (DNN), might be highly effective in the processing of raw data without the need for hand-crafted or rule-based feature engineering. The abundance of labeled ECG datasets makes the ECG an ideal substrate for developing deep learning-based AI algorithms.

In this thesis, we aimed to bridge the gap between technical AI research and clinical implementation and bring these algorithms closer to effecting real change in clinical practice. **Chapter 1** outlines four prerequisites for successful development and implementation of ECG-AI: asking the right clinical questions, ensuring rigorous quality control, addressing potential bias using explainability techniques, and conducting proper implementation studies. Furthermore, we identified four opportunities for ECG-AI applications: screening for unrecognized diseases, optimizing diagnostic workflows, providing decision support in treatment planning, and detecting novel ECG features.

As a first step in exploring the value of DNNs for ECG interpretation, we developed a deep learning algorithm in **Chapter 2** to optimize the diagnostic ECG workflow. This algorithm accurately triages ECGs from normal to acute, determining which cases require expert consultation and within what timeframe. Subsequently, in our pursuit of clinically applicable AI for the ECG, we conducted an implementation study with the triage algorithm in a hospital setting, demonstrating its safety and efficacy in terms of clinical outcomes (**Chapter 3**).

To address one of the prerequisites for successful implementation of ECG-AI, we investigated whether these algorithms could express uncertainty in **Chapter 4**. Addressing this aspect is crucial, as DNNs always provide a diagnosis or prediction, even when the input data is noisy or contains abnormalities the algorithm has not encountered before. We proposed a method to quantify the algorithm's certainty in its predictions, serving as a safeguard to determine which ECGs should undergo automatic analysis and which should be referred to an expert.

In addition to addressing the estimation of uncertainty, we encountered two other challenges while using deep neural networks (DNN) – the lack of explainability and the requirement for extensive datasets. To overcome these hurdles, we developed a novel method harnessing the power of DNN to interpret ECGs in an explainable manner. By training a variational auto-encoder on 1.1 million median beat ECGs, we successfully decomposed the ECG morphology into 32 explainable factors, which we refer to as the FactorECG (**Chapter 5**). Our findings demonstrated that this explainable approach not only performs as effectively as the conventional 'black box' DNNs for ECG interpretation but also shows promise in novel applications, such as detecting reduced ejection fraction. Furthermore, in **Chapter 6**, we discussed why the FactorECG method offers improved explainability compared to the heatmap-based techniques previously used.

As datasets with thousands of ECGs are not available for many clinically relevant questions, we evaluated the feasibility to transfer the knowledge that DNNs learned on *big data* to *small data*. In **Chapter 7**, we developed a deep learning algorithm to detect phospholamban (*PLN*) p.Arg14del variant carriers and used conventional heatmaps to visualize which features were used by the algorithm. Afterwards, we applied our novel explainable method, the FactorECG, to predict which *PLN* p.Arg14del variant carriers develop malignant ventricular arrhythmia (**Chapter 8**). We also applied the method in patients with dilated cardiomyopathy to predict ventricular arrhythmias (**Chapter 9**) and in patients that received cardiac resynchronization therapy to predict response to treatment and mortality (**Chapter 10**). Remarkably, both algorithms outperformed currently used prediction models while requiring only the 12-lead ECG as input.

Finally, in **Chapter 11**, we reflect on the progress made regarding the prerequisites and opportunities of ECG-AI. While we have achieved promising results for all identified opportunities and unveiled numerous future possibilities, significant challenges persist before the successful implementation of ECG-AI in clinical practice can become a reality. These challenges encompass refining the FactorECG algorithm, conducting large-scale randomized controlled trials and implementation studies, and developing software platforms to seamlessly integrate algorithms into the clinical workflow. Importantly, future studies must investigate whether the implementation of ECG-AI leads to improved patient outcomes and the necessary reduction in financial and logistical burdens on healthcare systems.

