Nils Gumpfer, Dimitri Grün, Jennifer Hannig, Till Keller and Michael Guckert*

Detecting myocardial scar using electrocardiogram data and deep neural networks

<https://doi.org/10.1515/hsz-2020-0169>

Received April 17, 2020; accepted August 30, 2020;

published online October 2, 2020

Abstract: Ischaemic heart disease is among the most frequent causes of death. Early detection of myocardial pathologies can increase the benefit of therapy and reduce the number of lethal cases. Presence of myocardial scar is an indicator for developing ischaemic heart disease and can be detected with high diagnostic precision by magnetic resonance imaging. However, magnetic resonance imaging scanners are expensive and of limited availability. It is known that presence of myocardial scar has an impact on the well-established, reasonably low cost, and almost ubiquitously available electrocardiogram. However, this impact is non-specific and often hard to detect by a physician. We present an artificial intelligence based approach – namely a deep learning model – for the prediction of myocardial scar based on an electrocardiogram and additional clinical parameters. The model was trained and evaluated by applying 6-fold cross-validation to a dataset of 12-lead electrocardiogram time series together with clinical parameters. The proposed model for predicting the presence of scar tissue achieved an area under the curve score, sensitivity,

specificity, and accuracy of 0.89, 70.0, 84.3, and 78.0%, respectively. This promisingly high diagnostic precision of our electrocardiogram-based deep learning models for myocardial scar detection may support a novel, comprehensible screening method.

Keywords: artificial intelligence; convolutional neural networks; deep learning; ECG classification; ischaemic heart disease; myocardial scar.

Introduction

Ischaemic heart disease (IHD) is a major cause of death in countries of Western life-style and therefore not only causes personal suffering but also has a high socio-economic impact (Benjamin et al. 2019; World Health Organization 2018). Early detection of IHD with subsequent treatment initiation can reduce mortality and alleviate the course of the disease. As a preliminary stage of IHD, development of myocardial scar (MS) is a potential indicator and early detection of MS together with appropriate therapeutic measures can prevent IHD or at least reduce effects of the disease.

Non-invasive state of the art methods for diagnosis of MS all suffer from drawbacks impeding their use in comprehensible screening. Magnetic resonance imaging (MRI) is considered as the gold standard for diagnosing MS by achieving high sensitivity and specificity (Winau et al. 2018). Intramyocardial late Gadolinium enhancement (LE) of the contrast agent can be visualised by MRI and has been shown to serve as an indicator for scar tissue (Kim et al. 1999). The drawbacks of the application of MRI as a screening method are its limited availability, high costs as well as the time demanding imaging process with a high noise exposure affecting the comfort of the patients (Oikarinen et al. 2013).

Presence of MS has an impact on the electrocardiogram (ECG) signals, but the effects are heterogeneous and complex in interpretation so that health-care professionals do not achieve the required diagnostic sensitivity in clinical routine. For this reason, the abundantly available ECG cannot be applied as an alternative for MRI (Inoue et al. 2017; Markendorf et al. 2019).

Jennifer Hannig, Till Keller and Michael Guckert: These authors share last authorship on this work.

***Corresponding author: Michael Guckert,** Cognitive Information Systems, KITE-Kompetenzzentrum für Informationstechnologie, Technische Hochschule Mittelhessen -University of Applied Sciences, 61169Friedberg, Germany; Department of MND-Mathematik, Naturwissenschaften und Datenverarbeitung, Technische Hochschule Mittelhessen-University of Applied Sciences, Wilhelm-Leuschner-Straße 13, 61169 Friedberg, Germany, E-mail: michael.guckert@mnd.thm.de.

Nils Gumpfer and Jennifer Hannig, Cognitive Information Systems, KITE-Kompetenzzentrum für Informationstechnologie, Technische Hochschule Mittelhessen – University of Applied Sciences, 61169 Friedberg, Germany. <https://orcid.org/0000-0001-8644-9885> (N. Gumpfer). <https://orcid.org/0000-0002-2789-5540> (J. Hannig) **Dimitri Grün and Till Keller,** Department of Internal Medicine I, Cardiology, Justus-Liebig-University Gießen, 35390 Gießen, Germany. <https://orcid.org/0000-0002-8668-636X> (D. Grün). <https://orcid.org/0000-0002-0895-6491> (T. Keller)

Scoring systems for the evaluation of ECG recordings have been developed since the early 1970s, for example the Selvester QRS score (Selvester et al. 1971; Selvester et al. 1985). However, it has been shown that this score can only achieve high sensitivity and specificity for the detection of large sized MS, whereas the diagnostic performance of small or medium sized MS is limited (Rosengarten et al. 2013). Numerous studies used machine learning (ML) or deep learning (DL) algorithms for ECG-based diagnosis of cardiac diseases. For example, detection of arrhythmia has been achieved with classic ML approaches such as support vector machines (SVM) (Albuquerque et al. 2018). For identification of acute myocardial infarction, several convolutional neural network (CNN) models have been proposed, achieving high diagnostic performance (Acharya et al. 2017; Baloglu et al. 2019; Liu et al. 2018; Strodthoff and Strodthoff 2019). A CNN is a bio-inspired neural network architecture which is based on an internal structure with analogies to the visual cortex of the animal eye (Hubel and Wiesel 1962; LeCun et al. 1989).

For the automated detection of MS, so far only the classic ML method SVM has been used (Dima et al. 2013; Goovaerts et al. 2019). This method requires pre-extraction of features to make ECG recordings accessible for algorithmic processing. These features either have to be defined or identified using signal-processing procedures such as the time-domain morphology algorithm (Mazomenos et al. 2012) or variational mode decomposition (Dragomiretskiy and Zosso 2014).

