Toward Trustworthiness of Deep Learning Models for 12-Lead ECGs

Dissertation
for the award of the degree
"Doctor rerum naturalium" (Dr.rer.nat.)
of the Georg-August-Universität Göttingen

within the doctoral program in Computer Science (PCS)
of the Georg-August University School of Science (GAUSS)

submitted by
Theresa Bender

from Bad Hersfeld
Göttingen, 2023

Thesis Committee

Prof. Dr. Ulrich Sax
Department of Medical Informatics, University Medical Center Göttingen

Prof. Dr. Marcus Baum
Institute of Computer Science, University of Göttingen

Prof. Dr. Dagmar Krefting
Department of Medical Informatics, University Medical Center Göttingen

Members of the Examination Board

Reviewer: Prof. Dr. Ulrich Sax
Department of Medical Informatics, University Medical Center Göttingen

Second Reviewer: Prof. Dr. Ulrich Parlitz
RG Biomedical Physics, Max Planck Institute for Dynamics and Self-Organization

Third Reviewer: Prof. Dr. Thomas Tolxdorff
Institute of Medical Informatics, Charité – Universitätsmedizin Berlin

Further members of the Examination Board:

Prof. Dr. Marcus Baum
Institute of Computer Science, University of Göttingen

Prof. Dr. Dagmar Krefting
Department of Medical Informatics, University Medical Center Göttingen

Prof. Dr. Niels Grabe
Institute of Pathology, University Medical Center Göttingen

PD Dr. Jan Kantelhardt
Institute of Physics, Martin Luther University Halle-Wittenberg

Date of the oral examination: 23.11.2023

Acknowledgements

I'm extremely grateful to my supervisor and chair of committee, Prof. Dr. Ulrich Sax, for his guidance and feedback. I am also thankful for my advisory board, who generously provided knowledge and expertise.

Many thanks to Dr. Nicolai Spicher, for all the fruitful discussions and motivation. Additionally, I would like to thank the Biosignal Processing Group and the Promoclub of the Department of Medical Informatics, for their feedback sessions and moral support.

Lastly, I'd like to thank my friends and family, especially my parents, who kept my spirits and motivation high during this journey.

Abstract

A 12-lead electrocardiogram (ECG), a common examination tool in cardiology, represents the electrical activity of the heart as waveforms. Predictions and classifications with deep learning (DL) algorithms show great potential to aid clinicians in the diagnosis and treatment of patients. However, since clinicians are responsible for the treatment and thus the outcome of single patients, they need to understand the reasoning behind these model's decisions. Important criteria for the acceptance of DL models in clinical settings are covered by aspects of trustworthiness, such as safety and privacy.

In this work, new methods and tools are developed to evaluate and quantify technical aspects of trustworthiness on a pre-trained deep neural network (DNN) for 12-lead ECG classification of six clinically relevant abnormalities. The open source DNN by Ribeiro et al. indicated a good performance on test data and was trained on a large data set. It is systematically analyzed for its reproducibility, explainability, robustness, and generalizability with multiple public and clinical data sets.

For this, F1-scores are calculated and evaluated for different groups, and quantitative measurements for relevance scores of post-hoc explainable artificial intelligence (XAI) methods are analyzed. Moreover, raw ECG data recorded in clinical routine is exported and integrated into the local research infrastructure to evaluate the generalizability of the model in clinical settings.

The results of the DNN with the original test data set can be reproduced with errors in the range of rounding errors. The DNN exhibits similarly high performance on the PTB-XL and CPSC 2018 public data sets, as well as on a large export of resting ECGs from Schiller devices acquired at the University Medical Center Göttingen. Applying XAI to the DNN reveals features similar to cardiological textbook knowledge, such as lead V1

being most important and missing P-waves in atrial fibrillation, and this is validated on all data sets. The noise annotations of PTB-XL are further analyzed regarding their influence on the performance of the pre-trained DNN. The results indicate that the DNN is able to detect atrial fibrillation in 12-lead ECGs with high accuracy, even in the presence of data quality issues, according to human experts. The experiments that concern performance and explainability are repeated on roughly $150,000$ local recordings and yield similar results on these real-world data.

These exemplary analyses of the trustworthiness of a DNN provide promising results and will be further investigated. Considering several aspects of trustworthiness, it is possible to foster trust in DNNs for clinical applications.

Contents

Acknowledgements i

Abstract iii

Contents v

Related Publications ix

Ethics xi

1. Introduction 1

1.1. Research Questions 2

2. Background 5

2.1. Electrocardiogram 5

2.1.1. Atrial Fibrillation 7

2.1.2. Left Bundle Branch Block 8

2.2. Deep Learning 8

2.2.1. DNN by Ribeiro et al. 11

2.3. Trustworthy AI 11

2.4. Statistical Analyses 14

2.5. Data Sets 15

2.5.1. CPSC2018 15

2.5.2. PTB-XL 15

2.5.3. UMG 12-lead ECGs 16

2.5.4. Data Availability 17

3. Reproducibility and Transparency 19

3.1. Methods 19

3.2. Results 21

3.2.1. Methods Reproducibility 21

3.2.2. Results Reproducibility 21

3.3. Discussion 23

4. Explainability 25

4.1. Methods 25

4.1.1. Binned and Average Relevance Scores Over Class 28

4.1.2. Average Relevance Scores Over Class and Lead 29

4.1.3. Entropy Analysis of Relevance Scores Over Class and Lead 29

4.1.4. Average Relevance Scores Over Class, Lead, and Beats 29

4.1.5. Qualitative Analysis of XAI Relevance Scores 30

4.1.6. Comparison Between Databases 30

4.1.7. Comparison Between XAI Methods 31

4.2. Results 31

4.2.1. Average Relevance Scores Over Class 31

4.2.2. Average Relevance Scores Over Class and Lead 32

4.2.3. Entropy Analysis of Relevance Scores 33

4.2.4. Average Relevance Scores Over Class, Lead, and Beats 34

4.2.5. Qualitative Analysis 35

4.2.6. Comparison of Databases 36

4.2.7. Comparison of XAI Methods 38

4.3. Discussion 38

4.3.1. Comparison of XAI Methods 42

4.3.2. Comparison of Databases 43

4.3.3. Artifacts 43

4.3.4. Limitations 43

4.3.5. Key Findings 43

5. Robustness 45

5.1. Methods 45

5.2. Results 47

5.3. Discussion 48

6. Generalizability **49**
6.1. Methods . . . 49
6.2. Results . . . 50
6.2.1. DL Performance . . . 50
6.2.2. DL Explainability . . . 51
6.3. Discussion . . . 53

7. Discussion **55**
7.1. Reproducibility and Transparency . . . 55
7.2. Explainability . . . 56
7.3. Robustness . . . 58
7.4. Generalizability . . . 58
7.5. Conclusion & Outlook . . . 60

List of Figures **61**

List of Tables **63**

List of Listings **65**

Glossary **67**

References (Own Publications) **69**

References **71**

Appendices **81**

A. Article A **83**

B. Article B **91**

C. Article C **105**

D. Abstract D **111**

E. Source Code **115**

Related Publications

Journal Publications

1. **Bender T**, Seidler T, Bengel P, Sax U, Krefting D. Application of Pre-Trained Deep Learning Models for Clinical ECGs. Stud Health Technol Inform 2021, doi: 10.3233/SHTI210539. [1]
2. **Bender T**, Beinecke JM, Krefting D, Muller C, Dathe H, Seidler T, Spicher N. and Hauschild AC. Analysis of a Deep Learning Model for 12-Lead ECG Classification Reveals Learned Features Similar to Diagnostic Criteria. IEEE J Biomed Health Inform 2023, doi: 10.1109/JBHI.2023.3271858. [2]
3. **Bender T**, Gemke P, Idrobo-Avila E, Dathe H, Krefting D, Spicher N. Benchmarking the Impact of Noise on Deep Learning-Based Classification of Atrial Fibrillation in 12-Lead ECG. Stud Health Technol Inform 2023, doi: 10.3233/SHTI230321. [3]
4. Gemke P, **Bender T**, Idrobo-Ávila E, Dathe H, Krefting D, Kacprowski T, Spicher N. Quantifying Alterations Over Time in ST-Segment/T-Wave Amplitudes During Elective Percutaneous Coronary Intervention. 2023 Computing in Cardiology (CinC), Atlanta, GA, USA, 2023, pp. 1-4, doi: 10.22489/CinC.2023.112.

Talks/Poster

1. **Bender T**, Beinecke J, Hauschild AC, Krefting D, Spicher N. Towards Explaining Decisions of a Deep Learning Model for AF Detection in 12-lead ECGs. Joint Annual Conference of the Austrian, German and Swiss Societies for Biomedical Engineering (BMT) 2022, Innsbruck, doi: 10.1515/bmt-2022-2001.

2. **Bender T**, Beinecke J, Hauschild AC, Krefting D, Spicher N. Analyzing a Deep Learning Model for 12-Lead ECG Classification with XAI. 67th Annual Meeting of the German Association of Medical Informatics, Biometry, and Epidemiology e.V. (GMDS) 2022, online, doi: 10.3205/22 gmds050.

3. Spicher N, **Bender T**, Idrobo-Ávila EH, Focke NK, Krefting D. Analyzing EEG networks throughout the lifespan. Third International Summer Institute on Network Physiology (ISINP) 2022, Como.

4. **Bender T**, Beinecke J, Dathe H, Hauschild AC, Krefting D, Spicher N. Opening the black box: Investigating deep learning models for 12-lead ECG classification. Third Infinity 2022, Göttingen.

5. Idrobo-Ávila E, Bognár G, Krefting D, Gemke P, **Bender T**, Kovács P, Spicher N. Quality assessment for multimodal biosignals acquired intraoperatively. 68th Annual Meeting of the German Association of Medical Informatics, Biometry, and Epidemiology e.V. (GMDS) 2023, Heilbronn, doi: 10.3205/23 gmds101.

6. Barth A, **Bender T**, Gemke P, Dathe H, Krefting D, Spicher N. Quantifying Baseline Noise in 12-Lead ECG. 68th Annual Meeting of the German Association of Medical Informatics, Biometry, and Epidemiology e.V. (GMDS) 2023, Heilbronn, doi: 10.3205/23 gmds104.

Ethics

Retrospective experiments with small UMG data set (ECG-AF/LBBB-2021, 29 ECG recordings): Ethik-Kommission der Universitätsmedizin Göttingen, Prof. Dr. Jürgen Brockmöller, vote-no: 29/4/21, 21.04.2021

Retrospective experiments with large UMG data set (ECG-full-2021, $\approx 150,000$ ECG recordings): Ethik-Kommission der Universitätsmedizin Göttingen, Prof. Dr. Jürgen Brockmöller, vote-no: 13/5/23, 24.05.2023

CHAPTER 1

Introduction

An electrocardiogram (ECG) is one of the main diagnostic tools in cardiology. It records changes in heart function and represents the heart's electrical signals as waveforms. They are measured over specific time spans, usually several seconds, but they can also exceed days in cases of intensive care or home monitoring.

A recent survey shows that the most common focus of algorithms for ECG data is arrhythmia detection, due to the fact that it has high case rates and is associated with a high morbidity and mortality [4]. Among the various types of arrhythmia, atrial fibrillation (AF) is the most represented in the literature. Applications range from monitoring single-lead data from wearables [5, 6] to predictions and classifications on multiple leads [7, 8].

The waveform data are usually stored as a series of amplitudes at time points with equal distances (samples). Several standards are in use to store waveform data as well as metadata about the signals and the patient, such as the European Data Format (EDF) [9] and Digital Imaging and Communications in Medicine (DICOM) waveforms [10]. From this, features are extracted for further analysis as clinically meaningful measures. Usually, clinical features of a single or multiple signals are used, sometimes including annotated information on events during the selected time frame [11] or additional clinical data, such as relevant blood values [12].

In recent years, machine learning (ML) algorithms have become more common than traditional feature-based approaches for the classification of and prediction on ECG signals [13]. The automatic adaptation of weights or coefficients through optimization functions allows ML algorithms to not only aid clinicians in decision-making according to predefined rules but also to extract hidden information by finding new relevant patterns in the data [14]. Examples of common algorithms are support vector machine (SVM)s, random forests, and decision tree ensembles. Deep Learning (DL) is a more specific subcategory of ML, and many deep neural network (DNN)s are currently being implemented for ECG analysis. DNNs comprise of big amounts of layers and are thus able to find more features in their input, usually without the requirement for expert knowledge.

While the performance of these automatic classifications and predictions is promising, the medical domain relies heavily on the trustworthiness of the model [15]. Since clinicians are responsible for the treatment and thus the outcome for single patients, they need to understand the reasoning behind each decision they make. This is especially true if the treatment is based on information such as diagnoses suggested by Artificial Intelligence (AI) applications.

Trustworthiness comprises of several requirements such as performance and safety, as well as ethical aspects such as fairness and privacy. Technical requirements according to Li et al. [15] include reproducibility and transparency, robustness against erroneous inputs, and generalization in terms of making accurate predictions on unseen data.

This thesis aims to analyze the trustworthiness of DL algorithms for 12-lead ECG recordings from clinical settings, including analyses on uncurated routine data from the University Medical Center Göttingen (UMG). Analyzing clinical ECGs with DL models has the potential to improve clinical workflows, assess data quality, and open possibilities for finding new parameters. Therefore, it would be highly beneficial to prepare the ground for the use of these methods in clinical settings, starting with medical research.

1.1. Research Questions

This thesis focuses on DL algorithms for 12-lead ECGs. There are many large public data sets available, as well as a large amount of private, local recordings. Figure 1.1 provides an outline of this thesis along the technical requirements of AI trustworthiness. The respective research questions are presented as follows along with detailed explanations.

1. Which factors contribute to reproducible results of DL models?

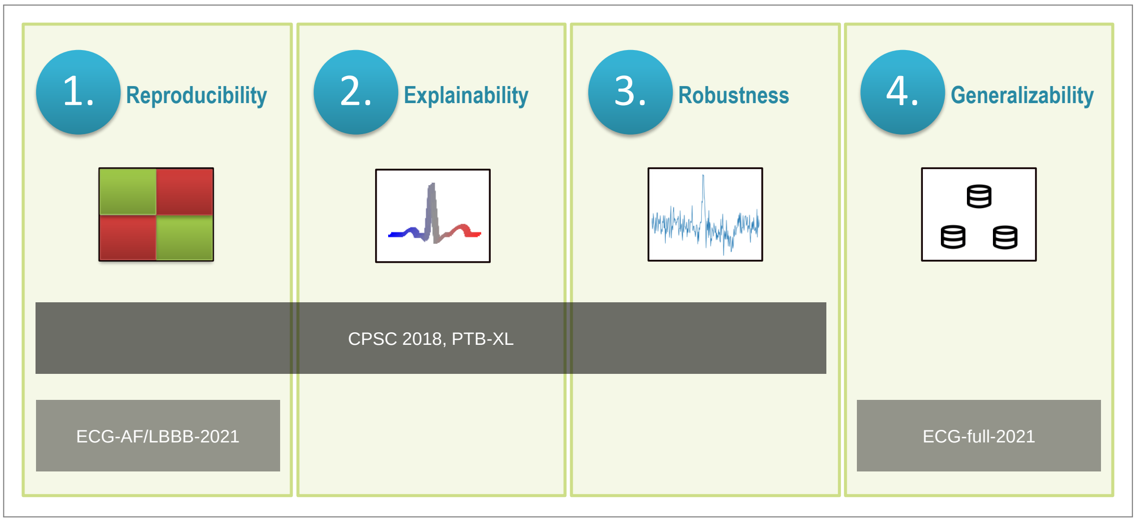


Figure 1.1.: Outline of this thesis investigating the trustworthiness of deep learning algorithms on 12-lead ECG data. In each step, different public (dark gray) and private (light gray) data sets are analyzed.

When applying a DL model to a new data set, the first problem to tackle is whether the performance on each input can be replicated. In ML in general, reproducibility faces specific challenges due to the large number of parameters, and the use of randomness during training [16].

As a first step toward trustworthiness, the reproducibility in terms of methods and results following the definition of Goodman et al. [17] is analyzed for this algorithm.

? 2. How can post-hoc methods objectively explain a model's decisions on ECG data?

Although explainable AI (XAI) has rapidly advanced over the last few years especially for imaging data, but also for text and tabular input, applications for time series data such as biosignals are rare and still require further research [18]. For ECG classification in particular, XAI methods are usually applied qualitatively [19–21].

Commonly used visualizations present individual recordings with corresponding XAI information, such as pseudo-colored overlays. This qualitative evaluation of single records lacks information for characterizing models and their limitations, which would be required for a successful integration of DL algorithms in clinical practice [22]. Thus, this study develops quantitative methods to hopefully offer insights into the possible features used by such models, thus improving their explainability.

? 3. How robust are DL algorithms against typical ECG noise?