Nederlandse samenvatting

Het 12-afleidingen elektrocardiogram (ECG) is een veelgebruikt diagnostisch instrument in de klinische praktijk en speelt een cruciale rol bij het identificeren van diverse cardiale afwijkingen. Het nauwkeurig interpreteren van het ECG blijft een complexe taak met aanzienlijke variabiliteit tussen verschillende waarnemers. Bovendien zijn experts om het ECG te interpreteren niet overal beschikbaar. Deze uitdagingen hebben geleid tot de introductie van algoritmen voor de geautomatiseerde interpretatie van het ECG. Huidige versies hebben echter nog niet het niveau van nauwkeurigheid bereikt dat vergelijkbaar is met dat van een cardioloog. Een aanzienlijke verbetering van deze algoritmen is echter te verwachten door de ontdekking dat een specifiek type artificieel intelligentie (AI) algoritme, genaamd diep neuraal netwerk (DNN), zeer effectief kan zijn in het verwerken van ruwe gegevens zonder de noodzaak van handmatig gecreëerde of *rule-based* signaaleigenschappen. De overvloed aan gelabelde ECG-datasets maakt het ECG een ideaal substraat voor de ontwikkeling van zulke *deep learning* gebaseerde AI.

In dit proefschrift hebben we als doel gesteld om de kloof te overbruggen tussen technisch AI onderzoek en klinische implementatie van ECG-AI algoritmen om deze algoritmen dichter bij een daadwerkelijke verandering in de klinische praktijk te brengen. **Hoofdstuk 1** schetst vier voorwaarden voor een succesvolle implementatie van ECG-AI: het vaststellen van de relevante klinische vraagstukken, het waarborgen van strenge kwaliteitscontrole, het aanpakken van mogelijke vooringenomenheid met behulp van *explainability* en het uitvoeren van degelijke implementatiestudies. Bovendien hebben we de vier meest kansrijke toepassingen voor ECG-AI geïdentificeerd: screening van nog niet vastgestelde ziekten, optimalisatie van diagnostische *workflows*, het bieden van ondersteuning bij het kiezen van de juiste behandeling en het detecteren van nieuwe ECG-kenmerken.

Als eerste stap in het verkennen van de waarde van DNN's voor ECG-interpretatie hebben we in **Hoofdstuk 2** een *deep learning* algoritme ontwikkeld om de diagnostische workflow te optimaliseren. Dit triage algoritme kan ECG's nauwkeurig classificeren van normaal tot acuut, waarbij wordt bepaald welke gevallen een consult van de cardioloog vereisen en binnen welk tijdsbestek. Vervolgens hebben we in ons streven naar klinisch toepasbare AI voor het ECG een implementatiestudie uitgevoerd met het triage-algoritme in een ziekenhuisomgeving, waarbij de veiligheid en werkzaamheid ervan werden aangetoond in termen van klinische uitkomsten (**Hoofdstuk 3**).

Een tweede voorwaarde voor succesvolle implementatie van ECG-AI, strenge kwaliteitscontrole, werd in **Hoofdstuk 4** onderzocht. Daar bestuderen we of ECG-AI-algoritmen onzekerheid kunnen uitdrukken. Dit is een cruciaal aspect, omdat DNN's altijd een diagnose of voorspelling geven, zelfs wanneer het ECG van slechte kwaliteit is of afwijkingen bevat die het algoritme niet eerder heeft gezien. We hebben een methode ontwikkeld om de mate van zekerheid van het algoritme in zijn voorspellingen te kwantificeren, als een drempelwaarde om te bepalen welke ECG's automatisch mogen worden geanalyseerd en welke altijd aan een cardioloog moeten worden voorgelegd.

Naast het schatten van onzekerheid, kwamen we twee andere uitdagingen naar voren: het gebrek aan *explainability* en de vereiste van zeer grote datasets. Om deze obstakels te overwinnen, hebben we een nieuwe methode ontwikkeld om de kracht van DNN's te benutten bij het interpreteren van ECG's op een *explainable* manier. Door een *variational auto-encoder* te trainen op 1,1 miljoen *median beat* ECG's, waren we in staat om de ECG-morfologie te ontleden in 32 factoren, die we het FactorECG noemen (**Hoofdstuk 5**). Onze bevindingen toonden aan dat deze verklaarbare aanpak niet alleen even effectief presteert als de conventionele 'black box' DNN's voor ECG-interpretatie, maar ook veelbelovend is voor nieuwe toepassingen, zoals het detecteren van een verminderde ejectiefractie op basis van enkel het ECG. Bovendien hebben we in **Hoofdstuk 6** besproken waarom de FactorECG-methode verbeterde *explainability* biedt in vergelijking met de eerder gebruikte *heatmap*-technieken.