In this paper, we present a DL model for the detection of MS that uses raw ECG recordings and corresponding clinical parameters for each patient (e.g. cardiovascular risk factors), without prior knowledge about medical interpretations of patterns. Such a model may be further elaborated to allow future clinical use as a simple, rapid, and cheap screening test for MS.

Parameters of the model were calibrated using supervised learning on the given set of ECG recordings. During this process, the model approximates the underlying mathematical decision function.

Results

In this section, we first describe the data pipeline used for data cleansing and preprocessing. Second, we present two DL architectures. One for a model based on raw ECG recordings and a second architecture based on raw ECG recordings and additional clinical parameters. For extensive information about the applied methods, we refer to section Materials and methods.

Data pipeline

Data for both training and application of the trained models is supplied through a data pipeline consisting of several preprocessing steps arranged in two separate branches for ECG recordings and for clinical parameters, respectively (Figure 1).

First, ECG recordings stored in XML HL7v3 format (Health Level Seven, Inc. 2020) are loaded in which then, artefacts, e.g. zero lines at the beginning and end of the recordings, are removed by cropping. Then the values are scaled to locate them in a narrower range, which makes them more accessible for deep learning algorithms. In the next step, data augmentation procedures (subsampling) are applied to compensate for the relatively small number of training records before data is saved in the last step of the data pipeline.

Parallel to preprocessing the ECG recordings, clinical parameters run through a second branch of the pipeline. After initial processing, i.e. loading and categorisation, attributes are transformed into one-hot encoded vectors which are then concatenated into a single vector to be used as input for the neural network.

The pipeline provides data for two experimental setups in which models for an ECG-based model and for a combined model based on ECG data and clinical parameters are examined. Both strands with preprocessed ECG samples and encoded clinical parameters can be combined using an anonymised distinct identifier (record-ID). This enables to link clinical parameters to the corresponding ECG recording so that these two types of data can either be used together or separately.

Figure 1: Overview of the data pipeline for preprocessing.

ECG model

CNNs have originally been designed for pattern identification in images (LeCun et al. 1989) and have already been used for ECG analysis, yielding considerable performance for specific tasks (e.g. Strodthoff and Strodthoff (2019)). Therefore, CNN are a promising architectural choice for the task of detecting MS in ECG recordings.

As starting point of our DL model, we used the CNN architecture as proposed by Strodthoff and Strodthoff (2019), which was developed for identification of acute MI. The following characteristics of the architecture were optimised for our specific medical purposes by grid search optimisation:

- Number of convolutional layers
- Number of filters in convolutional layers
- Kernel/filter size of convolutional layers
- Pooling (yes/no) after convolutional layers
- Additional dropout between pooling and convolutional layers with varying dropout rates

Furthermore, we extended the initial starting configuration with additional fully-connected dense layers, experimentally adjusting their number, width, and the respective dropout rates.

The final optimised CNN (Appendix, Figure 3) consists of an input layer (width = 2000) followed by a successive series of convolutional and pooling layers reducing the input to a width of 64. In between these layers, dropout with a rate of 0.25 proofed to successfully prevent the model from overfitting. Four fully-connected layers (width = 64), each followed by a dropout rate of 0.1, aggregate the information into a final layer. This layer yields a probability distribution over two classes (presence of MS versus absence of MS) based on a softmax activation.

Combined model

Besides the experiments with ECG recordings alone, we investigated whether using ECG data together with clinical parameters of the individual patient in a combined DL model could improve diagnostic performance. Therefore, the CNN model described before was combined with a fully-connected feedforward network (FFN) which was applied to the records with clinical parameters (Appendix, Figure 4). The characteristics of this network were optimised analogously to that of the CNN. The final FFN structure has an input layer of 55 nodes followed by four fully-connected layers (width = 256). All layers but the first are connected with dropout rates of 0.5.

The outputs of the CNN and FFN form the 4-node input for a final FFN with three hidden layers (width = 4), each having a dropout rate of 0.25. As for the ECG-only model, the final output layer of the combined model contains two nodes for the overall prediction (presence of MS, absence of MS) and uses a softmax activation. The combined model was trained in a separate end-to-end training process and does not rely on pre-trained sub-models.

Experimental results

We now present performance measures for both models. Table 1 contains areas under the curve (AUC) of receiver operating characteristic analyses, sensitivities, specificities, and accuracies of the ECG-only (listed in the columns on the left hand side) and the combined model (listed on the right hand side). The measures were determined on subsample-level and on patient-level applying 6-fold cross-validation with 50 repetitions for each split, see paragraph Performance metrics of section Materials and methods.

In summary, the ECG model achieved a mean AUC, sensitivity, specificity, and accuracy on patient-level of 0.81 ± 0.06 , 70.0 ± 2.4 , 73.0 ± 20.7 , and $70.2 \pm 7.9\%$, respectively. The mean AUC, sensitivity, specificity, and accuracy on patient-level of the combined model are 0.89 ± 0.06 , 70.0 ± 20.4 , 84.3 ± 13.9 , and $78.0 \pm 7.0\%$, respectively.

With an outlook to a potential clinical application, the positive predictive value (PPV) and negative predictive value (NPV) of the ECG model would be 73.0 ± 17.5 and $72.3 \pm 11.3\%$, respectively. The PPV and NPV of the combined model would be 84.2 ± 12.2 and $78.2 \pm 14.0\%$, respectively.

The results of both models are discussed comparatively and related to alternative approaches in the following section. Graphs of the receiver operating characteristic (ROC) analyses can be found as Figure 5 in the Appendix.

Discussion

In this section, we discuss the main findings of this study, compare our results to state of the art approaches, and point out limitations as well as potential directions of future work.

Model performance

Our results presented in Table 1 show that combining ECG data and clinical parameters produced the best average

Table 1: Results of different model variants based on 6-fold cross-validation.