Compared with most data sets available, such as those acquired in clinical studies, uncurated real-world data can contain immense noise. Whether introduced by the device, such as through power line interference, or by the patient's behavior, such as motion during the measurement, noise can obscure relevant information in the signals. Starting with a chance of false findings and alarms [23], the signal-to-noise ratio (SNR) of the signal being too low can render diagnostically relevant features undetectable [24]. Many state-of-the-art algorithms for ECG classification extract semantic features derived from human expert knowledge. These algorithms are susceptible to noise, which leads to incorrect results [25]. By contrast, DL algorithms are based on agnostic features, which are derived end-to-end from a fully-automatic correlation analysis between waveform input and respective output classes.

A fundamental assumption in training these models is that training and test data sets stem from the same distribution. This leads to problems in the case of data set shifts, when the performance suddenly becomes worse with data that are acquired on different devices, by other users, or that contain noise. Although initial studies suggest that DL algorithms might be more robust to noise than traditional algorithms [26], the extent of its influence remains unclear and needs to be investigated.

? 4. How generalizable are DL models for 12-lead ECGs on uncurated data?

Data set shifts after training the model on a specific data set as mentioned above are also a problem of generalization [15]. A generalizable model performs similar on unseen data, therefore reducing risks and increasing reliability in real-world scenarios such as clinical routine [15].

After analyzing reproducibility, explainability and robustness mainly on publicly available data sets, generalizability on unseen data can best be tested on uncurated data from clinical routine.

CHAPTER 2

Background

The following sections provide an overview of technical and physiological concepts related to DL and electrocardiography as well as the data sets used in this thesis.

2.1. Electrocardiogram

An ECG measures the electrical activity of the heart. It is a routine procedure in clinical settings, such as cardiology and emergency care. ECG recordings can be acquired in different circumstances, such as while resting or during exercise, and they differ in their length. For example, a standard resting ECG is usually 10 seconds long, while Holter ECGs measure longer periods of approximately 24 hours. The following paragraphs describe several concepts that are frequently discussed in this work.

Cardiac Cycle: A *cardiac cycle* refers to one heartbeat, from its beginning to the start of the next beat. Each cycle contains phases of relaxation and contraction in the right and left atriums and ventricles, respectively. These contractions are a result of the *depolarization* of the heart, which pumps blood throughout the body. It is initiated by the *sinus node* and continues through the HIS bundle and the left and right bundle branches. After *repolarization* of these areas, the heart muscles can relax again. ECG data represents the cardiac cycle in amplitudes (electrical currents) at equally distanced timepoints (*samples*) from multiple directions (*leads*). The association of the phases of the cardiac cycle with the P-, Q-, R-, S- and T-waves of an ECG is depicted in Fig. 2.1.

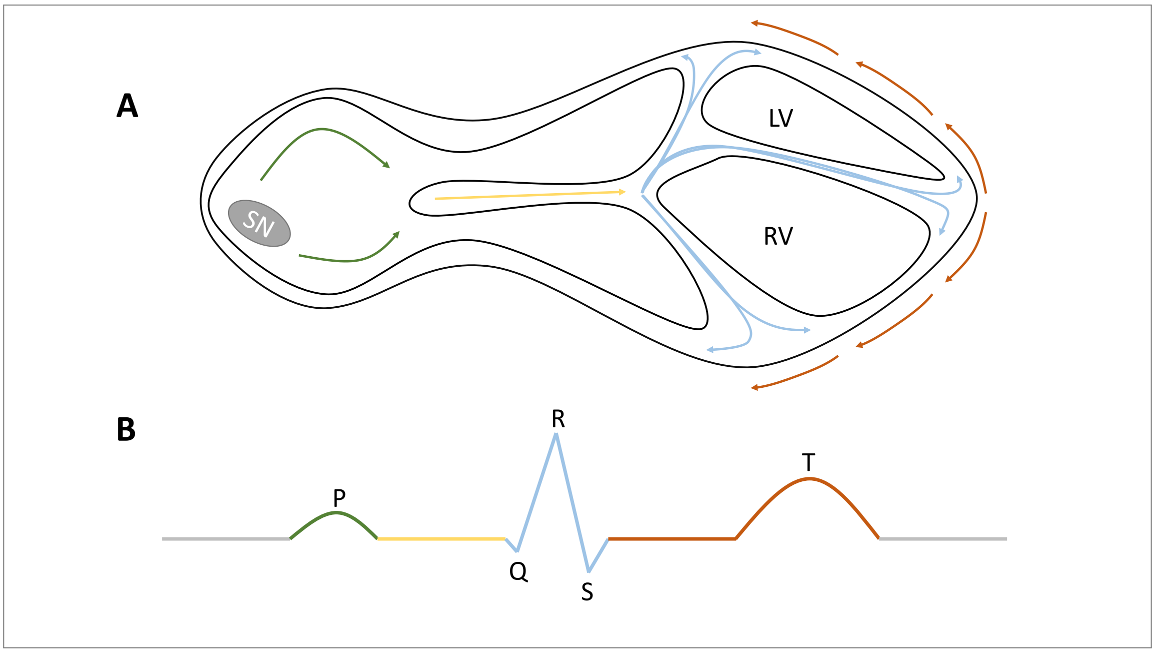


Figure 2.1.: The cardiac cycle and its ECG representation. **A** Schematic representation of a healthy heart. The sinus node initializes electrical impulses from the atrium (green) over the HIS bundle (yellow) through the left and right bundle branches (blue), followed by repolarization (red). **B** ECG representation of a heart beat. Colors correspond to areas in A. SN: sinus node; LV: left ventriculum; RV: right ventriculum. Adapted from [27].

Leads: A lead refers to the polarization of the heart, measured as a potential difference between at least two electrodes. More leads can provide more *views* on the polarization of the heart from a specific direction and thus a more detailed picture of the current state during the cardiac cycle, while fewer leads - down to one for most wearables - are used if only basic features, such as the heart rate, are derived. A standard resting ECG has 12 leads, including six chest leads and three standard and three augmented limb leads. Views of these 12 leads are presented in Fig. 2.2.

Limb leads are derived from electrodes on both arms and the left leg, providing a vertical, frontal view of the heart. The potential differences are measured bipolarly between two limb leads. Standard leads include I, II, and III as defined by Einthoven, and aVR, aVL, and aVF by Goldberger, which are derived from I-III [27].

Additionally, six electrodes are placed on the chest, opening up a horizontal view ”from above”. The corresponding chest leads are then measured unipolarly with each chest electrode against two limb leads as a neutral reference. Wilson defined these leads as V1-V6 from right to left [27].

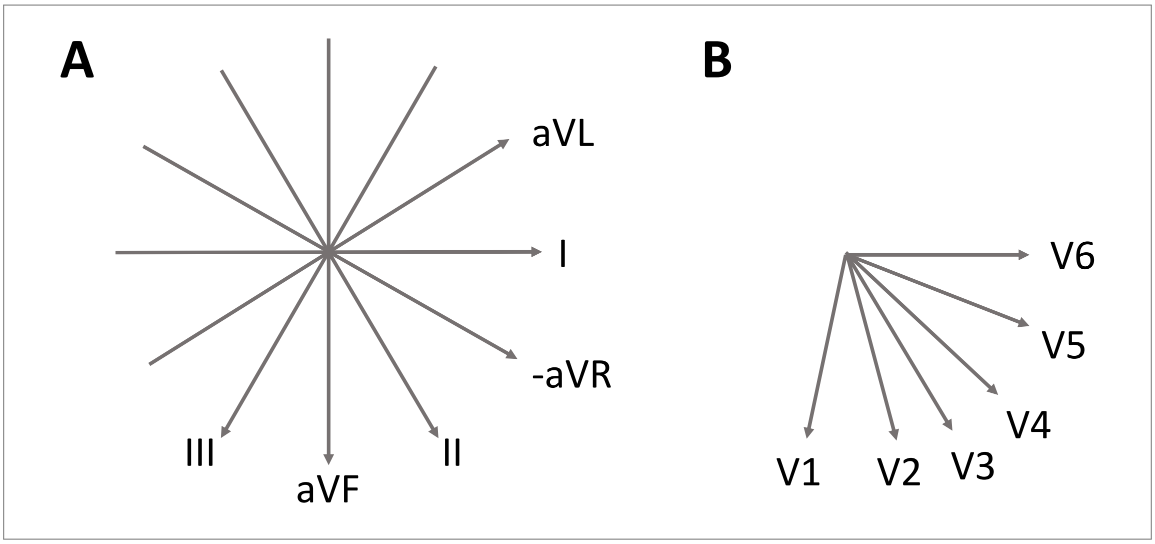


Figure 2.2.: Views on the heart with a 12-lead ECG. **A** Frontal view: Einthoven (I-III) and Goldberger (aVR, aVL, aVF) leads, derived from electrodes on the arms and the left leg. **B** Horizontal view: Wilson leads, derived from electrodes on the chest.

Abnormalities: The standard features extracted from ECGs include lengths and heights between different peaks and waves, which can be used to calculate further parameters such as heartbeat frequency and heart position. If these parameters differ from normative values, it can be interpreted as an abnormality, which in turn substantiates diagnoses. Two abnormalities discussed in this thesis are explained in the following sections. While they can be diagnosed through ECG acquisition with a reduced number of leads, the gold standard for diagnosis is 12-lead ECG [28].

2.1.1. Atrial Fibrillation

AF is a type of arrhythmia based on uncoordinated electrical impulses that originate from the atrium, which result from a non-functioning sinus node [4]. AF can be diagnosed from ECGs, and the criteria for diagnosis are missing P-waves, as they are usually initiated by the sinus node, and absolute irregular RR intervals [4]. Furthermore, repeating fibrillatory waves (f-waves) - uncoordinated electrical impulses in the atrium - mimic P-waves. They can usually be observed best in the chest leads, especially V1 [29].

2.1.2. Left Bundle Branch Block

Lesions in the HIS bundle or its derivatives lead to a downstream block of electrical impulses. When this compromises the left bundle branch, this abnormality is called a left bundle branch block (LBBB). The criteria for the presence of LBBB in ECGs include unusually wide QRS complexes of more than 120 milliseconds, while ST-segment and T-waves point in opposite directions [30]. Broadly notched or slurred R-waves occur in left-sided leads I, aVL, V5, and V6, while Q-waves are absent [30]. Furthermore, a negative terminal deflection in V1, such as an rS-complex with a tiny R-wave and a huge S-wave, is a clear diagnostic marker [31].

2.2. Deep Learning

AI is the broader term for all algorithms that "mimic" cognitive function. A subcategory of AI is ML, which can determine certain patterns from data by adapting weights or coefficients according to the input. Feature extraction with domain knowledge is a prerequisite for this type of AI. DL is a subcategory of ML with complex structures of many layers, allowing to discover patterns even from raw input data such as images without requiring domain knowledge [32].

The relationship of all three categories is pictured in Section 2.3, including another type of categorizing ML algorithms. They can be categorized by their type of training input as follows: data with labels according to the desired output are used in *supervised learning*, whereas data without labels as input are used in *unsupervised learning*. In cases where the algorithm learns by trial and error instead of using labels, this is called *reinforcement learning*. In this work, the focus is on supervised DL algorithms for classification tasks.

Perceptron: An example of an ML implementation is the *neural network*, which models the interconnected neurons of the human brain. The most basic variant of a neural network is the *perceptron* [33], which has only one output neuron and one hidden layer. A schematic overview of a perceptron is provided in Fig. 2.4. A perceptron consists of input neurons x_i with the respective weights w_i of their connections with the next layer, a bias with weight w_B, and an output neuron y. In between, there is a function that connects input and output, usually the sum of all input values and the bias multiplied

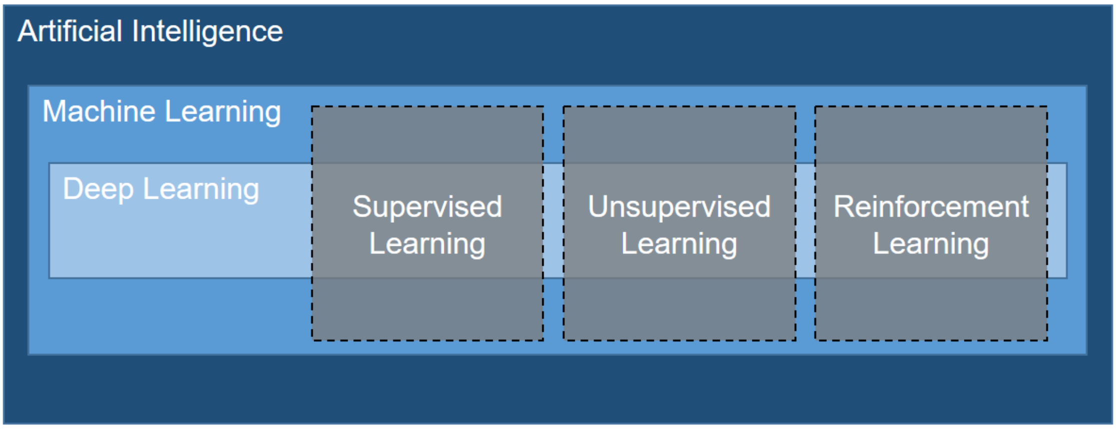


Figure 2.3.: Relationship between artificial intelligence, machine learning (ML), and deep learning (DL). Both ML and DL can be categorized further by type of training input (gray). Adapted from [14], Fig. 2.

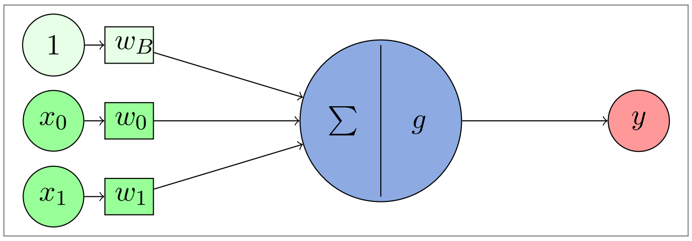


Figure 2.4.: Schema of a perceptron with one hidden layer (blue). The sum of bias 1 and inputs x_i multiplied with respective weights w_i is put into an activation function g, resulting in output y.

with their respective weights, resulting in

$$z = w_B + \sum_{i=1}^{m} x_i w_i, \tag{2.1}$$

where m is the number of input neurons.

This value z is then put into an activation function g [34]. The default activation is *linear*, where each output value is equal to the input ($y = z$). However, this only allows for linear separation of the input. To allow for more complex decisions, nonlinear functions such as *sigmoid* or *rectified linear unit (ReLU)* are frequently used. For example, a sigmoid activation results in the following output:

$$y = g(z) = \frac{1}{1 + e^{-z}}. \tag{2.2}$$

While sigmoid transforms values into the range of $[0, 1]$, ReLU acts as a linear function for positive values and sets all negative values to zero. Each of these and other activation functions has its strengths and weaknesses, therefore different activations can be found in different layers of a single network.

Deep Neural Networks: The more layers are added to a network, the more complex the decision making gets. *Deep neural networks* (DNNs) have at least three hidden layers and are categorized as DL algorithms. Additionally, each layer can have multiple neurons. If every neuron is connected to all neurons of the previous and next layer, the network is *fully connected.* With multiple neurons in the output layer, multi-class decisions can be modeled.

The value of neuron j in layer k of a fully connected network is

$$z_j^k = w_{B,j}^{k-1} + \sum_{i=1}^{m} x_i^{k-1} w_{i,j}^{k-1}, \tag{2.3}$$

with m as the number of neurons in the previous layer *k-1* and x being either the input neuron in case of the first layer, or the output y of the previous activation function.

Training: Neural networks can "learn" the separation of its input by being trained with exemplary data sets. In supervised learning scenarios, this *training data* comes with ground truth labels for the desired classification. Training a network means adapting its parameters, usually the weights of the different connections. While changes in weights alter the steepness in relation to the input values, the activation function can be shifted by changing the bias.

A network is trained in a series of *epochs.* After each epoch, the error between ground truth and model output is calculated. The goal is to optimize this error E by finding a parameter combination of weights $W = w_{i,j}^k$ where

$$W = argmin(E) = argmin(\frac{1}{n}\sum_{i=1}^{n}(y_i - y_{GT})^2), \tag{2.4}$$

with n as the number of output neurons y_i and their respective ground truths y_{GT}.

One can achieve this by adapting the weights through *backpropagation* [32]. A common method for this is *gradient descent*, where for each neuron, starting at the output layer, the steepest gradient for minimizing E is calculated based on the respective parameters, which are then adjusted and the error is propagated backwards to the previous layer.

Architecture: Most DNNs are built from more complex layers than perceptrons. One example that is frequently used for matrices as input, including images and biosignals, is the convolutional neural network (CNN) [32]. It includes layers with convolutional operations, which allow it to learn spatial features through shared weights. These layers are usually combined with ReLU activation functions.

Since gradients can become very low the more layers a network has, it is possible that weights in earlier layers only get slightly adjusted in each training step. This *vanishing gradient* problem can be addressed, for example, with *residual neural networks* (ResNets) [35]. ResNets add *skip connections* between different layers, meaning that the input of the first layer is added to the connected layer's value. This allows one to skip whole layers, especially if they do not contribute significantly to the output result.