Aangezien datasets met duizenden ECG's niet beschikbaar zijn voor veel klinisch relevante vragen, hebben we in **Hoofdstuk 7** geëvalueerd of we de kennis die DNN's hebben opgedaan uit zeer grote datasets, konden overdragen naar kleine datasets. We hebben een DNN ontwikkeld om dragers van de phospholamban (PLN) p.Arg14del-variant op te sporen en hebben conventionele *heatmaps* gebruikt om te visualiseren welke kenmerken door het algoritme werden gebruikt. Vervolgens hebben we onze nieuwe *explainable* methode, het FactorECG, toegepast om te voorspellen welke dragers van de PLN p.Arg14del-variant een kwaadaardige ventriculaire aritmie gaan ontwikkelen (**Hoofdstuk 8**). We hebben de methode ook toegepast op patiënten met gedilateerde cardiomyopathie om ventriculaire aritmieën te voorspellen (**Hoofdstuk 9**) en op patiënten die cardiale resynchronisatietherapie kregen om de respons op behandeling en sterfte te voorspellen (**Hoofdstuk 10**). Opmerkelijk genoeg presteerden beide algoritmen beter dan momenteel gebruikte voorspellingsmodellen, terwijl ze alleen het 12-afleidingen ECG als invoer nodig hadden.

Tot slot reflecteren we in **Hoofdstuk 11** op de vooruitgang die is geboekt met betrekking tot de mogelijkheden en voorwaarden van ECG-AI. Hoewel we veelbelovende resultaten hebben behaald voor alle geïdentificeerde toepassingen en talrijke toekomstige mogelijkheden hebben beschreven, blijven er aanzienlijke uitdagingen bestaan voordat de succesvolle implementatie van ECG-AI in de klinische praktijk werkelijkheid kan worden. Deze uitdagingen omvatten het verfijnen van het FactorECG-algoritme, het uitvoeren van grootschalige *randomized controlled trials* en implementatiestudies en het ontwikkelen van softwareplatforms om algoritmen naadloos te integreren in de klinische workflow. Belangrijk is dat toekomstige studies onderzoeken of de implementatie van ECG-AI leidt tot betere uitkomsten voor patiënten en tot de benodigde vermindering van de financiële en logistieke last voor de gezondheidszorg.

List of publications

Broek HT van den, Wenker S, **Leur RR van de**, Doevendans PA, Chamuleau SAJ, Slochteren FJ van, Es R van. 3D Myocardial Scar Prediction Model Derived from Multimodality Analysis of Electromechanical Mapping and Magnetic Resonance Imaging. *J Cardiovasc Transl* 2019;12:517--527.

Leur RR van de, Blom LJ, Gavves E, Hof IE, Heijden JF van der, 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.

Leur RR van de, Boonstra MJ, Bagheri A, Roudijk RW, Sammani A, Taha K, Doevendans PA, Harst P van der, Dam P van, Hassink R, Es R van, Asselbergs FW. Big Data and Artificial Intelligence: Opportunities and Threats in Electrophysiology. *Arrhythmia Electrophysiol Rev* 2020;9:146–154.

Bos MN, **Leur RR van de**, Vranken JF, Gupta DK, Harst P van der, Doevendans PA, Es R van. Automated Comprehensive Interpretation of 12-lead Electrocardiograms Using Pre-trained Exponentially Dilated Causal Convolutional Neural Networks. *2020 Comput Cardiol* 2020;00:1–4.

Vranken JF, **Leur RR van de**, Gupta DK, Orozco LEJ, Hassink RJ, Harst P van der, Doevendans PA, Gulshad S, Es R van. Uncertainty estimation for deep learning-based automated analysis of 12-lead electrocardiograms. *European Hear J - Digital Heal* 2021.

Leur RR van de, Taha K, Bos MN, Heijden JF van der, Gupta D, Cramer MJ, Hassink RJ, Harst P van der, Doevendans PA, Asselbergs FW, Es R van. Discovering and Visualizing Disease-Specific Electrocardiogram Features Using Deep Learning: Proof-of-Concept in Phospholamban Gene Mutation Carriers. *Circulation Arrhythmia Electrophysiol* 2021;14.