Metric	ECG model														Combined model	
	Cross-validation split						Cross-validation split									
	1	2	3	4	5*	6	Mean ± SD	1	2	3*	4	5	6	Mean ± SD		
AUC	Subsample-level	0.71	0.84	0.86	0.77	0.85	0.83	0.81 ± 0.05	0.85	0.89	0.99	0.86	0.93	0.83	0.89 ± 0.05	
	Patient-level	0.71	0.82	0.76	0.87	0.89	0.83	0.81 ± 0.06	0.88	0.89	0.99	0.87	0.93	0.80	0.89 ± 0.06	
	Subsample-level	66.0	87.0	44.0	85.0	65.0	71.0	69.7 ± 14.3	78.0	50.0	54.0	67.0	97.0	86.0	72.0 ± 16.8	
	Patient-level	73.0	75.0	45.0	86.0	70.0	71.0	70.0 ± 12.4	82.0	50.0	45.0	57.0	100.0	86.0	70.0 ± 20.4	
Specificity (%)	Subsample-level	68.0	81.0	92.0	51.0	88.0	77.0	76.2 ± 13.6	71.0	91.0	100.0	96.0	77.0	73.0	84.7 ± 11.4	
	Patient-level	50.0	82.0	100.0	42.0	89.0	75.0	73.0 ± 20.7	62.0	91.0	100.0	100.0	78.0	75.0	84.3 ± 13.9	
Accuracy (%)	Subsample-level	67.0	84.0	64.0	64.0	76.0	75.0	71.7 ± 7.3	75.0	74.0	74.0	85.0	88.0	78.0	79.0 ± 5.5	
	Patient-level	63.0	79.0	68.0	58.0	79.0	74.0	70.2 ± 7.9	74.0	74.0	68.0	84.0	89.0	79.0	78.0 ± 7.0	

Best model according to highest AUC on patient-level marked with *. AUC = area under receiver operating characteristic curve; NPV = negative predictive value; PPV = positive predictive value; SD = standard deviation.

performance over the six cross-validation runs with respect to each of the measures (AUC = 0.89, Sensitivity = 70%, Specificity = 84.3%, and Accuracy = 78%). The same holds for the maximal values and for the results of the single runs. Apparently, supplementary information in the form of clinical parameters can improve model performance.

We observed high measures for AUC, sensitivity and specificity but not necessarily in the same models. The model with the highest AUC can be seen as the model with the best compromise between specificity and sensitivity and depending on the application, models selected based on maximum sensitivity or maximum specificity can be more appropriate. Our approach computes models with good performance for all three perspectives.

The area under the receiver operating characteristic curve was used as performance measure as this statistical method is well established in medical studies evaluating diagnostic tests. With specific clinical requirements, such an unweighted compromise might not be ideal. For example, within a diagnostic rule-out approach optimal sensitivity is aimed, whereas in case of a diagnostic rule-in algorithm, a high specificity is of greater importance.

Due to the relatively small case number of $n = 114$, we augmented data by subsampling and applied 6-fold cross-validation. Therefore, our validation datasets contained data from 19 patients in each cross-validation run so that validation was performed with $n = 1900$ on subsample-level and $n = 19$ on patient-level. Predictions on subsample-level were condensed into patient-level predictions by computing the mean of all the subsample predictions of the corresponding patient. Surprisingly, standard deviation of model performance on subsample-level is close to that on patient-level despite the different size of the validation datasets. Furthermore, there is no significant difference between the model performance on subsample-level and on patient-level, indicating that subsampling together with accumulating predictions to patient-level is viable.

More elaborate statistical evidence, e.g. confidence intervals or test measures, cannot be reported as the influence of fluctuations is too strong for the sample size given (see discussion in Isaksson et al. (2008) and Bengio and Grandvalet (2004)). We therefore applied cross-validation with repeated training and validation (50 times for each of the six split-variants), reporting mean and variance values for sensitivity, specificity, accuracy, and AUC.

Using accuracy must be handled with care as it is considered to fail as performance metric for imbalanced class distributions. Table 3 shows that patients with LE and those without are almost equally distributed so that accuracy is applicable in our case.

Comparison to existing approaches

Table 2 relates the performance of our models to that of applying the Selvester score (Selvester et al. 1971; Selvester et al. 1985), human-based ECG evaluation, and two SVM-based approaches. The measures of both our models deliver comparably high performance measures and reach (ECG-based model) or exceed human performance (combined model) as published by Asch et al. (2006), Carpenter et al. (2015), and Markendorf et al. (2019).

Comparing our results with that of the other methods listed in Table 2, it must be taken into account that the underlying cohorts differ in size, structure and time of study. Considering the larger training sets used by Goovaerts et al. (2019) and by Dima et al. (2013) and the limited number of cases used by Chaudhry et al. (2017) and Bignoto et al. (2018), our results are similar. For a more substantiated comparison, the methods would have to be applied to the identical data, see Bengio and Grandvalet (2004). It is noteworthy that only Goovaerts et al. (2019) uses AUC as performance measure which is of high relevance in the medical context and is the only measure that does not depend on a single threshold.

In summary, our proposed model achieves reasonable performance on a relatively small case number used for training, which is a promising result, encouraging further research and validation.

Limitations

Over recent years, DL has frequently been used in the medical context. A recent meta-analysis has shown that

existing DL models detecting diseases by analysis of medical images achieve diagnostic performance equivalent to that of health-care professionals (Liu et al. 2019). However, only 25 studies out of 31587 under investigation could be included in this meta-analysis. The rest had to be left out for reasons such as missing external validation or lack of transparency. Limited amount of appropriate data is often one of the main obstacles for a comprehensive use of DL models in general and particularly in healthcare.