2.2.1. DNN by Ribeiro et al.

Currently, the most researched field in algorithms for ECG data is AI, especially DL, which is constantly yielding new algorithms [36]. Ribeiro et al. developed the DNN investigated in this work for the automatic classification of six cardiac disorders, including LBBB and AF. It yields independent probabilities for each of these disorders for an input matrix with exactly 12 leads and $4,096$ samples, expecting a sampling frequency of 400 Hz. Details of the ResNet model built on CNNs can be found in [37]. This DNN was trained on approximately two million data sets acquired in Brazil within a large telemedicine network. The pre-trained model, including its weights, is archived and published along with the test data set [38, 39].

2.3. Trustworthy AI

AI algorithms are usually evaluated on performance-based metrics only, such as accuracy. However, they do not assert factors relevant to many challenges including biases or malicious attacks. Hence, for a trustworthy AI algorithm, further aspects beyond performance need to be considered [15], as depicted in Fig. 2.5.

The main areas described by Li et al. [15] comprise of ethical and technical requirements. Ethical requirements such as privacy and fairness are crucial in medical scenarios, but are beyond the scope of this thesis. Here, the focus is on the technical requirements as described in the following paragraphs.

Reproducibility and Transparency: Good scientific practice is increasingly focusing on open science to allow reproducibility and transparency. For this, trained models and

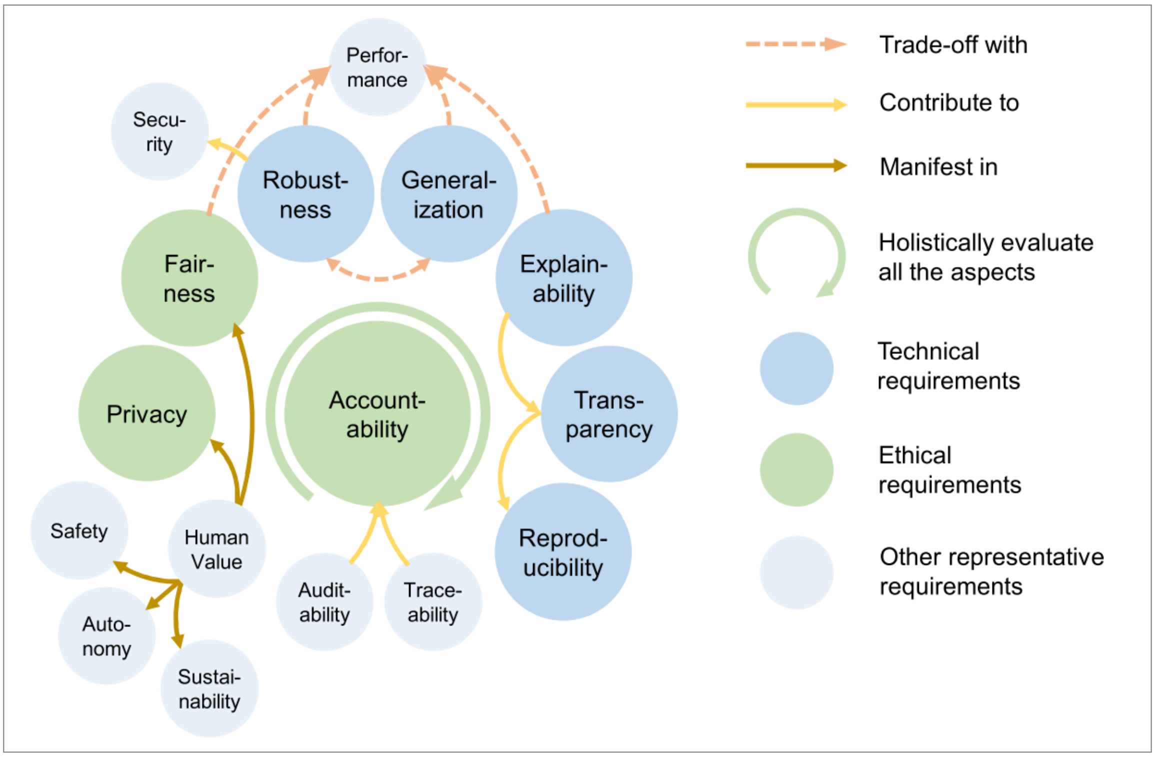


Figure 2.5.: The relation between different aspects of AI trustworthiness. Note that implicit interaction widely exists between aspects, and only representative explicit interactions are covered. From [15], Fig. 1.

test data should be published alongside the description of the model architectures in publications. In this work, both aspects of reproducibility following the definition of Goodman et al. [17] are evaluated:

- *Methods reproducibility:* Will the same data sets result in the same output of the DL model?

- *Results reproducibility:* Will the DL model perform similarly in another environment (other data, other physicians, and another healthcare system)?

Explainability: In this work, explainability of a DNN is evaluated with post-hoc XAI attribution methods Layer-wise Relevance Propagation (LRP) [40] and Integrated Gradients (IG) [41]. Other approaches are available for explaining models for biosignal data using ante-hoc methods, as in [42–44], that can be tailored to a specific model, but they are not suitable for pre-trained DNNs where no adaptation to the model itself is possible. Other methods, such as perturbation methods [45, 46], focus on occluding different parts of images and then analyzing the resulting changes in activations. These methods can also be used to calculate relevance scores for every input feature but, as demonstrated by [47], they produce noisier heatmaps compared with LRP methods.

IG attribute the prediction of a neural network f on unseen data to its input features x, using a baseline input $\tilde{x}$ for attribution calculation. The IG are defined as the path integral of the gradients along the straight-line path from $\tilde{x}$ and x, defined as $\tilde{x}+\alpha(x-\tilde{x})$ for $\alpha \in [0,1]$. The integrated gradient for the i-th input dimension is then defined as

$$\mathrm{IG}_i(x) := (x_i - \tilde{x_i}) \cdot \int_0^1 \frac{\partial f(\tilde{x} + \alpha(x - \tilde{x}))}{\partial x_i} d\alpha, \tag{2.5}$$

where $\frac{\partial f(x)}{\partial x_i}$ is the gradient of output $f(x)$ along the i-th dimension.

LRP tries to explain $f(x)$ by decomposing the output in such a way that

$$f(x) \approx \sum_{d=1}^{V} R_d, \tag{2.6}$$

where V is the input dimension. $R_d > 0$ would then indicate the presence of the structure which is to be classified and $R_d < 0$ would indicate its absence. Propagation of relevance scores works as follows: Let $R_j^{(\ell+1)}$ be a known relevance score of a certain neuron j in the $\ell+1$-th layer of a neural network. The decomposition of the relevance score $R_j^{(\ell+1)}$ in terms of messages $R_{i\leftarrow j}$ sent to neurons of the previous layer ℓ must hold the conservation property

$$\sum_i R_{i\leftarrow j}^{(\ell,\ell+1)} = R_j^{(l+1)}, \tag{2.7}$$

where $\sum_i$ describes the sum over all neurons in the ℓ-th layer of the neural network. Several decomposition rules have been proposed, each having its own advantages and disadvantages. Examples are the introduction of a stabilizer $\epsilon \geq 0$ [40], a different treatment of positive and negative activations with respective weights α and β, where $\alpha + \beta = 1$ [40], and redistributing relevance scores according to the square magnitude of the weights ω [48].

Robustness: DL models for classification tasks have the potential to handle signal noise efficiently due to their data-driven character, but the influence of such noise on the accuracy of these methods is still unclear. Therefore, different types of noise are analyzed for their influence of the accuracy of the model.

Generalizability: To assess the generalizability of a DNN, it should be applied to unseen data. For that purpose, a large uncurated local data set is acquired from the Department of Cardiology at UMG (see 2.5.3).

2.4. Statistical Analyses

The accuracy of a DNN can be described by different statistical measures. First, its classification results can be defined as follows:

- True Positives (TP): Number of samples correctly classified
- False Positives (FP): Number of samples wrongly classified
- True Negatives (TN): Number of samples correctly not classified
- False Negatives (FN): Number of samples wrongly not classified

With these, it can be defined how many classified inputs actually belong to the class with

$$precision = \frac{TP}{TP + FP}, \tag{2.8}$$

and how many members of a class were actually classified with

$$recall = \frac{TP}{TP + FN}. \tag{2.9}$$

A statistical measure for the correct classifications among the total number of classified inputs is

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{2.10}$$

Another evaluation metric frequently used for describing a models performance is the F1-score which can be calculated as harmonic mean of precision and recall with

$$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 * precision * recall}{precision + recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}. \tag{2.11}$$

The complexity of time series can be expressed with *Sample Entropy (SampEn)*, which is the negative natural logarithm of the probability that two sets of data points of lengh m matching pointwise within a tolerance r stay in this tolerance when including the next point, $m + 1$.

2.5. Data Sets

The analyses in this thesis are based on several data sets, either publicly available or unpublished, acquired in clinical routine at UMG. Public data sets can also stem from study contexts or clinical routines, but they are usually curated, as opposed to the uncurated UMG data. All data sets used in this thesis contain labels for the abnormalities AF, LBBB, and right bundle branch block (RBBB). They are described in further detail in the following subsections.

2.5.1. CPSC2018

The CPSC2018 database[1] was acquired in 11 Chinese hospitals containing 12-lead ECGs with a ground truth for one normal type and eight abnormalities provided by human experts [49]. The $9,831$ included recordings of $9,458$ patients have durations of $6-60$ seconds and were sampled at 500 Hz.

The following labels are used for the analysis:

- *Normal* - healthy subjects
- *AF*
- *LBBB*

2.5.2. PTB-XL

PTB-XL [50] is a public database acquired between October 1989 and June 1996 in Germany. The $21,799$ included recordings of $18,869$ patients have a duration of 10 seconds and were sampled at 500 Hz. Furthermore, a version downsampled to 100 Hz is provided for each recording. Additionally, the database contains metadata about each recording, including demographic data of the patient as well as information about the quality of the data, such as noise annotations. Compared with the CPSC database, the subjects' origin is disparate, and there is a chance of different clinical guidelines being in practice for the respective annotations by cardiologists.

The following labels are used for the analysis:

- *NORM* - healthy subjects
- *AFIB* - AF

[1] `https://storage.cloud.google.com/physionet-challenge-2020-12-lead-ecg-public/PhysioNetChallenge2020_Training_CPSC.tar.gz`

- *CLBBB* - (complete) LBBB

PTB-XL includes recordings annotated as incomplete LBBB as well, however, we did not include them in our analyses since the criteria differ slightly from a complete block, with a QRS duration of less than 120 milliseconds.

The following metadata columns are used for noise analyses:

- *baseline_drift*
- *static_noise*
- *burst_noise*
- *electrodes_problems*

In the majority of cases, they contain the name of a single lead (e.g. "aVL"), multiple leads ("I,aVR"), or ranges (e.g. "I-III"). Sometimes, more general labels such as "alles" (all) and "noisy recording" can be found in these columns.

2.5.3. UMG 12-lead ECGs

10-second 12-lead ECGs were recorded at 500 Hz in the Department of Cardiology with Schiller[2] devices and exported in DICOM format. Identifying personal data such as birthyear were removed, and identifiers were either pseudonymized (ECG-AF/LBBB-2021) or deleted (ECG-full-2021).

In contrast to CPSC2018 and PTB-XL, there is no label for healthy patients. Although, for comparison with arrhythmic abnormalities, sinus rhythm (SR) can be found in the data set.

The following labels of the built-in algorithm are used for the analysis of these UMG data sets:

- *VORHOFFLIMMERN* - AF
- *LINKSSCHENKELBLOCK* - LBBB
- *RECHTSSCHENKELBLOCK* - RBBB
- *SINUSRHYTHMUS* - SR

[2] https://www.schillermed.de/

ECG-AF/LBBB-2021 This smaller data set was selected by cardiologists from PDF printouts in health records fo 35 patients in total. 6 ECGs were removed, as they were no longer available in a digital format. The data set contains 19 recordings of patients with diagnosed AF and 17 recordings of patients diagnosed with LBBB.

ECG-full-2021 In 2021, a full export of all ECG recordings acquired with Schiller devices in the Department of Cardiology was conducted in DICOM format. A total of $146,505$ recordings are available, acquired in the time periods 07/2006 – 02/2008 and 01/2018 - 11/2021.

2.5.4. Data Availability

On the path to getting AI methods into clinical settings, the secondary use of clinical data in research contexts is becoming increasingly necessary [51]. Nevertheless, many devices used at the UMG still use proprietary formats to store the data. Therefore, as part of this thesis, the data has been made available for secondary research.

12-lead ECGs at UMG are by default only available in a proprietary format stored on Schiller servers, with visualizations and analyses by clinicians conducted on workstations (SEMA) provided by the manufacturer. Through SEMA, single recordings could already be exported in DICOM format, as tested for ECG-AF/LBBB-2021.

DICOM waveforms [10], developed in 2000, have been adopted by an increasing number of manufacturers of ECG hardware over the last few years. They are an open standard and can store samples of a biosignal, including metadata about the acquisition and the signal itself, as well as annotations for one or more channels. Next to Health Level 7 (HL7) Annotated Electrocardiogram (aECG), for example, DICOM is a common export format for ECG devices [52].

With an automated, regular export in mind, the manufacturer was contacted by members of the Department of Cardiology to explore further possibilities for the SEMA exports. With the acquired information, a full DICOM export can be triggered, which is imported into the medical data integration center (MeDIC) infrastructure.

This export is integrated as a new data source in the MeDIC infrastructure [53] to enable data usage applications for all researchers, prospectively combined with other data acquired at UMG. Furthermore, for easier access and analysis of the data after successful application, XNAT[3] is chosen as an integrated platform for biosignal and other clinical data. Similar to DICOM, which this platform supports as an import format, it was originally built for imaging data but has also been extended for biosignals [54].

[3] https://www.xnat.org/

The export of approximately $150,000$ ECG recordings was successfully integrated into the MeDIC infrastructure and is now available for data usage applications. Anonymization procedures are already in place, while the pseudonymization process for integration with other clinical data is still ongoing.

Furthermore, a local XNAT instance is running on secured servers, filled with public and cohort study data sets, and ready to import the UMG DICOM files.

CHAPTER 3

Reproducibility and Transparency

> *This chapter is based on the conference publication Application of Pre-Trained Deep Learning Models for Clinical ECGs [1] (see Appendix A).*

As discussed previously, current trends for ECG analyses are in DL, where DNNs currently dominate the high ranks of many challenges in medical classification tasks. Recently, 12-lead ECG recordings were employed to build a DNN to detect ECG abnormalities [37]. The authors stated that the model outperformed resident medical doctors. However, the question remains of how these results translate to other countries and to other settings, such as inpatient care.

The aforementioned DNN for the automatic detection of certain cardiovascular diseases on 12-lead ECG data is applied to pseudonymized ECG recordings from inpatient care. This study evaluates the added value of such a DL model by Ribeiro et al. (see Subsection 2.2.1) compared with existing built-in analysis with respect to clinical relevance and analyzes its methods and results reproducibility [17].

3.1. Methods

For *method reproducibility*, the original paper, the Zenodo repository, and the corresponding source-code repository [1] were checked for metadata on the runtime environ-

[1] https://github.com/antonior92/automatic-ecg-diagnosis

```python
zero_padding = np.zeros((96, 12))
firstiteration = True
for d in ds:
    # correct unit (μV->mV) and sampling frequency (500Hz to 400Hz)
    resampled = signal.resample_poly(d.waveform_array(0), 4, 5)/1000
    concat = np.concatenate([resampled, zero_padding], axis=0)
    if firstiteration:
      x = np.expand_dims(concat, axis=0) # (1,4096,12) matrix
      firstiteration = False
    else:
      x = np.concatenate([x, np.expand_dims(concat, axis=0)], axis = 0)
```

Listing 3.1: Preprocessing of ECG-AF/LBBB-2021. Each recording d in the loaded data set ds is resampled and zero-padded. Afterwards, x contains all preprocessed recordings, according to the format which is required by the used model.

ment settings. The model is implemented in the university's JupyterHub and executed on the provided test data. As the model gives probabilities for each class, the thresholds used are required to reproduce the results and evaluate possible differences. As they are not explicitly provided in the paper, threshold values found in the current version of the code are used (*generate_figures_and_tables.py*), with 0.124, 0.07, 0.05, 0.278, 0.39, and 0.174 for first-degree atrioventricular block (1dAVb), RBBB, LBBB, sinus bradycardia, AF, and sinus tachycardia, respectively.

For assessing the *results reproducibility*, the UMG data set ECG-AF/LBBB-2021 (see Subsection 2.5.3) is loaded with the program library pydicom[2] (version 2.1.2). Preprocessing includes resampling from 500 to 400 Hz and zero-padding to 4,096 samples, as shown in Listing 3.1.