Haghighi K, Gardner G, Vafiadaki E, Kumar M, Green LC, Ma J, Crocker JS, Koch S, Arvanitis DA, Bidwell P, Rubinstein J, **Leur RR van de**, Doevendans PA, Akar FG, Tranter M, Wang H-S, Sadayappan S, DeMazumder D, Sanoudou D, Hajjar RJ, Stillitano F, Kranias EG. Impaired Right Ventricular

Henkens MTHM, Martínez HL, Weerts J, Sammani A, Raafs AG, Verdonschot JAJ, **Leur RR van de**, Sikking MA, Stroeks S, Empel VPM van, Rocca HB, Stipdonk AMW van, Farmakis D, Hazebroek MR, Vernooy K, Bayés-de-Luna A, Asselbergs FW, Bayés-Genís A, Heymans SRB. Interatrial Block Predicts Life-Threatening Arrhythmias in Dilated Cardiomyopathy. *J Am Heart Assoc* 2022;11:e025473.

Sammani A, **Leur RR van de**, Henkens MTHM, Meine M, Loh P, Hassink RJ, Oberski DL, Heymans SRB, Doevendans PA, Asselbergs FW, Riele ASJM te, Es R van. Life-threatening ventricular arrhythmia prediction in patients with dilated cardiomyopathy using explainable electrocardiogram-based deep neural networks. *Ep Europace* 2022.

Zepeda-Echavarria A, **Leur RR van de**, Sleuwen M van, Hassink RJ, Wildbergh TX, Doevendans PA, Jaspers J, Es R van. Electrocardiogram Devices for Home Use: Technological and Clinical Scoping Review. *JMIR Cardio* 2023;7:e44003.

Vries NM de, Zepeda-Echavarria A, **Leur RR van de**, Loen V, Vos MA, Boonstra MJ, Wildbergh TX, Jaspers JEN, Zee R van der, Slump CH, Doevendans PA, Es R van. Detection of Ischemic ST-Segment Changes Using a Novel Handheld ECG Device in a Porcine Model. *JACC: Adv* 2023;2:100410.

Taha K, **Leur RR van de**, Vessies M, Mast TP, Cramer MJ, Cauwenberghs N, Verstraelen TE, Brouwer R de, Doevendans PA, Berg MP van den, D'hooge J, Kuznetsova T, Teske AJ, Es R van. Deep Neural Network-based Clustering of Deformation Curves Reveals Novel Disease Features in Phospholamban Mutation Carriers. *Int J Cardiovasc Img* 2023.

Leur RR van de, Sleuwen MTGM van, Zwetsloot PPM, Harst P van der, Doevendans PA, Hassink RJ, Es R van. Automatic Triage of 12-lead ECGs using Deep Convolutional Neural Networks: A First Implementation Study. *submitted* 2023.

Leur RR van de, Brouwer R de, Bleijendaal H, Verstraelen TE, Mahmoud B, Perez-Matos A, Dickhoff C, Schoonderwoerd BA, Germans T, Houweling A, Zwaag PA van der, Cox MG, Tintelen JP van, Riele AS te, Berg MP van den, Wilde AA, Doevendans PA, Es RA de B and R van, Es R van. ECG-only Explainable Deep Learning Algorithm Predicts Risk of Malignant Ventricular Arrhythmia in Phospholamban Cardiomyopathy. *submitted* 2023.

Calcium Cycling Is an Early Risk Factor in R14del-Phospholamban Arrhythmias. *J Personalized Medicine* 2021;11:502.

Jessen H, **Leur RR van de**, Doevendans P, Es R van. Automated Diagnosis of Reduced-Lead Electrocardiograms Using a Shared Classifier. *2021 Comput Cardiol (CinC)* 2021;48:1–4.

Vessies MB, Vadgama SP, **Leur RR van de**, Doevendans PA, Hassink RJ, Bekkers E, Es R van. Interpretable ECG classification via a query-based latent space traversal (qLST). *Arxiv* 2021.

Siegersma KR, **Leur RR van de**, Onland-Moret NC, Leon DA, Diez-Benavente E, Rozendaal L, Bots ML, Coronel R, Appelman Y, Hofstra L, Harst P van der, Doevendans PA, Hassink RJ, Ruijter HM den, Es R van. Deep neural networks reveal novel sex-specific electrocardiographic features relevant for mortality risk. *European Hear J - Digital Heal* 2022.

Wouters PC, **Leur RR van de**, Vessies MB, Stipdonk AM van, Ghossein MA, Hassink RJ, Doevendans PA, Harst P van der, Maass AH, Prinzen FW, Vernooy K, Meine M, Es R van. ECG-based deep learning improves outcome prediction following cardiac resynchronization therapy. *European Heart Journal* 2022.