In case of the approach presented in this article, the limited number of 114 patients and the heterogeneity within the data also constituted a key limitation for model development. Due to this limitation, we exclusively performed an internal validation (in-sample validation) using 6-fold cross-validation, see Altman and Royston (2000) for a description of different types of validation sets. An external validation (out-of sample validation) with data derived from new, previously unknown patients is currently missing and will be critical to ensure the high performance measures achieved in this study.

It should further be mentioned that the data partly includes redundant information, e.g. different leads of the ECG may contain similar patterns, or that the clinical parameters body-mass-index (BMI) and body-surface-area (BSA) are not completely independent.

Furthermore, all data used for training and validation was derived from patients scheduled for an MRI examination. This means that ECG recordings, MRI diagnosis, and clinical parameters used in our approach were obtained in patients with a relevant risk for cardiovascular diseases. It is uncertain whether the performance achieved

Table 2: Comparison of our models to published methods for detection of myocardial scar.

Method	Reference	AUC	Sensitivity (%)	Specificity (%)	Accuracy (%)	Patients (n)
ECG model	Proposed method (mean performance)	0.81	70.0	73.0	70.2	114
ECG model	Proposed method (highest AUC)	0.89	70.0	89.0	79.0	114
ECG + clin. param. model	Proposed method (mean performance)	0.89	70.0	84.3	78.0	114
ECG + clin. param. model	Proposed method (highest AUC)	0.99	45.0	100.0	68.0	114
Selv. QRS score (compared to LE in MRI)	Chaudhry et al. (2017)	N/A	57.0	48.0	N/A	60
Selv. QRS Score (compared to LE in MDCT)	Bignoto et al. (2018)	N/A	84.8	88.8	85.3	75
Human-based ECG eval. (Q-waves)	Asch et al. (2006)	N/A	48.4	83.5	N/A	66
Human-based ECG eval. (Q-waves)	Carpenter et al. (2015)	N/A	36.3	79.8	66.1	498
Human-based ECG eval. (Q-waves)	Markendorf et al. (2019)	N/A	70.0	40.0	N/A	149
Human-based ECG eval. (fQRS)	Markendorf et al. (2019)	N/A	46.0	59.0	N/A	149
SVM-based ECG eval. (Q-waves)	Dima et al. (2013)	N/A	87.3	91.2	89.2	260
SVM-based ECG eval. (fQRS)	Goovaerts et al. (2019)	0.95	86.0	89.0	88.0	723

Bold: Model with highest mean AUC as compromise between sensitivity and specificity. AUC = area under receiver operating characteristic curve; ECG = electrocardiogram; fQRS = fragmented QRS complex; LE = late enhancement; MRI = magnetic resonance imaging; MDCT = multidetector computed tomography; NPV = negative predictive value; PPV = positive predictive value; SVM = support vector machine.

for the cohort used in this study can be maintained for a putative healthy cohort.

Another limitation is the black box nature of the DL model proposed in this study. We currently do not elucidate which patterns of the ECG or features of the clinical parameters contribute to the final decision of the DL model and how this decision is reached.

Future work

Currently, enrolling patients and therefore collecting further medical data is an ongoing process of our project and will allow us to use a more comprehensive dataset for model training and hyperparameter optimisation in the future. A larger set of medical data will also enable to hold-out data for external validation. Therefore, our main focus will be to test the generalisation performance of our DL model by reproducing the promising performance measures achieved in the presented study for this out-of-sample data. Furthermore, an interesting aspect for validation will be the comparison of the diagnostic performance of our DL model with established diagnostic approaches used in daily clinical routine.

Another important issue for future work is to elucidate the specific ECG patterns and relevant clinical parameters leading to the diagnostic prediction of the DL model. Potential methods from the emerging field of explainable artificial intelligence (XAI) such as layer-wise relevance propagation (LRP) (Bach et al. 2015), local interpretable model-agnostic explanations (LIME) (Ribeiro et al. 2016) and deep learning important features (DeepLIFT) (Shrikumar et al. 2017) may provide transparency and traceability and thereby dissolve the black-box nature of the DL model. Such methods can further facilitate the use of DL methods in clinical routine due to an increase of trust (Andras et al. 2018).

Conclusion

In spite of the comparatively small case number in the presented study, the developed DL models reached promising performance measures to detect MS. The procedure can be applied to ECG data and clinical parameters directly and does not require any pre-extraction of features, which enables simple and flexible application. The potential reduction of expensive MRI and the possible use as screening tool for MS justify further experimentation. Future implementations that provide validation and explainability of DL

predictions are essential to increase credibility and trust in this promising technology and will help to establish its use for medical application.

Materials and methods

Medical data

Data on clinical parameters, ECG and MRI are used from 114 patients enrolled in the Kerckhoff Biomarker Registry for the training and evaluation of the deep neural networks.

Clinical cohort: Participants of whom data was used in this project have been recruited from an ongoing cardiovascular imaging sub-study, a longitudinal prospective cohort study that started 2016 as part of the Kerckhoff Biomarker Registry (BioReg). This study included patients with clinical indication for cardiovascular magnetic resonance (CMR) and who were older than 18 years. All patients gave written informed consent and the study was approved by the local ethics committee. Patients enrolled with suspected coronary artery disease (CAD) or with suspected progress of an established CAD who therefore underwent CMR stress testing were eligible for further investigation within the present project. For the present analyses, only patients in whom an XML ECG and an MRI dataset were available were used, leading to a study sample of 114 patients.