The model results are compared with the built-in analysis as well as the clinical diagnosis. The model's performance is assessed by sensitivity (recall) and specificity, precision, and F1-score, following the original publication of the DNN. Here, the recordings that indicate the respective other disorder are used as negative samples. Finally, diverging results in either the DNN classification, built-in classification, or diagnosis are evaluated by a cardiologist for clinical soundness and relevance.

[2] https://pydicom.github.io

3.2. Results

3.2.1. Methods Reproducibility

No metadata description is provided in the journal article [37], but it refers to the open source repository on GitHub. In this repository, library versions are given in a specific requirements file that contains all Python libraries used, although the correct version for the training of the model is not indicated.

No information about the employed operating system or hardware environment could be discovered. Thus, some confusion exists regarding the employed TensorFlow version. In the environment used in this study, which uses the latest versions by default, essentially all Python libraries have been updated since the original publication of the model.

Applying the model to the original test data of 827 ECG recordings produces similar but not equal results. The comparison with the reported abnormality probabilities by the authors reveals differences in approximately 88% of the values (4381). Noteworthily, a switch from TensorFlow 2.3.1 to 2.2 results in one more value differing. However, differences are in the order of rounding errors in floating point values (i.e., $\approx$ 1e$-$7). When compared with class thresholds, none of these differences result in a different classification.

3.2.2. Results Reproducibility

For the 19 AF recordings of the ECG-AF/LBBB-2021 data set, 17 are classified correctly with AF with probabilities significantly above the classification threshold. Two recordings are not classified, as their probabilities are below 0.2 and thus not close to the threshold. An overview is provided in Fig. 3.1.

Three of the AF recordings are additionally classified as RBBB with a probability of almost 0.8. 1dAVb is detected in seven of the LBBB patients, with a minimum probability of 0.27, which is considerably higher than the threshold of 0.124. These findings are confirmed by experts. One LBBB patient is also clearly classified as AF, with a probability of ≈ 0.6, and identified as a false positive.

For LBBB, the ”perfect detection” with an F1-score of 1 could be reproduced, and for AF, the F1-score is 0.919, which is even higher than the published score of 0.870. A summary of all performance scores is provided in Table 3.1.

Overall, the model performance in terms of F1-score for the initial 29 recordings is similar to or higher than that for the original test data, while the built-in algorithm exhibits higher scores for AF but lower ones for LBBB.

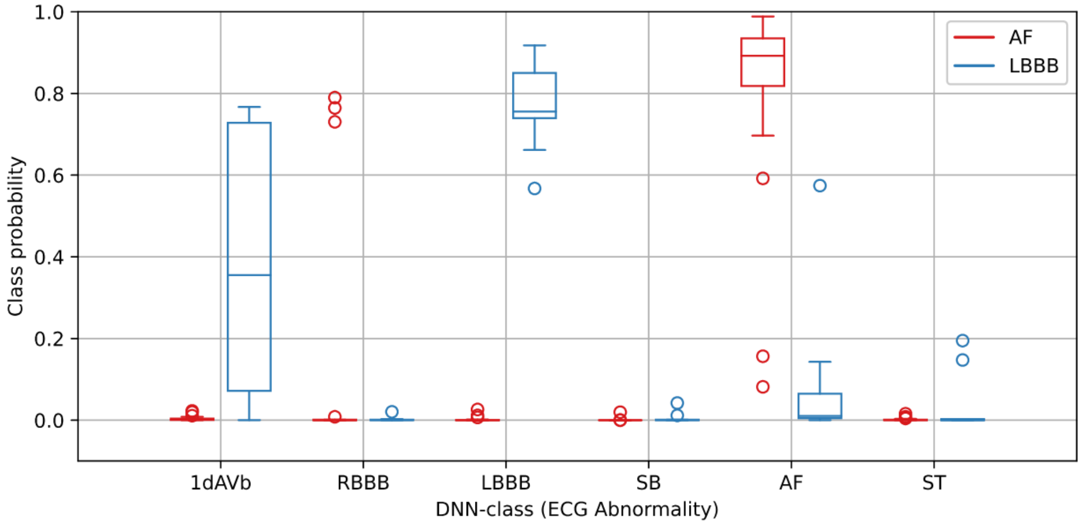


Figure 3.1.: DNN classification results on ECG-AF/LBBB-2021. Separated into cohorts annotated with either AF (red, n=10,001) or LBBB (blue, n=3,495) as annotated by the device's built-in algorithm. Considered abnormalities are: 1st degree AV block (1dAVb), right bundle branch block (RBBB), left bundle branch block (LBBB), sinus bradycardia (SB), atrial fibrillation (AF) and sinus tachycardia (ST). [1]

Table 3.1.: Comparison of DL model results including F1-scores. Calculated for both LBBB and AF. Considered analyses: built-in algorithm of ECG devices as well as DNN classification on the same data set, in addition to scores published by Ribeiro et al. for original test data [37]. [1]

		LBBB	**AF**
Built-in	**P**	10	19
	FP	0	0
	FN	1	0
	F1-Score	0.947	1.000
DNN (local data set)	**P**	10	19
	FP	0	1
	FN	0	2
	F1-Score	1.000	0.919
DNN (Ribeiro et al.)	**P**	30	13
	FP	0	0
	FN	0	3
	F1-Score	1.000	0.870

3.3. Discussion

The DL method proposed by Ribeiro et al. cannot be fully reproduced numerically; differences in the class probabilities are found in the order of rounding errors. Different rounding methods are known to affect the reproducibility of numerical methods, when mathematical system libraries are used rather than static libraries [55]. Interestingly, different TensorFlow versions produce slightly different results in the otherwise identical runtime environment, which might be due to different mathematical optimization procedures. These issues can be avoided through a container-based provision of the method, or at least a full description of the meta data [56, 57]. Noteworthily, much crucial information was already provided in the source code by the authors; only information about the Python version and the used operating system would have been required for full method reproducibility. However, in ECG-AF/LBBB-2021, the predicted class probability is always clearly above or below the threshold; therefore, the numerical differences do not have any influence on the classification results. The results reproducibility for this data set is excellent. The performance parameters can be reproduced with ECG-AF/LBBB-2021, although they must be downsampled, padded, and rescaled.

While the DL model does not perform better on ECG-AF/LBBB-2021 than the built-in method, due to its free availability and applicability to data from different devices, it definitely provides added value, at least for multi-center studies where heterogeneous ECGs are required to be analyzed consistently. Furthermore, due to its good performance, the model can be implemented in self-training modules on ECG analysis for medical students. Added value to the clinical routine is not particularly clear, as the built-in method is comparable while offering more classes (e.g., left anterior fascicular block) or annotations about changes caused by ischemia as well as previous infarcts.

The limitations of this pilot study include a relatively low number of samples and missing healthy controls. It should be noted that the built-in method might have a high sensitivity for AF according to clinicians using such devices; therefore, it might be a bias that all selected ECG recordings are also annotated accordingly by the built-in method.

CHAPTER 4

Explainability

> *This chapter is based on the journal publication Analysis of a Deep Learning Model for 12-Lead ECG Classification Reveals Learned Features Similar to Diagnostic Criteria [2] (see Appendix B) and the conference abstract Towards Explaining Decisions of a Deep Learning Model for AF Detection in 12-lead ECGs (see Appendix D).*

After analyzing the reproducibility of the DNN by Ribeiro et al., the explainability of this model is examined with post-hoc XAI methods. In the following, quantitative methods for the analysis of the XAI output are defined as a first attempt to open the black box of the Ribeiro model.

4.1. Methods

Explainable attribution methods are applied to a pre-trained DNN for abnormality classification in 12-lead electrocardiography (see Subsection 2.2.1) to understand the relationship between model prediction and learned features. Data is classified from two public databases, first CPSC2018, then PTB-XL to check for validation of the proposed methods (cf. Section 2.5). Preprocessing is similar to Listing 3.1, although the loading of the data set and the length of recordings differ for each source. The algorithm for CPSC data is given exemplarily in Listing 4.1.

```
    header_files = get_headerfiles(input_dir)
    num_files = len(header_files)
    firstiteration = True
    for i in range(num_files):
      if(i in idx_x):
        recording, header = load_challenge_data(header_files[i])s
        # correct input shape to sample,lead
        recording = np.swapaxes(recording,0,1)
        # correct sampling frequency from 500Hz to 400Hz
        resampled = signal.resample_poly(recording, 4, 5)/1000
        if(len(resampled) < 4096):
          zero_padding = np.zeros((4096-len(resampled), 12))
          resampled = np.concatenate([resampled, zero_padding], axis=0)
        if firstiteration:
          x = np.expand_dims(resampled[0:4096], axis=0)
          firstiteration = False
        else:
          x = np.concatenate([x, np.expand_dims(resampled[0:4096], axis
  =0)], axis = 0)
```

Listing 4.1: Preprocessing of CPSC data. Each recording in the index list *idx_x* is reshaped, resampled and zero-padded, and returned in variable x.

Regarding data processing[1], each ECG signal is fed into the Ribeiro model for classification, which results in a matrix with dimensions $N \times 6$, assigning probabilities for six ECG abnormalities. In the following, the classification probability is defined as $\{C_n \in \mathbb{R} \mid 0 \leq C_n \leq 1\}$, which indicates the prediction score of the model with sigmoid activation. The iNNvestigate package [58], which implements multiple XAI methods, is used to compute relevance scores for each sample of the input ECG.

To use this library on the Ribeiro model, the sigmoid activation of the model's last layer must be replaced, as presented in Listing 4.2. While sigmoid activation does not change the rank order of the predicted classes, it might obfuscate the true confidence of the model's individual class predictions[2]. Thus, it is replaced with a linear activation.

In this study, the XAI methods IG and LRP are used, with the IG implementation having a baseline input of zero and an interval size of $m = 64$. These attribution methods assign a *relevance score* R_n to each sample n of the classified signals. The computation of this relevance score with iNNvestigate is presented in Listing 4.3.

Relevance scores allow to analyze what the DNN learned during training, with three quantitative methods, namely average relevance scores over (a) classes, (b) leads, and (c) average beats. Additionally, qualitative analyses and a comparison of XAI methods

[1] All computations are implemented using Python v3.6.8 and the libraries iNNvestigate v1.0.9, Tensorflow v1.12.0, neurokit2 V0.1.7, and h5py v2.10.0.

[2] `https://github.com/albermax/innvestigate/issues/84`, accessed: October 14, 2022

```python
import keras
import copy

def loadmodel(input_dir):
  # load Ribeiro model - https://zenodo.org/record/3765717
  model = keras.models.load_model(input_dir + ”model.hdf5”, compile=
False)
  for l in model.layers:
    l.name = ”%s_workaround” % l.name
  # create model with new names
  model = keras.models.Model(input=model.input, output=model.output)
  model.compile(loss='binary_crossentropy', optimizer=keras.optimizers.
Adam())
  orig_model = copy.copy(model)
  # replace sigmoid activation with linear
  # (s. https://github.com/keras-team/keras/issues/7190)
  bn_layer = model.get_layer(index=-1)
  bn_layer.activation = keras.activations.get('linear')
  bn_prev_layer = model.get_layer(index=-2)
  bn_output = bn_layer(bn_prev_layer.output)
  bn_model = keras.Model(inputs=model.inputs, outputs=bn_output)
  return orig_model, bn_model
```

Listing 4.2: Loading the Ribeiro model for classification of six ECG abnormalities. The function returns the original model, as well as a copy with a linear instead of sigmoid activation in the last layer.

```python
import numpy as np
import innvestigate
import innvestigate.utils as iutils

def analyze(x, model, bn_model, method, neuron):
  relevances = np.empty((len(x),4096,12))
  relevances_l = np.empty((len(x),4096,12))
  # create XAI analyzers with both models, output neuron (class) will be selected by index
  analyzer = innvestigate.create_analyzer(method, model, neuron_selection_mode='index')
  analyzer_l = innvestigate.create_analyzer(method, bn_model, neuron_selection_mode='index')
  # loop through dataset to avoid running out of memory
  for aidx, dataset in enumerate(x):
    # analyzer expects same input as model: (n,4096,12)
    a = analyzer.analyze(np.expand_dims(dataset, axis=0), neuron_selection=neuron) # neuron: 2 for LBBB or 4 for AF
    relevances[aidx] = a[0]
    a2 = analyzer_l.analyze(np.expand_dims(dataset, axis=0), neuron_selection=neuron)
    relevances_l[aidx] = a2[0]
  return relevances, relevances_l
```

Listing 4.3: Calculating relevance scores of data set x with both models returned by *loadmodel()* (Listing 4.2). XAI method and output neuron (class) are given as additional variables.

are performed. An overview of the workflow is provided in Fig. 4.1. All analyses are described in further detail in the following subsections.

4.1.1. Binned and Average Relevance Scores Over Class

First, relevance scores for all 200 normal, 200 LBBB, and 200 AF recordings are analyzed separately and bin the values for their respective classes. This allows us to compare the overall distribution of $R_{n,j,k}$ for the different classes.

All leads of each recording n are then aggregated into

$$M_n := \frac{1}{J\ K} \sum_{k=0}^{K-1} \sum_{j=0}^{J-1} R_{n,j,k}, \tag{4.1}$$

where $K = 12$ as number of leads and $J = 4,096$ as number of samples. $R_{n,j,k}$ takes positive or negative values; hence, a higher M_n is associated with a higher prediction score, which is termed *completeness* in [59]. Here, the prediction score is the output of the model with linear activation.

```python
import numpy as np
import antropy as ant
from scipy import signal

for j in range(12):
  a = np.swapaxes(relevances[i],0,1)[j]
  a_result = ant.sample_entropy(signal.detrend(a))
  patient.append(a_result)
```

Listing 4.4: Calculating the sample entropy of IG *relevances* for each recording i.

4.1.2. Average Relevance Scores Over Class and Lead

The relevance scores for each lead k and recording n are aggregated in

$$M_{n,k} := \frac{1}{J}\sum_{j=0}^{J-1} R_{n,j,k}, \tag{4.2}$$

where $J = 4,096$. This allows for a comparison of the distribution of $R_{n,j,k}$ with regard to class and ECG leads, and thus, of the importance of the individual ECG leads for the DNN. This is required as the different leads exhibit different morphologies and signal shapes that might cancel out in the first analysis.

4.1.3. Entropy Analysis of Relevance Scores Over Class and Lead

For a quantitative analysis based on complexity, the sample entropy (*SampEn*, m = 2, r = 0.2 std, N = 4,096) of the IG relevance scores for AF classification is calculated with regard to lead and label and aggregate the results as boxplots. The corresponding Python code is presented in Listing 4.4.

4.1.4. Average Relevance Scores Over Class, Lead, and Beats

In the first proposed quantitative analysis methods, time information is lost. However, for explaining the DNN's decision, this information is crucial as one needs to compare whether the agnostic features trained by the DNN reflect clinical features, such as missing P-waves and unusually wide QRS-complexes.

Analyzing individual ECG recordings provides only anecdotal evidence. Therefore, a two-step averaging procedure is performed that averages the information over several recordings while preserving time information. First, for each ECG record and lead, with the concept of ”average beats” [60] the whole signal is split into individual heart beats with the *ecg_segment()* function of neurokit2. They are averaged into a single,

```python
import numpy as np
import neurokit2 as nk

ecg = np.swapaxes(raw_data,1,2)[patient][lead]
ecg_cleaned = nk.ecg_clean(ecg, sampling_rate=400)
segments = nk.epochs_to_df(nk.ecg_segment(ecg_cleaned, rpeaks=None, sampling_rate=400, show=False))
result = segments.groupby('Label').agg({'Index': ['min', 'max']})
beats = len(result["Index"])
duration = int(len(segments)/beats)
avg_beat = np.zeros(duration)
avg_rel = np.zeros(duration)
# sum values of all beats
for idx in range(beats):
    start = result["Index"]["min"][idx]
    end = result["Index"]["max"][idx]
    avg_beat += raw_data[patient][start:end+1, lead]
    avg_rel += relevances[patient][start:end+1, lead]
```

Listing 4.5: Calculation of the average beat with average relevance scores for a specific *lead* of a single ECG recording (*raw_data*). The first and last beat must further be zero-padded if they would overlap with the full *duration*.

time-aligned representative beat for each lead. Then, with the exact same indices of the heartbeats, the same steps are performed on the relevance scores $R_{n,j,k}$. This yields an *average relevance score*, as presented in Listing 4.5. All average beats and relevance scores are then averaged for a given class. All segments are of equal size for one recording; hence, segments that overlap at the start or end of the recording are filled with zeros. Finally, amplitudes are normalized to $[-1, 1]$. For scatter plot visualizations, relevance scores are upsampled by a factor of 5.

4.1.5. Qualitative Analysis of XAI Relevance Scores

The results of all processed ECG signals are visualized as heatmap-colored scatter plots for each lead, after a normalization of the output to $[-1, 1]$, keeping the center of the values at zero. Furthermore, these relevance score plots are evaluated by an experienced cardiologist.

4.1.6. Comparison Between Databases

To evaluate the generalizability of the presented processing pipeline, the results are evaluated on another, distinct publicly available data set, namely PTB-XL [50] (see Subsection 2.5.3).