Leur RR van de, Bleijendaal H, Taha K, Mast T, Gho JMIH, Linschoten M, Rees B van, Henkens MTHM, Heymans S, Sturkenboom N, Tio RA, Offerhaus JA, Bor WL, Maarse M, Haerkens-Arends HE, Kolk MZH, Lingen ACJ van der, Selder JJ, Wierda EE, Bergen PFMM van, Winter MM, Zwinderman AH, Doevendans PA, Harst P van der, Pinto YM, Asselbergs FW, Es R van, Tjong FVY, consortium C-C collaborative. Electrocardiogram-based mortality prediction in patients with COVID-19 using machine learning. *Neth Heart J* 2022;30:312–318.

Leur RR van de, Bos MN, Taha K, Sammani A, Yeung MW, Duijvenboden S van, Lambiase PD, Hassink RJ, Harst P van der, Doevendans PA, Gupta DK, Es R van. Improving explainability of deep neural network-based electrocardiogram interpretation using variational auto-encoders. *European Hear J - Digital Heal* 2022.

Leur RR van de, Hassink RJ, Es R van. Variational auto-encoders improve explainability over currently employed heatmap methods for deep learning-based interpretation of the electrocardiogram. *European Hear J - Digital Heal* 2022.

Curriculum vitae

Rutger van de Leur was born on 9 August 1994 in Arnhem to Joris van de Leur and Liesbeth Jans. He grew up in Dieren, Gelderland and attended the Stedelijk Gymnasium Arnhem, along with his younger sister, Cecile van de Leur. He graduated summa cum laude in 2012 from secondary school. After moving to Utrecht, Rutger pursued studies in both Medicine and Mathematics at Utrecht University.

During his academic journey, Rutger actively participated in research projects. Initially, he conducted research on delirium in the intensive care unit as part of the Master's Honors Programme, under the guidance of Prof. Dr. A. Slooter. This research ignited his interest in medical statistics, artificial intelligence, epidemiology, and programming. In 2017, he began another Master's program in Epidemiology, alongside Medicine, with a particular focus on medical statistics. His elective courses included attending a Data Science Summer School at the London School of Economics and a Deep Learning course at the Amsterdam School of Data Science. Additionally, Rutger served as a part-time reservist in the Royal Netherlands Army from 2014 to 2017. In 2020 and 2021, he successfully obtained degrees in both Medicine and Epidemiology from Utrecht University.

In 2018, as part of the research project of his Master Epidemiology, Rutger started his research on deep learning for automated ECG analysis under supervision of dr. R. van Es, dr. R.J. Hassink and prof. dr. P.A. Doevendans. In 2018 they were awarded a ZonMw and Dutch Heart Foundation grant of 1.2 million euros to continue their research on this subject. That grant enabled him to fund his PhD project.

As a PhD student, Rutger supervised over 20 Master's students from the Artificial Intelligence program at the University of Amsterdam and the Medicine program at Utrecht University. Furthermore, he actively presented at multiple international conferences and co-authored more than 20 publications. He was also a co-applicant for three awarded grants, with a total budget of 1.8 million euros. Towards the end of his PhD, Rutger participated in a nine-month valorization program at UtrechtInc, working to bridge the gap between research and implementation.

Rutger will start as a cardiologist-in-training December 1st 2024 at the Diakonessenhuis Utrecht, after having worked at the Meander Medical Center in Amersfoort and the UMC Utrecht. Additionally, he co-founded and will be involved in Cordys Analytics B.V., a spin-off of UMC Utrecht that focuses on the valorization of artificial intelligence algorithms for ECG analysis.

Dankwoord

Naast mijn drie jaar promotieonderzoek, ben ik sinds 2019 al als student al begonnen met het doen van onderzoek binnen de Cardiologie van het UMC Utrecht. Alle projecten beschreven in dit proefschrift zouden niet mogelijk zijn geweest zonder de hulp van een groot aantal betrokkenen. Zonder iemand tekort te doen, wil ik hieronder een aantal mensen in het bijzonder bedanken.