Clinical variables: At baseline, the following used variables were obtained using questionnaires, patient history, examination and available medical documentation: sex, age, body-mass-index (Appendix, Formula 1), body-surface-area (m^2) (Appendix, Formula 2), smoking status, arterial hypertension (defined as systolic blood pressures ≥ 140 mm Hg and diastolic ≥ 90 mm Hg or the intake of antihypertensive medication), diabetes mellitus (defined as fasting concentration of blood sugar ≥ 126 mg/dl, non-fasting blood sugar ≥ 200 mg/dl or the intake of antidiabetic medication), dyslipidemia, presence of familial disposition to heart disease, and presence of CAD defined as atherosclerotic changes of the coronary vessels. Table 3 provides an overview of the distribution of cardiovascular risk factors in the used cohort. As only patients with relevant risk scheduled for an MRI were enrolled, the resulting cohort might not be representative for a general population. For comparison, prevalences described in the literature are given as Table 7 in the Appendix.

ECG data: In each patient of the evaluated cohort, a 12-lead ECG was obtained on the same day as the MRI examination. The resulting ECG recording contains time-series of each lead of 10s length and is stored in the XML-based HL7v3 format (Health Level Seven, Inc. 2020). The ECG dataset was generated using a commercially available electrocardiogram device (Cardiovit AT-102P, Schiller-Reomed AG, Obfelden, Switzerland) with the following parameters: Frequency range: 0.05–150 Hz; Measuring range: ± 300 mV; Sampling rate: 500 data points per second/5000 in 10 s (per lead); Digital resolution: 5 μ V/18 bit.

MRI data: Each evaluated patient underwent a clinically indicated MRI evaluation using standardised procedures. In this imaging

Table 3: Patient characteristics.

Variable	Overall patients (<i>n</i> = 114)	Unit
Age (mean ± SD)	65.85 ± 13.44	years
Sex (M/F)	86/28	n
Body-surface-area (mean ± SD)	2.01 ± 0.21	m ²
Body-mass-index (mean ± SD)	27.47 ± 4.59	kg/m ²
Arterial hypertension	92 (81%)	n
Diabetes mellitus	22 (19%)	n
Dyslipidemia	70 (61%)	n
Familial disposition of heart diseases	30 (26%)	n
Smoking (active)	24 (21%)	n
Smoking (ex)	48 (42%)	n
Coronary artery disease	58 (51%)	n
Chronic heart failure	11 (10%)	n
Late enhancement	54 (47%)	n

SD = standard deviation.

examination, structural conditions of the myocardium are assessed based on the temporal and spatial distribution of Gadolinium as contrast agent (Dulce et al. 1993; Saeed et al. 1989). Intact cell membranes cannot be passed by this contrast agent, what leads to higher contrast of structurally impaired tissue (e.g. scar) (Kim et al. 1999). This enables to identify myocardial areas showing so-called late contrast enhancement (LE) as areas with myocardial scar. This definition of final diagnosis was done in all patients by experienced cardiologists blinded to ECG data. The MRI dataset was generated using a clinically established MRI device (Magnetom Skyra 3T, Siemens Healthcare GmbH, Erlangen, Germany) with a stress MRI protocol using Adenosin for induction of pharmacological stress.

Algorithmic procedure

In the following subsection, we describe in detail components of the data pipeline presented within section Results. For an overview of the data pipeline, see Figure 1. Furthermore, we present the performance metrics used for evaluation of the DL models and provide details about the training procedure.

Cropping: By cropping, the ECG recordings were cleaned from artefacts that for instance may have been caused by inappropriate electrode contact. In the ECG recordings used in this work, artefacts were only present at start and end of the recordings, what simplified their elimination by removing the corresponding intervals.

Scaling: The ECG recordings contained values measured in the range of -2000 – 2000 μ V. We scaled the values to mV, yielding a range of -2 – 2 mV. This reduced the absolute distance between minimum and maximum values, what is a recommended preprocessing step for data to be fed into neural networks (Bishop 1995, p. 298f.).

Augmentation: In order to prevent the neural network from overfitting on the comparably small dataset of 114 records, we increased the amount of unique ECG samples by data augmentation via subsampling. This is medically justified as the presence of MS and the

thereby associated impaired conductivity of the myocardium should lead to permanent changes in the shapes of the ECG waveforms present in all subsamples. Therefore, shorter samples than a standard of 10 s contain features relevant for distinguishing pathological ECG with patterns caused by MS from those without such patterns.

In order to capture at least two heartbeats in each sample, a subsample size of 4 s was used. This approach has already been used previously (Strodthoff and Strodthoff 2019). Figure 2 illustrates the subsampling process used in this work. Subsampling generates new, distinct, non-identical samples from an existing sample using equidistant strides of an extraction window to extract overlapping subsamples. The proportion of overlap varies according to the window size w , the sample size s and the subsampling factor f . based on these parameters, the stride distance d is defined as

$$d = \left\lfloor \frac{s - w}{f} \right\rfloor,$$

where, $f \in \mathbb{N}_{>0}$.

The clinical parameters for each subsample were copied from the original dataset sample without further adaption. We used a window-size of 2000 timesteps and a subsampling factor of 100. The stride distance varied between 33 and 150, depending on the size of the cropped samples, which were in the range of 2664–5000 time steps.

Categorisation and one-hot encoding: All clinical parameters were one-hot encoded and combined to a 55-valued vector which served as input for the FFN. In a previous step, all parameters based on continuous float values were categorised using the ranges listed in Table 6.

K-fold cross-validation: The models presented in this work were evaluated using k-fold cross-validation (KFCV) (Stone 1974, 1978),

Figure 2: Subsample extraction from 12-lead ECG. An extraction window of size w is slid over the ECG sample of size s in equidistant strides of size d .

Figure 4: Combined model (ECG and clinical parameters). GAP = Global average pooling layer;FC = fully-connected dense layer.