4.1.7. Comparison Between XAI Methods

Since both the IG and LRP methods differ substantially in their approaches to calculating relevance scores for the input, using both methods should help to uncover critical information about why the DNN has made certain decisions. Hence, IG results are compared with LRP results using the following LRP decomposition rules implemented in iNNvestigate [58]:

a) The ϵ-LRP decomposition with $\epsilon = 1e - 07$;

b) The $\alpha\beta$-LRP decomposition with $\alpha = 1$ and $\beta = 0$;

c) The ω^2-LRP decomposition;

d) The combination of $\alpha\beta$-LRP decomposition with $\alpha = 1$ and $\beta = 0$ for convolutional layers and ϵ-LRP decomposition with $\epsilon = 0.1$ for fully connected layers.

The sigmoid function (used in the output layer) maps from $\mathbb{R}$ to $\mathbb{R}^+$ and thus inverts the signs of all negative values, as well as scales all values into the interval $[0, 1]$. This results in only small and positive values being backpropagated by LRP, possibly resulting in small and only positive relevance scores. Thus, these relevance scores are compared with those obtained using a linear output in the last layer. Since both activations yield similar results when compared visually in heatmaps, linear activation is used in the analyses, to avoid a possible sign flip.

4.2. Results

After processing recordings with the DNN, C_n is the probability that a recording n exhibits the interrogated abnormality. Applying an XAI method results in a relevance score $R_{n,j,k} \in \mathbb{R}$ for each input sample of a classified ECG with $j = \{0, 1, \ldots 4,095\}$ representing the sample index, and $k = \{0, 1, \ldots 11\}$ representing the lead.

4.2.1. Average Relevance Scores Over Class

The mean of the distributions of IG relevance scores $R_{n,j,k}$ for each class (Fig. 4.2) is close to zero. This indicates that the majority of ECG samples are not relevant for the DNN's decision. The distributions for both abnormalities are almost similar to normal recordings, although they are slightly broader and shifted to positive values. For LBBB, in the range of $[0.0, 0.10]$, there are a larger number of more positive relevance scores compared with normal recordings (Fig. 4.2b).

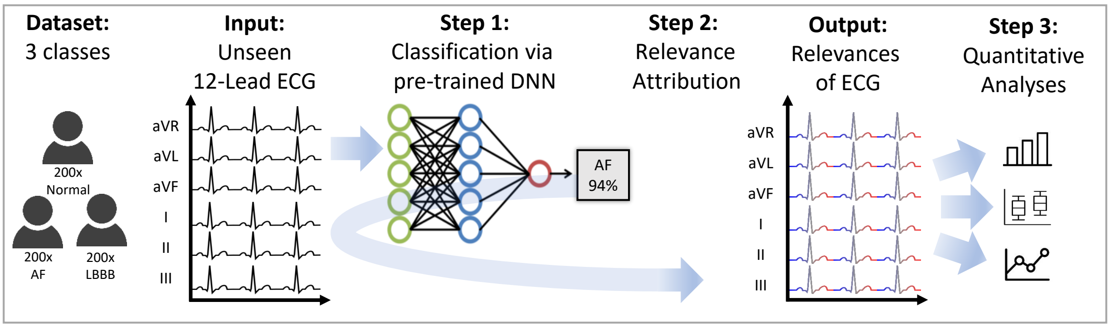
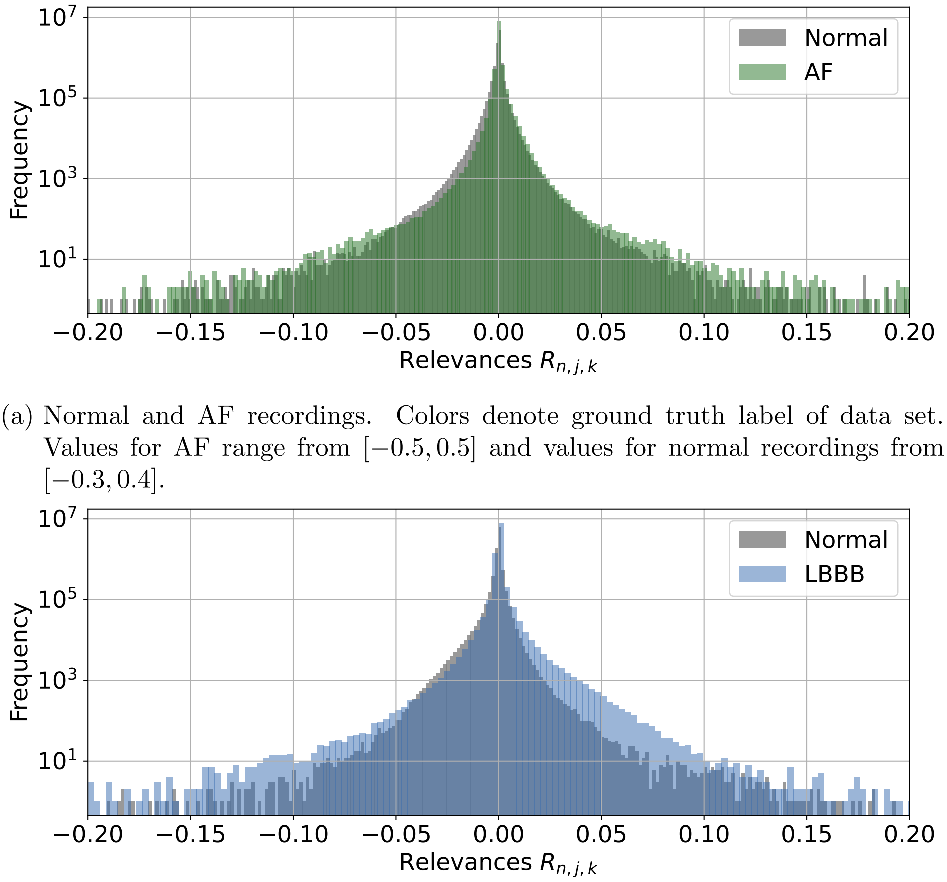


Figure 4.1.: Overview of the processing pipeline which is applied separately to data stemming from two different databases (CPSC/PTB-XL): For each database, the data set consists of 200 healthy controls (Normal) that are compared to patients showing AF and LBBB. Each (unseen) 12-lead ECG is fed into the pre-trained DNN and subsequently results are explored with the XAI methods, yielding a relevance score for each input sample, indicated here by blue (negative relevance score), grey (neutral), and red values (positive relevance score). Novel analysis methods are proposed for these scores, allowing to gain insight into the DNN's reasoning. [3] © 2023 IEEE

The aggregated relevance scores of individual recordings are again centered close to zero and rather equally distributed (Fig. 4.3). In general, LBBB exhibits larger values in positive and negative directions compared with AF. While the median value is always very close to zero, the mean value of relevance scores increases with increasing C_n. For AF classification (Fig. 4.3a), a large number of normal recordings correctly classified as not showing AF have a C_n near 0 while correctly classified AF recordings are near 1. In between is a transition area with nine false negative classifications in the range of $[0.1, 0.39[$. The remaining seven false negatives exhibit M_n values close to zero. LBBB has similar properties to AF, although there is no visible transition area and the values are not as close to 1 (Fig. 4.3b).

4.2.2. Average Relevance Scores Over Class and Lead

When the model results of each lead k for AF classification are analyzed (Fig. 4.4a), the mean relevance scores have medians of 0.0002 and -0.0001 and are in the range of $[-0.0002, 0.0010]$ and $[-0.0014, 0.0012]$ for AF and normal recordings, respectively. For LBBB classification (Fig. 4.4b), the medians are 0.0001 and -0.0002 and the ranges are $[-0.0008,\ 0.0016]$ and $[-0.0009, 0.0022]$ for LBBB and normal recordings, respectively. For each lead, the mean relevance scores are higher for both abnormalities compared with normal recordings, with lead V1 exhibiting the highest difference in median values.



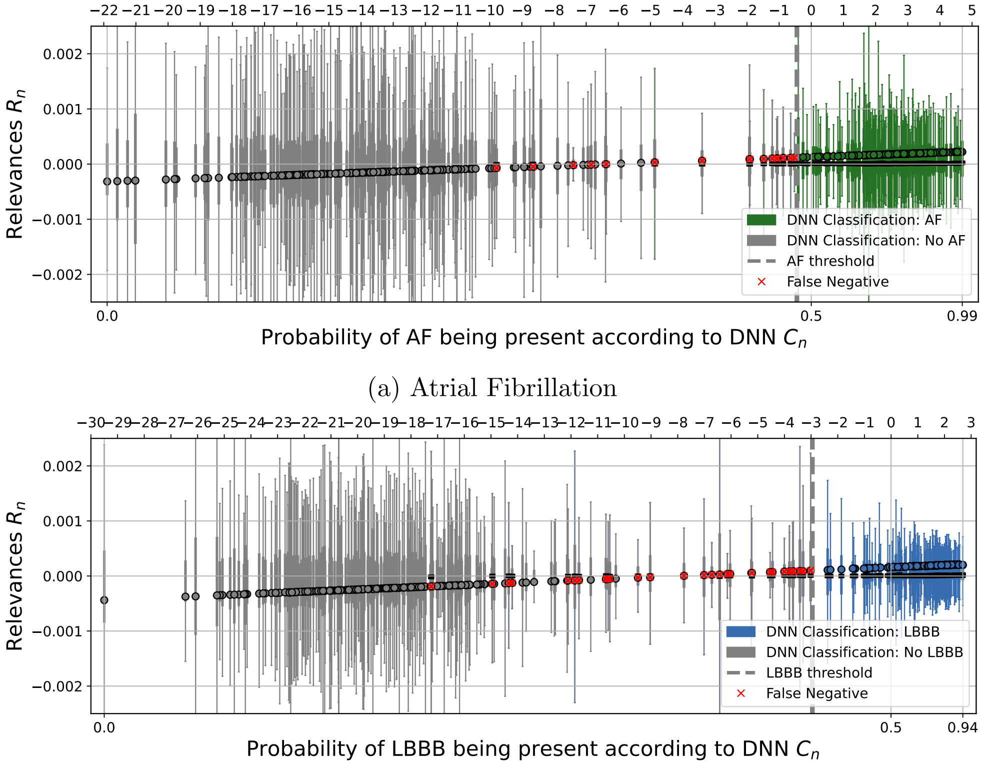
(a) Normal and AF recordings. Colors denote ground truth label of data set. Values for AF range from $[-0.5, 0.5]$ and values for normal recordings from $[-0.3, 0.4]$.

(b) Normal and LBBB recordings. Colors denote ground truth label of data set. Values for LBBB range from $[-0.6, 0.9]$ and values for normal recordings from $[-0.4, 0.5]$.

Figure 4.2.: Distribution of IG relevance scores $R_{n,j,k}$. To increase visibility, x-axes are limited to $[-0.20, 0.20]$. [3] © 2023 IEEE

4.2.3. Entropy Analysis of Relevance Scores

Analyzing the relevance scores of model probabilities for AF classification with *SampEn*, as depicted in Fig. 4.7, reveals similar ranges of $[0.03, 0.80]$ and $[0.06, 0.82]$ for AF and SR patients, respectively. Moreover, 11 out of 12 leads exhibit lower median values for AF patients, with the highest difference being 0.15 (lead V5). Through visual inspection, a clustering of relevance scores is observed in the area of QRS complexes. During measurement noise (e.g., inadequate skin-electrode contact), clusters with high absolute values and interchanging signs agglomerate.



(a) Atrial Fibrillation

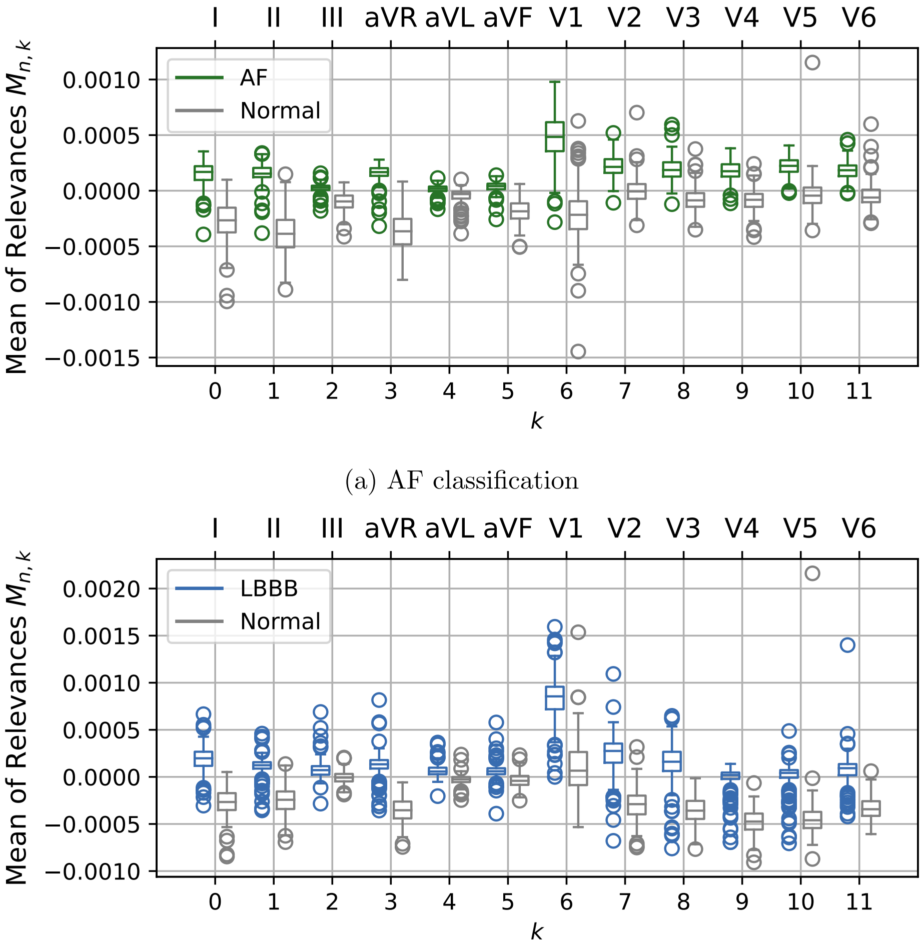
(b) Left Bundle Branch Block

Figure 4.3.: Distribution of R_n computed with IG for each recording as single boxplot. The bottom x-axis represents sigmoid activation output of the DNN, while the upper x-axis represents the output with linear activation. Boxplot colors denote DNN classification results and red crosses indicate false negatives. [3] © 2023 IEEE

4.2.4. Average Relevance Scores Over Class, Lead, and Beats

When classifying AF, QRS complexes are the most relevant areas, especially the R-peaks (Fig. 4.5). For normal recordings, high negative values can be observed for the area of P-waves. Negative values of normal recordings are higher compared with positive values of AF recordings.

For the LBBB classification, QRS complexes are also most relevant (Fig. 4.6). Furthermore, the concentration of high absolute relevance scores on specific waves or peaks is clearer, such as the negative T-wave in LBBB, which is assigned with negative relevance scores when positive in normal recordings. By contrast, for AF, many smaller relevance scores with higher variance are distributed over the whole beat.



(a) AF classification

(b) LBBB classification

Figure 4.4.: Distribution of $M_{n,k}$ computed with IG w.r.t. ECG leads, colors denoting ground truth label. For AF classification (a) and LBBB classification (b) boxplots show that the abnormal mean is higher for each lead with the highest difference in V1. [3] © 2023 IEEE

4.2.5. Qualitative Analysis

Clusters of high absolute relevance scores can be observed in the area of QRS complexes during visual inspection of single recordings, which are visualized as a heatmap in Fig. 4.8. For LBBB, IG seems to focus on negative S-waves and prolonged ST-segments in lead V1. Occasionally, broad and notched R-waves are also marked as relevant. By contrast, for AF recordings, the relevant parts are usually R-waves and, in rare instances, areas with missing P-waves.