Mijn promotor, **prof. dr. P.A. Doevendans.**

Beste **Pieter**. Ik herinner me nog goed hoe ik als zenuwachtige student ons eerste project probeerde te ‘pitchen’. Het vertrouwen en de steun die je toen toonde heeft uiteindelijk de subsidie mogelijk gemaakt waarop het hele proefschrift gebouwd is. Ik wil je graag bedanken voor alle snelle feedback tijdens mijn promotietraject, de ‘out-of-the-box’ ideeën tijdens de consortiummeetings en hoe we allerlei logistieke en praktische problemen uit de weg konden helpen met een ‘ok, PD’. Bedankt voor alles.

Mijn copromotoren, **dr. R. van Es** en **dr. R.J. Hassink**.

Beste **René**, als student kwam ik bij jou omdat ik graag ‘iets technisch’ binnen de Cardiologie wilde gaan doen als onderdeel van de Master Epidemiologie. Het enthousiasme en de brede blik die je toen toonde hebben me snel overgehaald om in jouw groep te beginnen, eerst met electroporatie maar tijdens onze vele dartende brainstormsessies in de Villa uiteindelijke bij AI. Dank dat je me de kans hebt geboden samen de subsidieaanvraag te schrijven die dit project mogelijk heeft gemaakt, en alle hulp en zeer laagdrempelige betrokkenheid nadien. Het blijft indrukwekkend hoe je pragmatisme en innovatie weet te combineren. Ik zie er naar uit samen de implementatie van AI voor het ECG mogelijk te maken.

Beste **Rutger**, ik ben heel erg blij jou als copromotor in het team gehad te hebben. Het was altijd fijn om weer nuchter naar de zaken en klinische implicaties te kunnen kijken, en dan toch weer mooie en innovatieve toepassingen te vinden. Ondanks vele pogingen is het idiopathisch ventrikelfibrilleren cohort (tot nu toe) zelfs niet met AI te kraken helaas. Die heb je nog tegoed. Ik kijk er erg naar uit als 'Rutger van Rutger' in het Utrechtse aan de opleiding te beginnen en nog verder samen te werken.

Ik wil de leden van de leescommissie, **prof. dr. F.W. Asselbergs, prof. dr. D.W. Donker, prof. dr. M. Meine, prof. dr. D.L. Oberski en dr. M. van Smeden,** graag bedanken voor het kritisch beoordelen van mijn proefschrift.

Ook dank aan alle leden van het 'Hart voor Duurzame Zorg' consortium (**Pieter Glijnis, Joris Jaspers, Rien van der Zee, Thierry Wildbergh** en alle betrokkenen vanuit de **Hartstichting en ZonMw/NWO**) voor de waardevolle input tijdens onze consortiummeetings.

Daarnaast wil ik graag mijn dank uitspreken voor alle collega's in het UMC Utrecht die een bijdrage hebben geleverd aan de vele samenwerkingsprojecten die we hebben uitgevoerd. In het bijzonder EFOlogen **Mathias Meine, Peter Loh, Nick Clappers, Jeroen van der Heijden en Anneline te Riele**. Dank voor jullie input en ideeën voor de meest uitlopende projecten. Ik ben jullie nog steeds dankbaar voor de bereidwilligheid om deze, toen nog jonge, student te helpen door ontelbare ECG's te labelen. Ik kan me nog goed herinneren hoe mooi ik het vond toen ik zelf, al was het pas tijdens EFO ASAS, ook doorkreeg hoe ik een ECG écht moest lezen. De kennis die jullie hebben over ECGs zal niet snel door AI worden gevat. Ook gaat mijn dank uit naar **Maarten Jan Cramer, Dirk van Osch en Arco Teske** voor alle input bij de echoprojecten, waarvan ik hoop dat we ze nog verder kunnen uitbouwen. **Linda van Laake**, die circadian rhythm in the QT interval paper gaat er komen!

Prof. dr. P. van der Harst, beste Pim, ik ben dankbaar voor je vertrouwen en het mogelijk maken en ondersteunen van mijn projecten binnen het UMC. Vanaf dat je halverwege mijn promotietraject instapte als nieuw hoofd van de afdeling heb ik (nog steeds) bewondering voor jouw immer snelle gedachtegang. Ik hoop dat we nog mooie onderzoekslijnen mogen uitzetten samen, en de onderzoeksgroep verder kunnen uitbouwen.