Acknowledgements: This project is supported by Research Campus of Central Hessen (FCMH) via Flexi Funds. The used clinical population is based on a cohort that is part of the Kerckhoff Biomarker Registry (BioReg) that is financially supported by the Kerckhoff Heart Research Institute (KHFI) and the German Center for Cardiovascular Research e.V. (DZHK). The sponsors had no influence on the study design, statistical analyses or draft of the paper. We thank Sabine Hurka for help with data analysis, Andreas Rolf and the

clinical team of the Campus Kerckhoff of the Justus-Liebig-University Gießen for help with acquisition of clinical data. **Author contribution:** All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission. **Research funding:** None declared. **Conflict of interest statement:** The authors are employees either of Technische Hochschule Mittelhessen (N.G., J.H., M.G.) or Justus-Liebig-University Gießen (D.G., T.K.) and

Figure 5: Receiver operating characteristic (ROC) curves of best cross-validated models. Each plot shows the ROC curves of the six best models determined by 6-fold cross-validation on subsample-level (A) and on patient-level (B). Additionally, the summarised ROC (SROC) curves were calculated based on the sensitivities and specificities of the best models using the method proposed by Littenberg and Moses (1993). AUC = area under ROC curve.

have applied for a patent on a method and apparatus for prediction and localisation of structural changes in myocardial tissue (DRN 2020021415065300DE).

Appendix

Formulae

$$\text{BMI} = \frac{\text{weight (kg)}}{\text{height}^2 \text{ (m)}} \quad (1)$$

$$\text{BSA} = \sqrt{\frac{\text{height (cm)} * \text{weight (kg)}}{3600}} \quad (2)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (4)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (5)$$

Table 4: Training configuration.

Parameter/component	Configuration
Keras (Chollet 2015)	version 2.2.4
TensorFlow (Abadi et al. 2015)	version 1.13.0
Optimiser	Adam (Kingma and Ba 2015)
Learning rate	0.001
Loss function	Binary cross-entropy
Activation function	Exponential linear units (Clevert et al. 2016)
Weight initialiser	He uniform (He et al. 2015)
Epochs	150
Cross-validation folds	6
Cross-validation repetitions	50
Convolution filters	64
Convolution kernel size	3
Convolution stride	1
Convolution padding	Same
Max pooling kernel size	2
Max pooling stride	1
Graphical processing units	8× NVIDIA(R) Tesla(R) V100-SXM2 16 GB
Central processing units	4× Intel(R) Xeon(R) E5-2698 (à 20 cores)
Main memory	512 GB

Table 5: Hyperparameter values used for grid search.

Hyperparameter	Values
Number of convolutional layers	2, 3, 4, 5 , 6, 7, 8
Number of convolution filters	16, 32, 64 , 128, 256, 512
Convolution kernel size	3 , 5, 7
Max pooling	no pooling, pooling after each conv. layer
Number of dense layers of the CNN	2, 3 , 4, 5, 6, 7, 8
Number of dense layers for clinical parameters	2, 3, 4 , 5, 6, 7, 8
Number of dense layers after concatenation	2, 3, 4 , 5, 6, 7, 8
Width of dense layers of the CNN	4, 8, 16, 32, 64 , 128, 256
Width of dense layers for clinical parameters	4, 8, 16, 32, 64, 128, 256
Width of dense layers after concatenation	4 , 8, 16, 32, 64, 128, 256
Dropout rate between convolutional layers	0.1 , 0.25, 0.5, 0.75
Dropout rate between dense layers of the CNN	0.1, 0.25 , 0.5, 0.75
Dropout rate between dense layers for clinical parameters	0.1, 0.25, 0.5 , 0.75
Dropout rate between dense layers after concatenation	0.1, 0.25 , 0.5, 0.75

The optimal hyperparameter values marked bold.
CNN = Convolutional neural network.

Table 6: Parameter categorisation based on intervals.

Parameter	Category	Interval
Age	Age#1	[0,45)
	Age#2	[45,55)
	Age#3	[55,65)
	Age#4	[65,75)
	Age#5	[75,85)
	Age#6	[85,∞)
BMI	BMI#1	[0,20)
	BMI#2	[20,25)
	BMI#3	[25,30)
	BMI#4	[30,35)
	BMI#5	[35,40)
	BMI#6	[40,∞)
BSA	BSA#1	[0.0,1.6)
	BSA#2	[1.6,1.8)
	BSA#3	[1.8,2.0)
	BSA#4	[2.0,2.2)
	BSA#5	[2.2,2.4)
	BSA#6	[2.4,∞)

BMI = body-mass-index; BSA = body-surface-area.

Table 7: Comparison of cardiovascular risk factor prevalences in cohort and German population.

Risk factor	Cohort (%)	Germany	
		(%)	Source
Smoking (active)	21	30	Lampert et al. (2013)
Smoking (ex)	42	28	Lampert et al. (2013)
Obesity	22	24	Mensink et al. (2013)
Arterial hypertension	81	32	Neuhauser et al. (2013)
Diabetes mellitus	19	7	Heidemann et al. (2013)
Dyslipidemia	62	19	Scheidt-Nave et al. (2013)