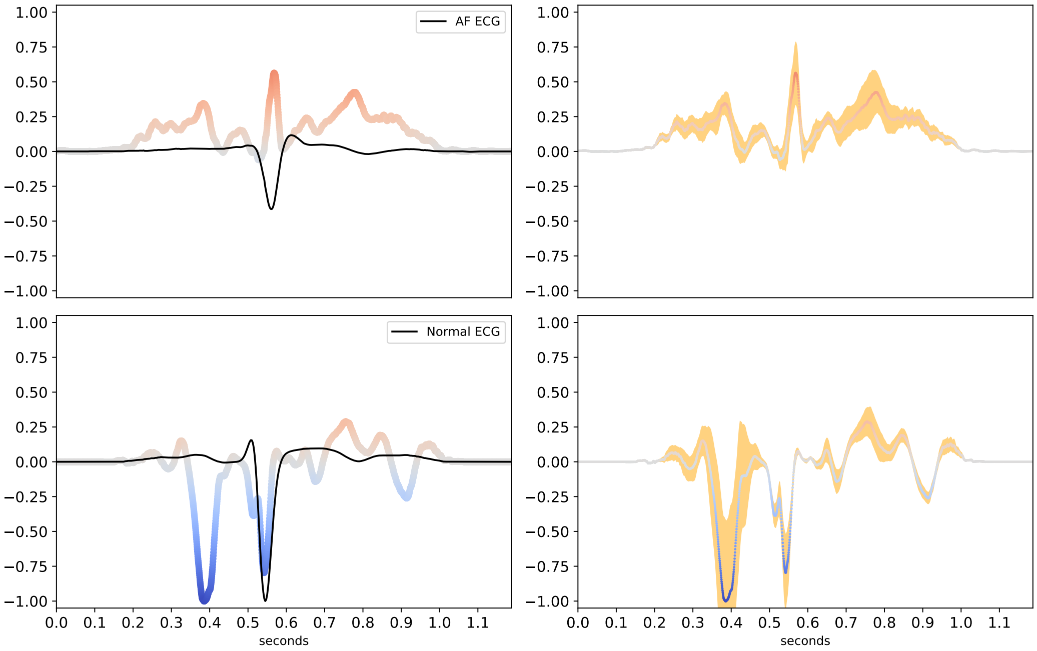


Figure 4.5.: Left column: Average beats (black curves) and IG relevance scores for lead V1 in AF classification. Abnormal ECGs show positive relevance scores (red) distributed over the whole P-QRS-T-cycle, negative relevance scores (blue) on normal recordings cover QRS complexes and especially P-waves. Right column: Instead of average beats, the variance of relevance scores across recordings is shown (orange). [3] © 2023 IEEE

When examining individual recordings, it can be observed that in cases of artifacts, such as baseline drifts or noise, the IG relevance scores are usually accumulated mainly in these areas. This can be seen in multiple false negative classifications, such as recordings A1017 (lead V1, Fig. 4.9), A0745 (V6), and A0205 and A0502 (both multiple leads, mainly: V1-6). In some cases, the classification is still correct despite the focus on artifacts, such as A0639 (V1) being classified as AF with ≈ 0.904.

4.2.6. Comparison of Databases

All quantitative methods exhibit similar results for PTB-XL data, as demonstrated for average beats in AF classification in Fig. 4.10. In particular, the distribution of relevance scores for LBBB recordings is narrower and shifted closer to positive values than for CPSC data (Fig. 4.11).

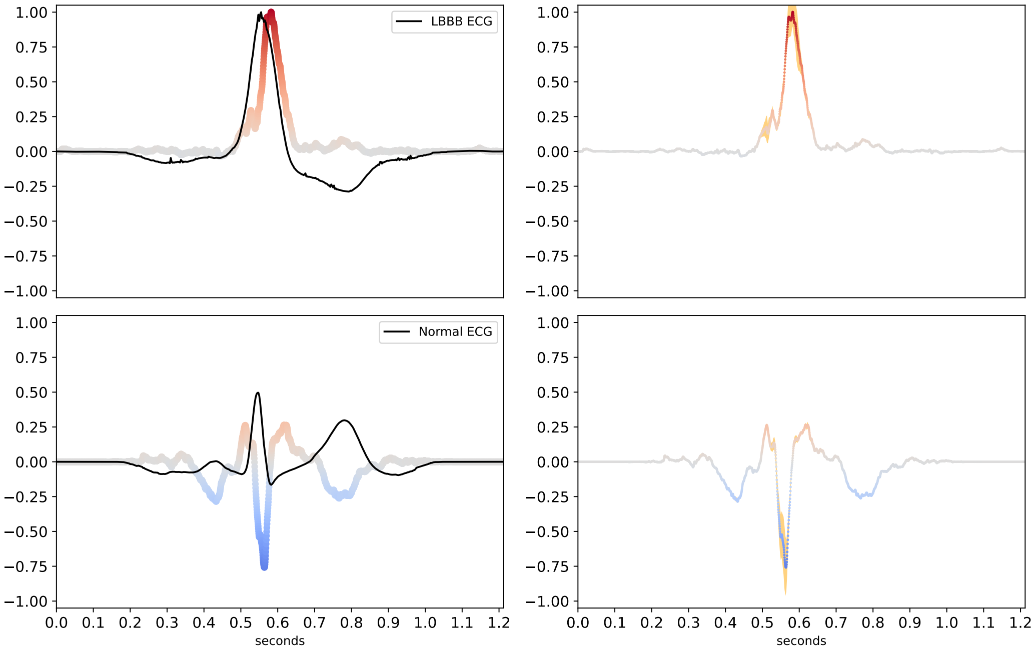


Figure 4.6.: Left column: Average beats and IG relevance scores for lead aVL in LBBB classification. Abnormal ECGs show positive relevance scores (red) on QRS complexes; negative scores (blue) on normal recordings can be seen on P- and T-waves. Right column: Instead of average beats, the variance of relevance scores across recordings is shown (orange). [3] © 2023 IEEE

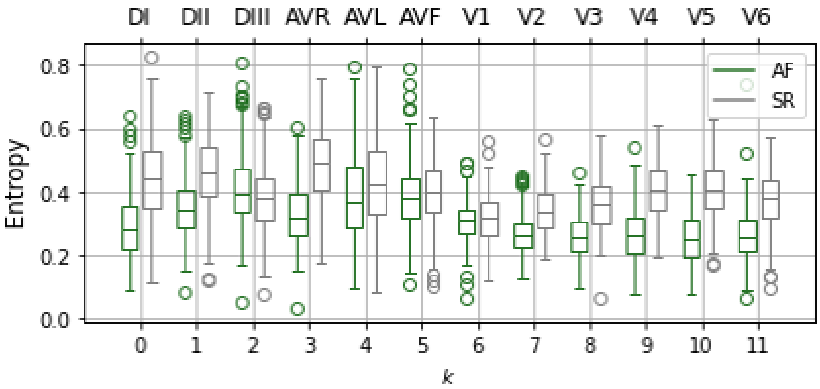


Figure 4.7.: SampEn entropy of IG relevance scores for AF classification with regard to lead and label.

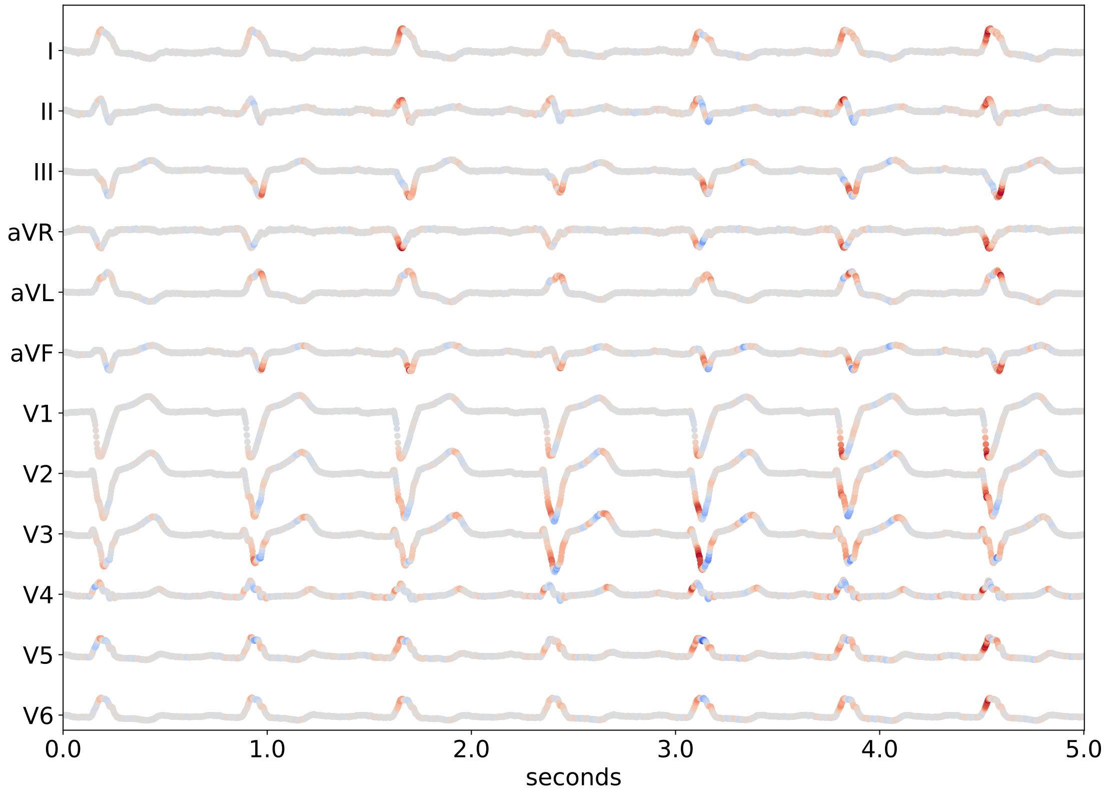


Figure 4.8.: Positive (red) and negative (blue) relevance scores calculated with IG on a correctly classified ECG (C_n(LBBB): ~ 0.871) from CPSC data set (ID A0977). Relevance scores normed to $[-1, 1]$ per lead. [3] © 2023 IEEE

4.2.7. Comparison of XAI Methods

IG and all considered LRP methods yield diverging results for the given data set. As seen in Fig. 4.12, for example, LRP methods ϵ and $\alpha\beta$ distribute high absolute relevance scores, especially around R-peaks, while ω^2 exhibits higher absolute values on waves in between as well as on artifacts. IG can also concentrate high absolute relevance scores around artifacts, but it generally exhibits more high absolute values, especially on R-peaks, when the leads of single recordings are compared with each other.

4.3. Discussion

Analyzing individual recordings (Fig. 4.3) reveals similar distributions of C_n for both classifications. Additionally, a distinct relationship between the averaged relevance scores M_n and the probability of the DNN C_n can be observed. A DNN classifier able

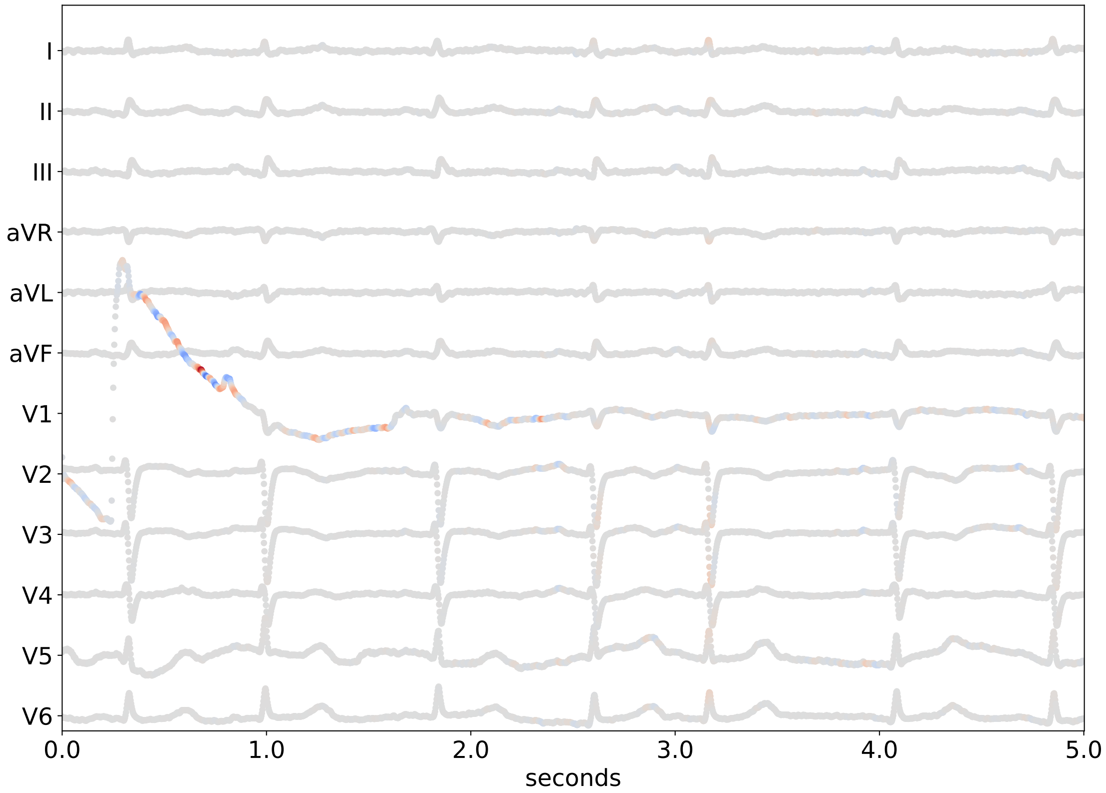


Figure 4.9.: Positive (red) and negative (blue) relevance scores calculated with IG on a false negative classified ECG (C_n(AF): $\sim$ 0.008) from CPSC data set (ID A1017). Relevance scores are clustered around the artifact in lead V1. [3] © 2023 IEEE

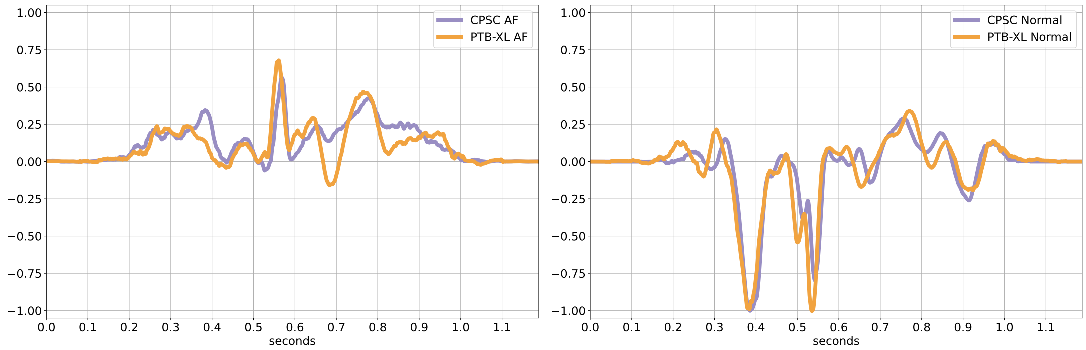


Figure 4.10.: IG relevance scores for lead V1 averaged over 200 ECGs extracted from CPSC (blue) and PTB-XL (orange). Figures depict AF recordings (left) and normal recordings (right), respectively. [3] © 2023 IEEE

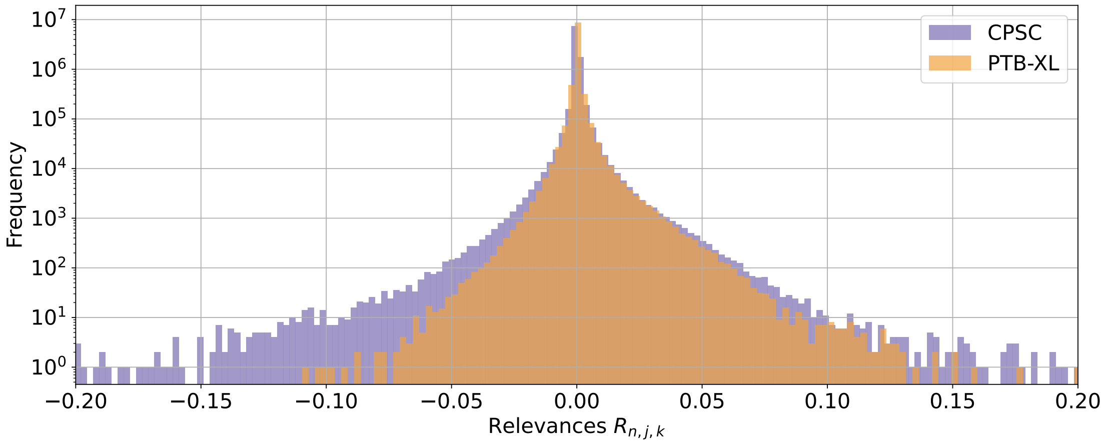


Figure 4.11.: Relevance scores of LBBB recordings from CPSC database (blue) compared to PTB-XL data (orange). To increase visibility, the x-axis is limited to $[-0.20, 0.20]$. Values for LBBB range from $[-0.11, 0.21]$ and values for normal recordings from $[-0.64, 0.56]$.[3] © 2023 IEEE

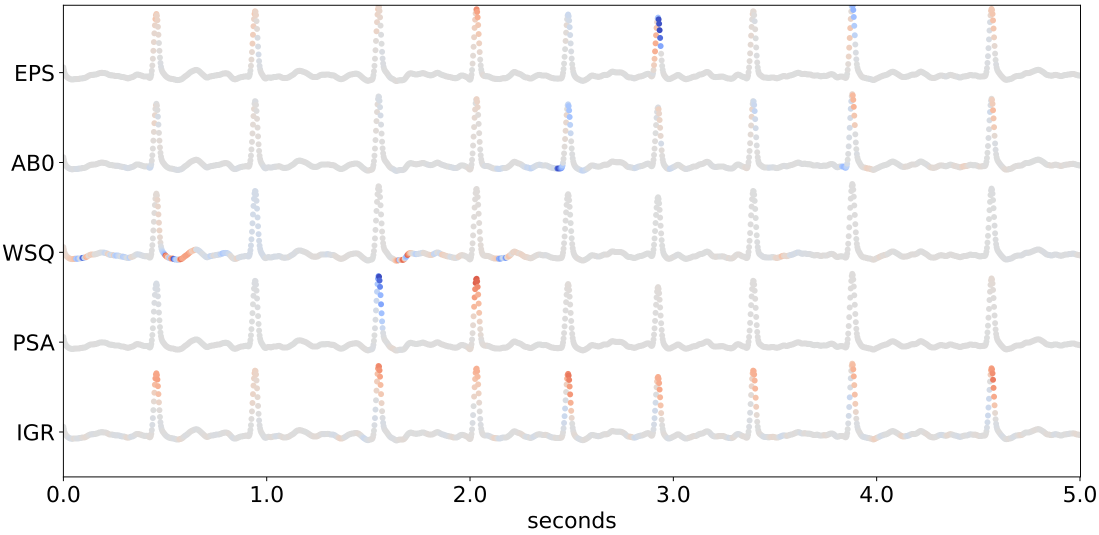


Figure 4.12.: Relevance scores calculated with five XAI methods normed to $[-1, 1]$ each on lead V6 of a correctly classified ECG (C_n(AF): ~ 0.987) from CPSC data set (ID A0086). EPS: LRPEpsilon, AB0: LRPAlpha1Beta0, WSQ: LRP-WSquare, PSA: LRPSequentialPresetA, IGR: IntegratedGradients. [3] © 2023 IEEE

to clearly separate a specific class would show a cluster nearby $C_n = 0$ and $M_n \ll 0$ for normal recordings as well as a cluster nearby $C_n = 1$ and $M_n \gg 0$ for the analyzed abnormality (AF/LBBB). For the Ribeiro model, the classes are not that clearly separable, which can generally be expected with a transition area between both clusters in which the DNN does not have high certainty in its decisions. Additionally, IG relevances have a tendency of exhibiting a higher complexity in SR than in AF patients during AF classification (cf. Subsection 4.2.3), suggesting a higher uncertainty of the DNN when tending toward a low C_n. Furthermore, many of the false negative classifications are slightly below the threshold, indicating that the thresholds might not be optimal for new data sets.