Alhoewel alle inhoudelijke bijdragen erg belangrijk waren, is de meest onmisbare bijdrage aan dit proefschrift toch wel gekomen van **Ronald Groenemeijer.** Zonder jou was het niet mogelijk om de ECG-database van het UMC Utrecht van de afgelopen 25 jaar te ontsluiten en hadden we dit onderzoek nooit kunnen doen. Ook dank aan iedereen van het **Stafsecretariaat, Leon van Iersel, Petra Berk, Geoffrey van der Plasse en datamanagers** van de divisie voor alle ondersteuning tijdens dit traject.

Mede Villabewoners, **Einar, Steven, Bas, Mira, Thijs, Marijn, Sanne, Philippe, Karim, Feddo, Melle, Bauke, Willem**, bedankt voor de gezelligheid op onze bijzonder ruime kamer en dat jullie mij opgenomen hebben als jonge student. Het blijft een ruimte die inspireert tot sport, van schaken tot darten, voor alles is er genoeg tijd. Begrijpen van het koffieschema zou niet misstaan als onderdeel van de sollicitatieprocedure voor een PhD positie.

Dan ook de andere mede-PhDers. **Arjan**, blij dat we veel samen hebben mogen werken, en dat ik je de wonderlijke wereld van Python heb kunnen bijbrengen - al weet ik niet of jij daar ook nog zo blij mee bent. **Karim**, de vele brainstormsessie via Teams tijdens COVID en later bij het dartboard, hebben tot hele mooie projecten geleid. Die straincurves krijgen we er ook nog wel onder. **Nicole, Hugo en Rosanne**, de Q-crew, het was mooi om samen een weekend te organiseren. Jammer dat jullie bijna allemaal uitgevlogen zijn, en tot in het Diak Nicole! **Melle**, ik ken geen enkele engineer die zo goed ECGs kan lezen als jij. Zonder jouw hulp en innovatieve ideeën was dit proefschrift nooit mogelijk geweest, heel blij dat is zo intensief met je heb mogen samenwerken en het stokje aan je mocht overdragen. **Alejandra**, muchas gracias! Indrukwekkend hoe je vanuit bijna niets en heel ecg-apparaat hebt weten te ontwerpen, produceren en testen in de kliniek. Succes met de laatste loodjes. **Luis en Bram**, dank voor jullie input tijdens je klinische werk, en mooi dat we nu zo veel samen kunnen werken. **Bauke**, mooi dat we samen de implementatie van AI ook gaan onderzoeken, dat is toch echt 'the next big thing', en op een mooie tijd in Londen! **Bas, Dino en Camila,** welkom bij de groep, goed dat jullie erbij zijn! **Aernoud** (zo veel gedaan en toch bescheiden, laat me vooral weten als we een keer aan de Amazone kunnen klussen), **Anne-Mar** (top dat we samen aan de opleiding mogen beginnen)**, Cheyenne, Evangeline, Fahima, Feddo, Iris, Klaske, Lieke, Lisa** (we vinden ooit nog iets in die IVF ECGs)**, Machteld, Manon, Marijke, Mark, Markella, Max, Mimount, Rob, Sanne, Sayonara, Steven, Thomas en Timion**, dank voor de gezelligheid tijdens de vele borrels, weekenden en lunches.

Uiteraard waren de onderzoeken uit dit boekje ook nooit mogelijk geweest door de bijdrages van het indrukwekkende leger aan studenten, onder andere **Meike, Nynke, Mathieu, Bjarne, Hidde, Balint, Jasper, Bram, Johanneke en Faye**. Ik zou **Max** en **Jeroen** in het bijzonder willen bedanken voor jullie belangrijke bijdragen aan dit proefschrift. Ook zou ik zou graag **Hidde, Fleur, Remco, Jan Walter, Ming en Michiel** uit het UMCG, AUMC en MUMC willen bedanken voor de mooie samenwerkingen.

Ik zou ook graag iedereen in het **Meander Medisch Centrum** willen bedanken voor de mooie tijd die ik daar heb gehad als ANIOS. In het bijzonder zou ik graag **Thierry** en **Marjan** willen bedanken voor alle hulp met het testen van het miniECG.

Dit proefschrift is ook het startpunt geweest van **Cordys Analytics**, waarmee ik hoop de brug tussen de kliniek en innovatie, die ook in dit boekje centraal staat, verder uit te bouwen. **John van den Berg**, veel dank voor het vertrouwen in Cordys en dat je als CEO het avontuur met René en mij aandurft. I would also like to express my gratitude to **Marcel, Nick, Kees, Parmenion, Jenna, Cristian and Guillemette** for your hard work as members of the Cordys team!