$$NPV = \frac{TN}{TN + FN} \quad (6)$$

$$PPV = \frac{TP}{TP + FP} \quad (7)$$

References

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2015). TensorFlow: large-scale machine learning on heterogeneous systems, Software available from tensorflow.org. <http://tensorflow.org/> (visited on 03/26/2020).
- Acharya, U.R., Fujita, H., Oh, S.L., Hagiwara, Y., Tan, J.H., and Adam, M. (2017). Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals. *Inf. Sci.* 415–416: 190–198.
- Albuquerque, V.H.C., Nunes, T.M., Pereira, D.R., Luz, E.J.D.S., Menotti, D., Papa, J.P., and Tavares, J.M.R.S. (2018). Robust automated cardiac arrhythmia detection in ECG beat signals. *Neural Comput. Appl.* 29: 679–693.
- Altman, D.G. and Royston, P. (2000). What do we mean by validating a prognostic model?. *Stat. Med.* 19: 453–473.
- Andras, P., Esterle, L., Guckert, M., Han, T.A., Lewis, P.R., Milanovic, K., Payne, T., Perret, C., Pitt, J., Powers, S.T., et al. (2018). Trusting intelligent machines: deepening trust within socio-technical systems. *IEEE Technol. Soc. Mag.* 37: 76–83.
- Asch, F.M., Shah, S., Rattin, C., Swaminathan, S., Fuisz, A., and Lindsay, J. (2006). Lack of sensitivity of the electrocardiogram for detection of old myocardial infarction: a cardiac magnetic resonance imaging study. *Am. Heart J.* 152: 742–748.
- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One* 10: 1–46.
- Baloglu, U.B., Talo, M., Yildirim, O., Tan, R.S., and Acharya, U.R. (2019). Classification of myocardial infarction with multi-lead ECG signals and deep CNN. *Pattern Recogn. Lett.* 122: 23–30.
- Bengio, Y. and Grandvalet, Y. (2004). No unbiased estimator of the variance of K-fold cross-validation. *J. Mach. Learn. Res.* 5: 1089–1105.
- Benjamin, E.J., Muntner, P., Alonso, A., Bittencourt, M.S., Callaway, C.W., Carson, A.P., Chamberlain, A.M., Chang, A.R., Cheng, S., Das, S.R., et al. (2019). Heart disease and stroke statistics 2014; 2019 update: a report from the American heart association. *Circulation* 139: e56–e528.
- Bignoto, T.C., Moreira, D.A.R., Habib, R.G., Barros Correia, E.D., Amarante, R.C., Jatene, T., Nunes, M.B.G., Senra, T., and Mastrocolla, L.E. (2018). Electrocardiography scar quantification correlates with scar size of hypertrophic cardiomyopathy seen by multidetector computed tomography. *Clin. Cardiol.* 41: 837–842.
- Bishop, C.M. (1995). *Neural networks for pattern recognition*. Oxford University Press, Inc., USA.
- Bradley, A.P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recogn.* 30: 1145–1159.
- Carpenter, A., Dastidar, A.G., Wilson, C., Rodrigues, J., Baritussio, A., Lawton, C., Palazzuoli, A., Ahmed, N., Townsend, M., Baumbach, A., et al. (2015). 7 Diagnostic accuracy of 12 lead ECG Qwaves as a marker of myocardial scar: validation with CMR. *Heart* 101: A1–A19.
- Chaudhry, U., Platonov, P.G., Jablonowski, R., Couderc, J.-P., Engblom, H., Xia, X., Wieslander, B., Atwater, B.D., Strauss, D.G., van der Pals, J., et al. (2017). Evaluation of the ECG based selvester scoring method to estimate myocardial scar burden and predict clinical outcome in patients with left bundle branch block, with comparison to late gadolinium enhancement CMR imaging. *Ann. Noninvasive Electrocardiol.* 22: e12440.
- Chollet, F. (2015). Keras. Software available from keras.io. <http://keras.io/> (visited on 26/03/2020).
- Clevert, D., Unterthiner, T., and Hochreiter, S. (2016). Fast and accurate deep network learning by exponential linear units (ELUs). In: *4th International conference on learning representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, conference track proceedings*. arXiv, online.
- Dima, S., Panagiotou, C., Mazomenos, E.B., Rosengarten, J.A., Maharatna, K., Gialelis, J.V., Curzen, N., and Morgan, J. (2013).