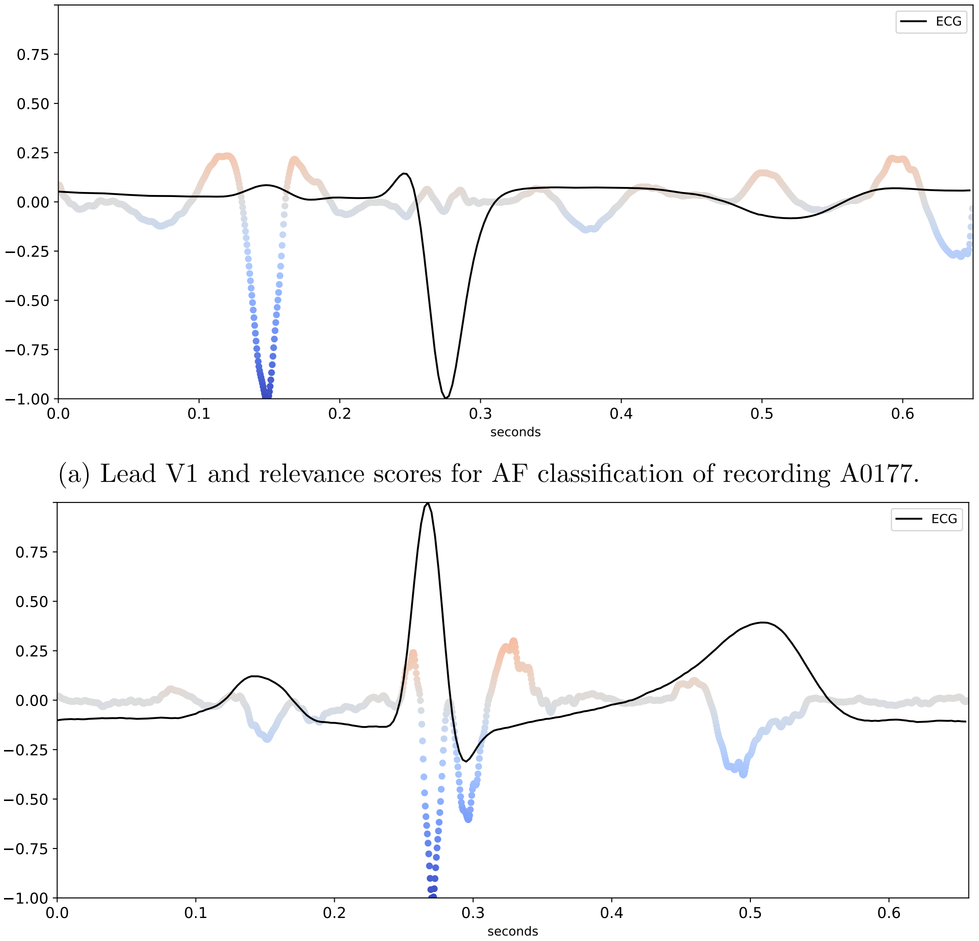
When analyzing individual leads, significant differences in relevance score distributions between abnormal and normal recordings are revealed (Fig. 4.4). This indicates which leads are most relevant for the DNN's decision. In general, for AF, the limb leads have lower relevance scores compared with the chest leads [29]. For AF as well as LBBB classifications, lead V1 has clear positive relevance scores, indicating that the model has trained clinically relevant features (cf. Section 2.1).

Time information is lost in average means as described before; therefore, a third analysis of the average beat and average relevance scores of a single lead could provide an even more detailed idea of the model's features. Although it is still not possible to uniquely identify the actual features learned by the DNN, positively relevant areas in the case of missing P-waves for AF classification indicate a good fit to clinical criteria [61]. Additionally, for the healthy controls, there are highly pronounced negatively relevant areas near P-waves, demonstrating that the DNN has learned that the existence of P-waves is a counter-sign for AF. As the analyses do not allow us to gain insights into the time scale, it cannot be quantified to which extend RR-interval variations impact the relevance scores.

Moreover, when analyzing the shape of an average relevance signal, which is continuously averaged over an increasing number of recordings, it can be seen that, for AF as well as normal ECGs, the variance of relevance scores is quite low. This indicates the robustness of the DNN, as it generates similar relevance scores despite the natural inter-patient variability in abnormal ECGs.

Regarding LBBB classification, high relevance scores around broadened QRS complexes indicate a good fit with clinical criteria [30]. The criterion of a T-wave displacement orthogonal to the major deflection of the QRS complex [30] can also be observed very well, although it results in only small positive relevance scores.

By contrast, for healthy controls, T-waves result in highly pronounced negatively relevant areas (e.g. Fig. 4.13b). Similarly, for AF classifications, P-waves are learned as a feature that indicates the absence of AF (e.g. Fig. 4.13a). Furthermore, the robustness of the relevance scores in terms of variance is even higher than that for AF.



(a) Lead V1 and relevance scores for AF classification of recording A0177.

(b) Lead aVL and relevance scores for LBBB classification of recording A0053.

Figure 4.13.: Average beats (black curve) and relevance scores for individual leads in a single normal recording representative for features found in average beats, correctly classified by the DNN: a) Highly negative relevance scores (blue) are found during the occurrence of the P-wave. b) Negative relevance scores (blue) are found during the P-/T-waves, and especially during occurrence of the P-wave of the QRS complex. [3] © 2023 IEEE

4.3.1. Comparison of XAI Methods

The presented results indicate that both XAI methods applied in this study, IG and LRP, are well suited to gaining insights into the reasoning of DNNs applied to biosignals. Additionally, the comparison of IG and LRP methods leads to the conclusion that IG produces the most distinct results.

4.3.2. Comparison of Databases

Analyses with PTB-XL obtain similar results to CPSC 2018. One noticeable difference is observed in the relevance score distribution of LBBB recordings, where less negative values for PTB-XL could be explained by the more specific label "Complete LBBB", which might be easier to classify. These more differentiated labels have potential for a comparison of model performance on complete and incomplete LBBBs.

4.3.3. Artifacts

The DNN tends to produce incorrect classifications when artifacts are present. This effect has been observed by others as well, such as [19]. Although it is not attempted in this work, artifact detection based on the presented approach could be a promising avenue for future work.

4.3.4. Limitations

However, a limitation of the conducted analyses based on IG is that from the relevance scores, no time-dependent information can be inferred. Especially for AF, it is unclear whether, for example, the R-peaks are marked as relevant because of their morphology or distance from one another. Therefore, the results can be rated as more robust for LBBB as a morphological abnormality than for AF as an arrhythmic and thus time-dependent abnormality.

4.3.5. Key Findings

In summary, this analysis suggests that the model by Ribeiro et al. learned features similar to cardiology textbook knowledge. IG relevance scores indicate that it learned features that point toward a disease, such as the abnormal QRS complex in LBBB, while other features, such as the T-wave pointing in the opposite direction, are not used for LBBB detection. Instead, the opposite of the feature, a T-wave pointing in the expected direction, is used as a feature for detecting healthy ECGs. The analysis and visualization methods for relevance scores proposed in this study facilitate a rapid and effective assessment of the DNN's learned features and have been confirmed by cardiologists.

CHAPTER 5

Robustness

> *This chapter is based on the conference publication Benchmarking the Impact of Noise on Deep Learning-Based Classification of Atrial Fibrillation in 12-Lead ECG [3] (see Appendix C).*

Analyzing the Ribeiro model for its robustness, the goal of this study is to benchmark the influence of four types of noise on its accuracy.

5.1. Methods

The data used in this study is a subset of PTB-XL (cf. Subsection 2.5.2), including the metadata provided by human experts regarding noise for assigning a signal quality to each ECG. The accuracy of the Ribeiro model (see Subsection 2.2.1) is analyzed with respect to the following two SNR metrics:

SNR based on annotations (SNR_a): For each ECG, the number of noisy leads is determined using the metadata fields described in Section 2.5.2. Using a custom script, this information is converted into numeric values ranging from 0 to $12 \in \mathbb{N}$ for each type of noise. The labels “alles” (all) and “noisy recording” are converted to 12. ECGs associated with other labels are removed as they are of a more qualitative nature (e.g. “leicht” [light]). Thus, for each signal, a qualitative, unit-less, linear SNR measure is

```python
import numpy as np

fund_energy = 0
remain_energy = 0
s = abs(np.fft.rfft(raw_data[patient][:,lead]))
W = np.fft.rfftfreq(4096, 1/400)
u = s.copy()
u[:] = 1
u[(W < 40)] = 0
u[(W > 150)] = 0
for f in range(1,len(s),1):
    fund_energy += (u[f] * s[f]) ** 2
    remain_energy += ((1 - u[f]) * s[f]) ** 2
snr = 10*np.log10(fund_energy/remain_energy)
```

Listing 5.1: Calculation of the measured SNR in dB, with the signal frequency band defined between 40 and 150 bpm.

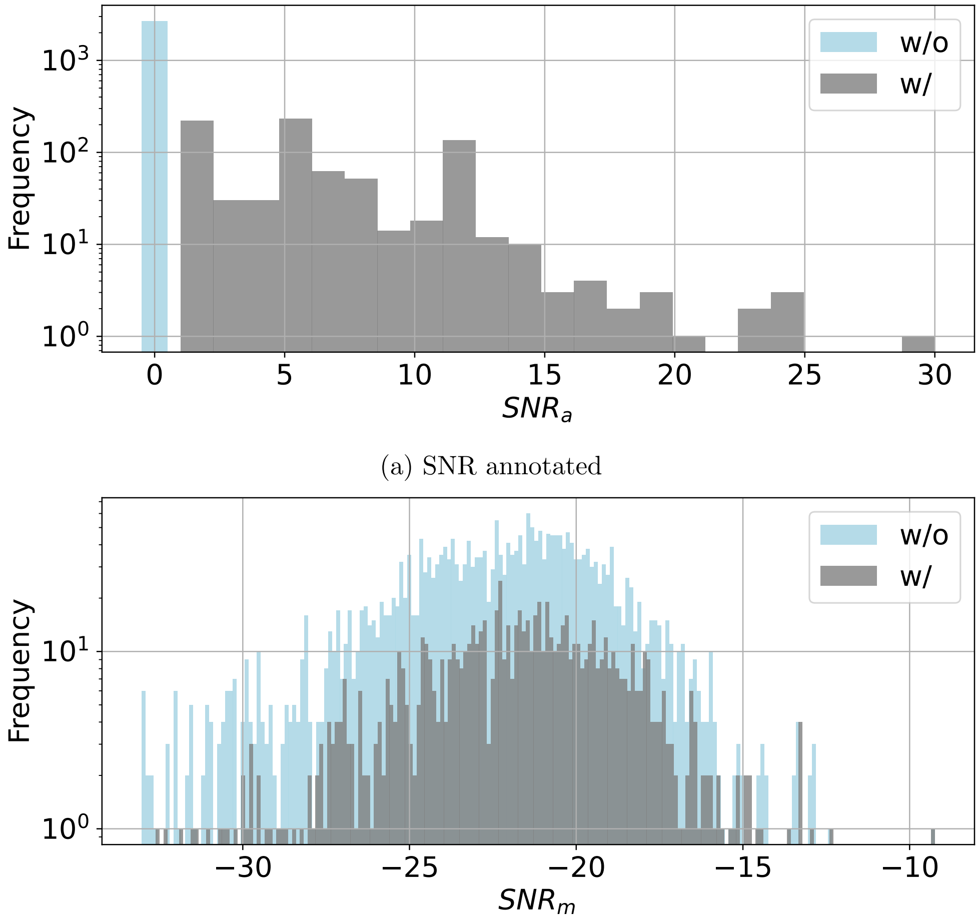
Table 5.1.: Properties of subset extracted from PTB-XL (left) and results of DL-based AF classification (right). ECGs are grouped according to annotations: In case there is one or more noise label in the metadata, an ECG is assigned to "with" label, else to "without". FP and FN denote False Positive and False Negative, respectively. [3]

Noise Label	AF	Healthy controls
without	1,097	1,581
with	417	419

Noise Label	DL: FP	DL: FN
without	0.04 %	3.96 %
with	0.24 %	7.06 %

computed, ranging from 0 if no noise is reported to $12 * 4 = 48$ if all leads are affected by all types of noise. As indicated in Tbl. 5.1, this information is used to split the data set in ECGs without a noise label and ECGs with a noise label. It must be underlined that a value of zero does not guarantee that the signal is free of noise; it just reflects that there is a potential for a noise-free ECG.

Measured SNR ($\mathbf{SNR_m}$): Furthermore, a quantitative SNR is computed for each ECG. Due to the limitations of the manual annotations and the fact that they are only available for 22% of the PTB-XL database [62], additionally a quantitative SNR measure is used for each signal. The Fourier transform of the signals as well as the ratio of energies is computed in two frequency bands, as proposed in [63]. Based on the expected heart rates during AF, the "signal" frequency band is defined as ranging from 40 to 150 beats per minute ($0.66 - 2.5$ Hz) and the "noise" frequency band as < 40 and > 150 beats per minute. Scaling with $10 \log 10$ results in an SNR expressed in the logarithmic decibel scale (dB), as presented in Listing 5.1.



(a) SNR annotated

(b) SNR measured

Figure 5.1.: Distribution of the values of both SNR metrics, with (grey) and without (blue) noise labels. [3]

5.2. Results

Fig. 5.1 depicts the distribution of $\mathrm{SNR_a}$ and $\mathrm{SNR_m}$ values. The majority of ECGs with noise labels have a value lower than 15 with the maximum being 29. This indicates that even in the duration of 10 seconds, different data quality issues per lead may occur. $\mathrm{SNR_m}$ values occur in the range of $[-33.03, -7.78]$ dB, with no clear difference between ECGs with and without noise labels.

Regarding the model performance, it can be observed from Tbl. 5.2 that the model can robustly identify AF even in cases where signals are labeled by human experts as being noisy on multiple leads. False positive and false negative rates are slightly worse

Table 5.2.: DNN accuracy w.r.t. the four types of noise. The variable n represents the number of signals with the given label. For comparison to signals without a label, n ECGs are randomly drawn 100 times and accuracy is given as mean ± standard deviation. [3]

Type Label	Baseline Drift ($n = 305$)	Static Noise ($n = 478$)	Burst Noise ($n = 156$)	Electrode Problems ($n = 6$)
without	96.8% ± 0.9%	96.8% ± 0.7%	96.9% ± 1.3%	96.3% ± 8.0%
with	97.7%	94.6%	94.9%	100.0%

for data labeled as noisy, with F1-scores of 0.949 without and 0.921 with noise labels. Noteworthily, data annotated as exhibiting baseline drift noise result in an accuracy that is very similar to that of data without.