Ten slotte, was mijn promotietijd (en de eerste stappen in de onderzoekswereld) nooit zo leuk geweest als ik de volgende mensen niet zou kennen. Ik ken velen van jullie al lang en ben dankbaar dat ik me met zoveel leuke mensen mag omringen. **Martijn**, de bassischool ligt inmiddels al lange tijd achter ons, maar nog steeds klinkt in onze gesprekken die nuchterheid uit Dieren door. **Bart**, van tennismaat tot co-gebruiker van Yoda, waarmee we beiden de wereld verkennen. Onze klussessies hebben er meermaals voor gezorgd dat ik mijn zinnen tijdens dit promotietraject kon verzetten. **Joost en Jurriaan**, ook wij hebben een geliefde auto die ons verbindt (in totaal nu 4/6e auto in bezit). Inmiddels drukbezette mannen, die gelukkig nog steeds tijd vrijmaken om Saba een veilig onderkomen te geven. Of zij ooit nog elektrisch gaat worden weet ik niet. **Beerforce 1,** sinds de middelbare school al vrienden en nog steeds vertrouwde gezichten, ondanks dat niemand meer in de buurt van Arnhem woont en we allemaal drukke levens hebben. Onze wandel- en wielrentochten door de Europese bergen, inclusief filosofische gesprekken, soms (helaas) met muziek, zijn toch wel een jaarlijks hoogtepunt. **Kikvorsmannen**, sinds de oprichting van DRAFT zijn we hechte vrienden waar ik altijd op kan rekenen. Ik hoop dat we nog veel mogen lachen samen en, ondanks dat we nu langzaamaan volwassen worden, elkaar nog veel blijven zien. Ik zou ook alle generaties huisgenoten van het **Haviksnest** willen bedanken voor de onvergetelijke tijd – van eerste studentenjaren tot de eerste (COVID)meetings, we hebben het allemaal meegemaakt. Tijdens mijn studie heb ik via JHG ook mooie vriendschappen mogen sluiten. Ik wil iedereen bedanken voor de gezelligheid, de fietstochten en eindeloze sparsessies over het mooie artsenvak. **Stefan Koudstaal** wil ik in het bijzonder bedanken voor het introduceren van deze student bij de cardiologie voor een onderzoeksproject, daar is uiteindelijk alles begonnen.

Mijn paranimfen, **Casper en Philippe**, dank dat jullie mijn paranimfen willen zijn. Door jullie weet ik zeker dat ik met een gerust hart, dit promotietraject kan afronden. **Cas**, we kennen elkaar al lang en in vele gedaantes: naast samen aspirant bij JHG, huisgenoot en sinds kort boomer, kunnen we nu ook paranimf aan het rijtje toevoegen. Dank voor de goede relativerende gesprekken (op de racefiets), ik hoop dat er nog velen mogen volgen! Pas op dat de boomer in je niet te veel en/of te vroeg naar boven komt, nu het volwassen leven begint. **Philippe**, op het moment dat wij veel te enthousiast in onze self-driving Tesla in Yosemite bijna de dood vonden, samen met een lokaal hertje, was een vriendschap geboren. Mooi om ook jouw paranimf geweest te zijn (al zal ik jouw niveau van prozaïsch taalgebruik nooit bereiken). Tot in het Diak!

Lieve **papa en mama**, dank voor een fijne liefdevolle opvoeding en alles daarna, het heeft mij zonder meer gebracht tot waar ik nu ben. Ik hou van jullie. Lieve **Ciel**, dank voor de vele goede gesprekken, je nuchtere kijk op zaken en het stellen van de juiste (moeilijke) vragen als dat nodig was. Ik hoop dat iemand met jouw zorgvuldigheid ook wordt verleid tot de wetenschap.

Lieve **Anne**, het blijft bijzonder hoe we allebei gegroeid zijn tijdens onze promotietrajecten. Dank om me altijd weer op de rit te krijgen als ik (weer) net iets te veel ballen in de lucht probeerde te houden. Ik hoop dat ik jou ook door het laatste deel van je promotietraject weet te loodsen. Je bent een geweldig goed iemand, ik ben heel blij om aan jouw zijde te zijn en blijf altijd zeggen wat je vindt! Ik hou van jou.