- On the detection of myocardial scar based on ECG/VCG analysis. *IEEE Trans. Biomed. Eng.* 60: 3399–3409.
- Downey, T.J., Meyer, D.J., Price, R.K., and Spitznagel, E.L. (1999). Using the receiver operating characteristic to assess the performance of neural classifiers. In: *IJCNN'99. International joint conference on neural networks. Proceedings (Cat. No. 99CH36339)*, Vol. 5. IEEE, New York, pp. 3642–3646.
- Dragomiretskiy, K. and Zosso, D. (2014). Variational mode decomposition. *IEEE Trans. Signal Process.* 62: 531–544.
- Dulce, M.C., Duerinckx, A.J., Hartiala, J., Caputo, G.R., O'Sullivan, M., Cheitlin, M.D., and Higgins, C.B. (1993). MR imaging of the myocardium using nonionic contrast medium: signal-intensity changes in patients with subacute myocardial infarction. *Am. J. Roentgenol.* 160: 963–970.
- Goovaerts, G., Padhy, S., Vandenberg, B., Varon, C., Willems, R., and Van Huffel, S. (2019). A machine-learning approach for detection and quantification of QRS fragmentation. *IEEE J. Biomed. Health Inf.* 23: 1980–1989.
- Hanley, J.A. and McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143: 29–36.
- He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: surpassing human-level performance on image net classification. In: *2015 IEEE International conference on computer vision, ICCV 2015, Santiago, Chile, December 7–13, 2015*. IEEE, New York, NY, USA, pp. 1026–1034.
- Health Level Seven, Inc. (2020). HL7 version 3 (schema). <http://www.hl7.org/> (visited on 03/26/2020).
- Heidemann, C., Du, Y., Schubert, I., Rathmann, W., and Scheidt-Nave, C. (2013). Prävalenz und zeitliche Entwicklung des bekannten Diabetes mellitus. *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz* 56: 668–677.
- Hubel, D.H. and Wiesel, T.N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160: 106–154.
- Inoue, Y.Y., Ambale-Venkatesh, B., Mewton, N., Volpe, G.J., Ohyama, Y., Sharma, R.K., Wu, C.O., Liu, C.-Y., Bluemke, D.A., Soliman, E.Z., et al. (2017). Electrocardiographic impact of myocardial diffuse fibrosis and scar: MESA (multi-ethnic study of atherosclerosis). *Radiology* 282: 690–698.
- Isaksson, A., Wallman, M., Göransson, H., and Gustafsson, M.G. (2008). Cross-validation and bootstrapping are unreliable in small sample classification. *Pattern Recogn. Lett.* 29: 1960–1965.
- Kim, R.J., Fieno, D.S., Parrish, T.B., Harris, K., Chen, E.-L., Simonetti, O., Bundy, J., Finn, J.P., Klocke, F.J., and Robert, M.J. (1999). Relationship of MRI delayed contrast enhancement to irreversible injury, infarct age, and contractile function. *Circulation* 100: 1992–2002.
- Kingma, D.P. and Ba, J. (2015). Adam: a method for stochastic optimization. *3rd International conference on learning representations, ICLR 2015, San Diego, CA, USA May 7–9, 2015, conference track proceedings*. arXiv, online.
- Lampert, T., von der Lippe, E., and Müters, S. (2013). Verbreitung des Rauchens in der Erwachsenenbevölkerung in Deutschland. *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz* 56: 802–808.
- LeCun, Y., Boser, B.E., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W.E., and Jackel, L.D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Comput.* 1: 541–551.
- Littenberg, B. and Moses, L.E. (1993). Estimating diagnostic accuracy from multiple conflicting reports: a new metaanalytic method. *Med. Decis. Making* 13: 313–321.
- Liu, W., Zhang, M., Zhang, Y., Liao, Y., Huang, Q., Chang, S., Wang, H., and He, J. (2018). Real-time multilead convolutional neural network for myocardial infarction detection. *IEEE J. Biomed. Health Inf.* 22: 1434–1444.
- Liu, X., Faes, L., Kale, A.U., Wagner, S.K., Fu, D.J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., et al. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit. Health* 1: e271–e297.
- Markendorf, S., Benz, D.C., Messerli, M., Grossmann, M., Giannopoulos, A.A., Patriki, D., Fuchs, T.A., Gräni, C., Pazhenkottil, A.P., Buechel, R.R., et al. (2019). Value of 12-lead electrocardiogram to predict myocardial scar on FDG PET in heart failure patients. *J. Nucl. Cardiol.*, <https://doi.org/10.1007/s12350-019-01841-6>.
- Mazomenos, E.B., Chen, T., Acharyya, A., Bhattacharya, A., Rosengarten, J., and Maharatna, K. (2012). A time-domain morphology and gradient based algorithm for ECG feature extraction. In: *2012 IEEE International conference on industrial technology, IEEE, New York, NY, USA*, pp. 117–122.
- Mensink, G., Schienkiewitz, A., Haftenberger, M., Lampert, T., Ziese, T., and Scheidt-Nave, C. (2013). Übergewicht und Adipositas in Deutschland. *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz* 56: 786–794.
- Metz, C.E. (1978). Basic principles of ROC analysis. *Semin. Nucl. Med.* 8: 283–298.
- Neuhauser, H., Thamm, M., and Ellert, U. (2013). Blutdruck in Deutschland 2008–2011. *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz* 56: 795–801.
- Oikarinen, H., Karttunen, A., Pääkkö, E., and Tervonen, O. (2013). Survey of inappropriate use of magnetic resonance imaging. *Insights Imag.* 4: 729–733.
- Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). Why should I trust you?: explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco, CA, USA August 13–17, 2016*, pp. 1135–1144.
- Rosengarten, J.A., Scott, P.A., Chiu, O.K.H., Shambrook, J.S., Curzen, N.P., and Morgan, J.M. (2013). Can QRS scoring predict left ventricular scar and clinical outcomes? *EP Europace* 15: 1034–1041.
- Saeed, M., Wagner, S., Wendland, M.F., Derugin, N., Finkbeiner, W.E., and Higgins, C.B. (1989). Occlusive and reperfused myocardial infarcts: differentiation with Mn- DPDP-enhanced MR imaging. *Radiology* 172: 59–64.
- Scheidt-Nave, C., Du, Y., Knopf, H., Schienkiewitz, A., Ziese, T., Nowossadeck, E., Gößwald, A., and Busch, M. (2013). Verbreitung von Fettstoffwechselstörungen bei Erwachsenen in Deutschland. *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz* 56: 661–667.
- Selvester, R.H., Wagner, G.S., and Hindman, N.B. (1985). The selvester QRS scoring system for estimating myocardial infarct size: the

- development and application of the system. *Arch. Intern. Med.* 145: 1877–1881.
- Selvester, R.H., Wagner, J.O., and Rubin, H.B. (1971). Quantitation of myocardial infarct size and location by electrocardiogram and vectorcardiogram. In: Snellen, H.A., Hemker, H.C., Hugenholtz, P.G., and Van Bommel, J.H. (Eds.). *Quantitation in cardiology*. Springer Netherlands, Dordrecht, pp. 31–44.
- Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In: *Proceedings of the 34th international conference on machine learning, ICML 2017, Sydney, NSW, Australia 6th to 11th August 2017*, pp. 3145–3153. *Proceedings of Machine Learning Research*, online.
- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Stat. Soc. B* 36: 111–147.
- Stone, M. (1978). Cross-validation: a review. *Series Stat.* 9: 127–139.
- Strodthoff, N. and Strodthoff, C. (2019). Detecting and interpreting myocardial infarction using fully convolutional neural networks. *Physiol. Meas.* 40: 015001, <https://doi.org/10.1088/1361-6579/aaf34d>.
- World Health Organization (2018). *European health report 2018: more than numbers-evidence for all*. WHO Regional Office for Europe, Copenhagen, Denmark.
- Winau, L., Nagel, E., Herrmann, E., and Puntmann, V.O. (2018). Towards the clinical management of cardiac involvement in systemic inflammatory conditions—a central role for CMR. *Curr. Cardiovasc. Imaging Rep.* 11: 11.