5.3. Discussion

The reasons for this behavior could be explained by a partial misinterpretation of baseline drift or static noise as P-waves. As indicated in Chapter 4, the DL model is trained such that P-waves and R-peaks have high relevance, similar to human perception, while numerous other features influence its decision. This multi-factor decision process could be robust to different kinds of noise, although this requires its presence during training.

However, since the distribution of SNR_m looks visually similar with or without noise labels, SNR_a might not be optimal for quality assessment on its own. A "no noise" label that explicitly identifies ECGs without data quality issues as well as more labels in general would be valuable additions for future experiments.

In conclusion, the issue of processing noisy ECG data can be addressed successfully by DL methods, which might not require as much preprocessing as many conventional methods.

CHAPTER 6

Generalizability

The analyses on public data sets regarding the performance and explainability of a DL model in previous chapters yielded promising results. However, these data sets are usually curated and might not reflect the quality of real-world data acquired in clinical practice.

In Chapter 3, the Ribeiro model was already tested on a small, selected data set from UMG and reached high F1-scores. To further investigate the suitability of the model for uncurated data, a full export of ECG data acquired at the UMG Department of Cardiology is initiated and analyzed in a similar manner to the public data sets.

6.1. Methods

The analyses presented in Chapters 3 and 4 are repeated with the data set ECG-full-2021 (cf. Subsection 2.5.3). I analyze a total of $3,495$ ECGs with LBBB and $10,001$ with AF labels recorded with Schiller devices from 2002 to 2021 at UMG with the built-in annotations as ground truth.

Table 6.1.: Comparison of DL model results including F1-scores for ECG-full-2021. Calculated for both LBBB and AF, with RBBB and SR as controls, respectively. Considered analyses: DNN classification and scores published by Ribeiro et al. [37]. FP and FN denote False Positive and False Negative, respectively.

		LBBB	**AF**
DNN (ECG-full-2021)	**FP**	721 (18.2 %)	74 (26.8 %)
	FN	113 (3.2 %)	2,679 (0.7 %)
	F1-Score	0.890	0.842
DNN (Ribeiro et al.)	**FP**	0	0
	FN	0	3
	F1-Score	1.000	0.870

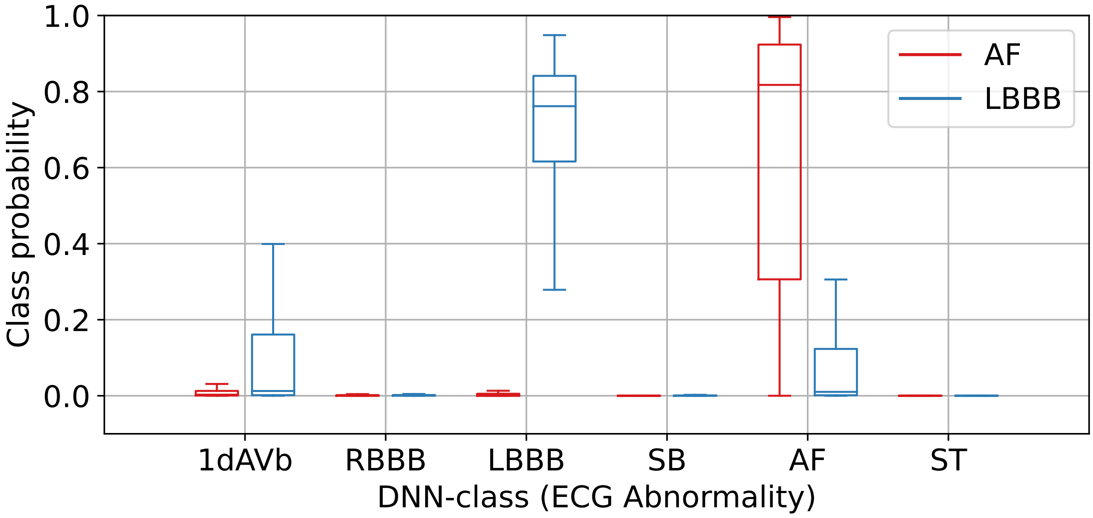


Figure 6.1.: DNN classification results on ECG-full-2021. Separated into cohorts annotated with either AF (red, n=10, 001) or LBBB (blue, n=3, 495) as annotated by the device's built-in algorithm. Considered abnormalities are: 1st degree AV block (1dAVb), right bundle branch block (RBBB), left bundle branch block (LBBB), sinus bradycardia (SB), atrial fibrillation (AF) and sinus tachycardia (ST).

6.2. Results

6.2.1. DL Performance

When a large uncurated data set is analyzed, the model results for LBBB are similar to the built-in labeling, with only 3.2% not classified as LBBB. For recordings labeled AF, the model classifies more than a quarter less than the built-in method. Performance

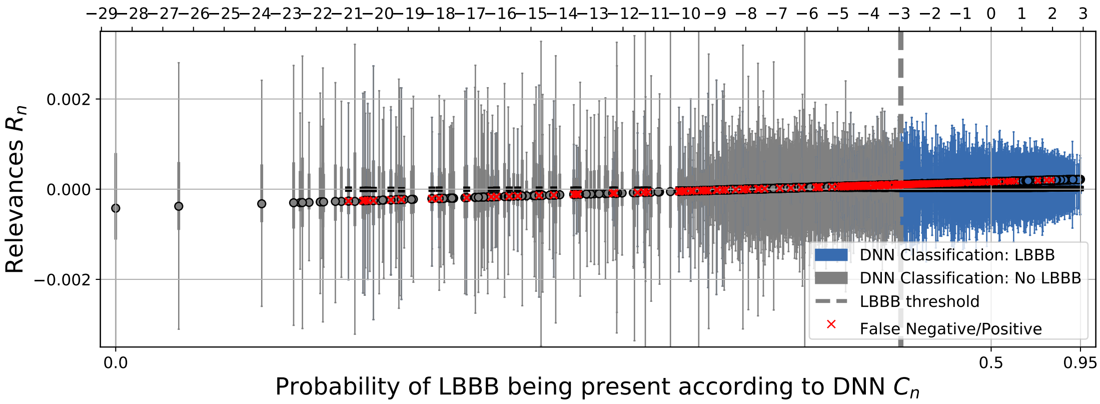


Figure 6.2.: Distribution of R_n computed with IG for each LBBB and RBBB recording as single boxplot (ECG-full-2021). The bottom x-axis represents sigmoid activation output of the model, while the upper x-axis represents the output with linear activation. Boxplot colors denote LBBB classification results and red crosses indicate false negatives and false positives.

metrics compared with the original publication are provided in Tbl. 6.1.

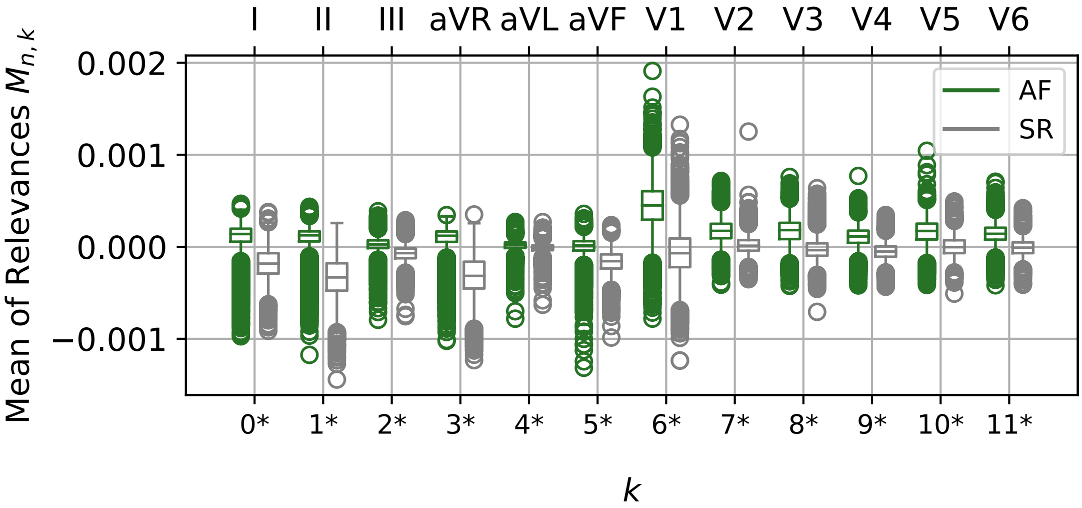
Upon analyzing model results for all abnormalities (Fig. 6.1), many LBBB recordings again have an additional 1dAVb classification (cf. Subsection 3.2.2). Furthermore, the distribution of classification probabilities is larger than that for the small, selected data set in [1].

6.2.2. DL Explainability

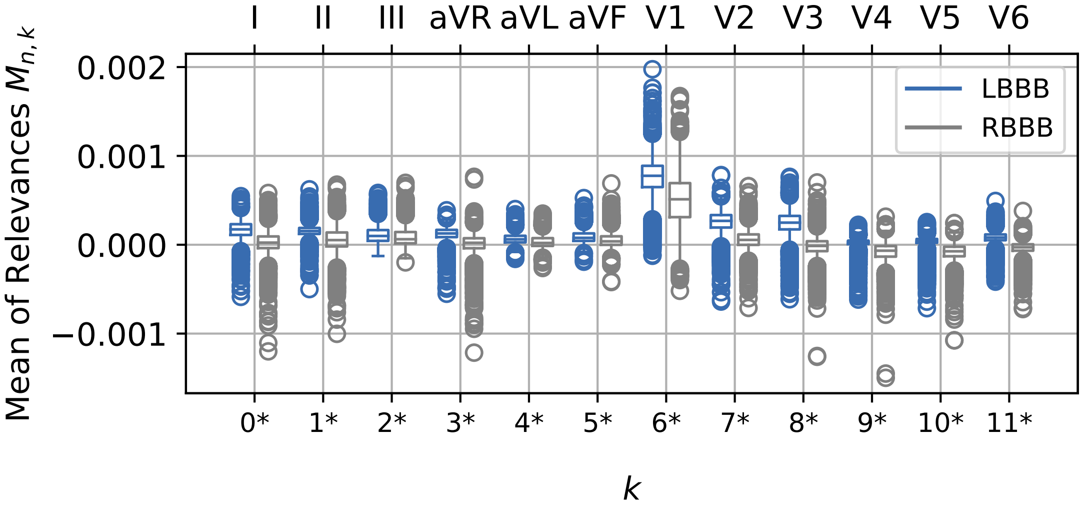
The quantitative analyses described in Section 4.1 are repeated with ECG-full-2021 to test the proposed analyses on real-world ECGs. Since the devices do not provide labels that correspond to normal ECGs, the rhythmic abnormality AF is compared with sinus rhythm (*SINUSRHYTHMUS*), while the morphological abnormality LBBB is compared with its counterpart RBBB (*RECHTSSCHENKELBLOCK*), which usually do not occur simultaneously.

The relevance scores exhibit similar trends as those for public data sets, as illustrated in Fig. 6.2. Mean values increase with higher classification probabilities, with most false classifications occurring when the mean is around zero. In contrast to healthy controls investigated for the CPSC and PTB-XL data sets, the RBBB recordings are closer to the classification threshold and almost overlap with the LBBB recordings. Overall, the classification results are less clear than those in the public data sets.

This can also be observed in the boxplots of mean relevance scores per lead, which



(a) AF classification: AF vs SR recordings.



(b) LBBB classification: LBBB vs RBBB recordings.

Figure 6.3.: Distribution of $M_{n,k}$ computed with IG w.r.t. ECG leads (ECG-full-2021), colors denoting ground truth label. For AF classification (a) and LBBB classification (b), boxplots show that the abnormal mean is higher for each lead.

are presented in Fig. 6.3. Again, the results correspond to clinical recommendations with, for example, lead V1 having the highest mean and difference to SR recordings for AF classification (see Fig. 6.3a). However, in Fig. 6.3b, a small difference is observed between the LBBB and RBBB recordings.

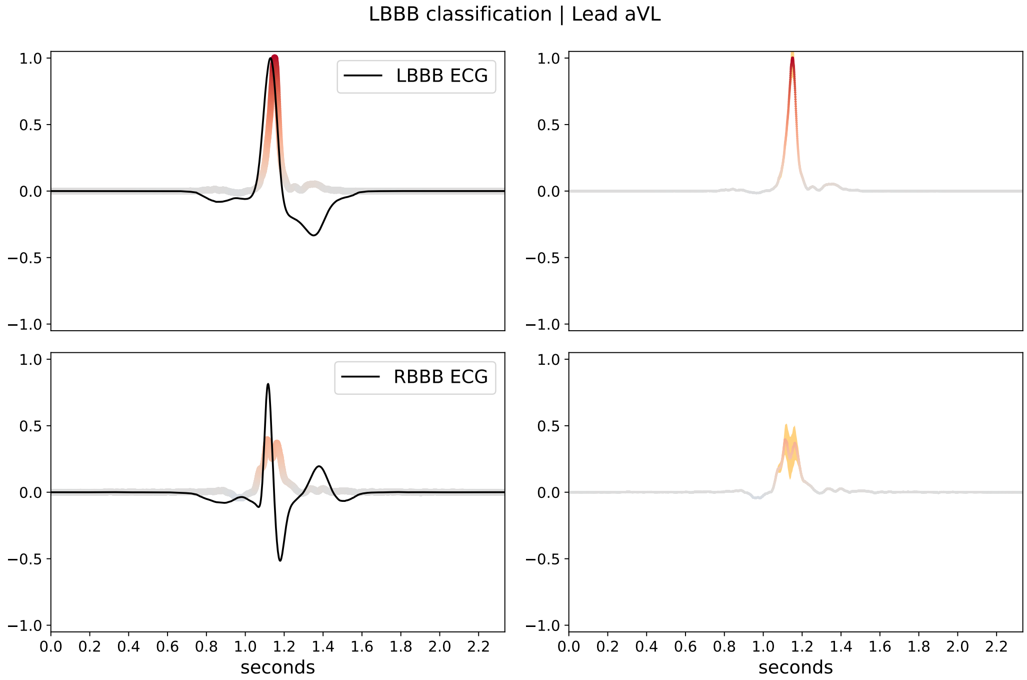


Figure 6.4.: Average beats and IG relevance scores for lead aVL in LBBB classification (ECG-full-2021). Left column: Average beats and IG relevance scores for lead aVL in LBBB classification. Abnormal ECGs depict positive relevance scores (red) on QRS complexes; negative scores (blue) on RBBB recordings can lightly be seen on P-waves. Right column: Instead of average beats, the variance of relevance scores across recordings is depicted (orange).

Furthermore, in the average beats and relevances presented in Fig. 6.4, the relevance scores are again high and positive on QRS complexes for both LBBB and RBBB. Only around the P-wave can some negative relevance scores be seen in RBBB recordings.

6.3. Discussion

When the built-in algorithm is used as the ground truth for a larger data set, the model shows lower performance scores. For AF recordings, this could be due to a possible over-labeling of the built-in algorithm noticed by the cardiologists.

Especially when classifying recordings that exhibit other abnormalities, such as RBBB in the LBBB classification in Fig. 6.2, more recordings tend to be found in the transition area, leading to more false classifications. It is possible that the model has problems distinguishing between these abnormalities since the clinical criteria are quite similar, including broad QRS complexes which the model seems to focus on.

However, areas of high relevance scores that have been identified in Chapter 4 could be reproduced on this uncurated routine data set, including a focus on lead V1 exhibited in Fig. 6.3 as well as QRS complexes in Fig. 6.4.

Overall, a ground truth based on the diagnosis is required to interpret these results correctly. Furthermore, if one wants to learn more about the counter-features that the model is looking for, such as the presence of P-waves in AF classification, criteria for labeling recordings as "normal" based on built-in annotations must be defined. As soon as the UMG ECG recordings are fully integrated with other clinical data in the MeDIC infrastructure, the diagnoses of clinicians could be used as ground truth.

Finally, the data will be made available for researchers who apply for them in an XNAT instance, which will allow them to run scripts such as those developed in this thesis with Jupyter extensions.

CHAPTER 7

Discussion

The aim of this thesis was to analyze the suitability for clinical applications of a DL model with good performance trained on a large data set. Experiments with both curated public and uncurated private ECG recordings were performed in terms of the reproducibility, explainability, robustness and generalizability of DNNs, with the example of a model for 12-lead ECG classification by Ribeiro et al. [37]. In the following subsections, the results of each experiment are discussed in detail.

7.1. Reproducibility and Transparency

Some pre-trained DL algorithms are shared publicly on platforms such as Zenodo[1], while most papers at least offer the model source code to reproduce results on other data sets. However, the original test data sets requried for method reproducibility are often not accessible, which hampers trust in DL models.

[1] https://zenodo.org/

> In Chapter 3, an uncurated data set from the Department of Cardiology at UMG was classified for two abnormalities with a ResNet by Ribeiro et al., published on Zenodo [38], to investigate the model's performance and reproducibility. The performance on 29 recordings with either AF or LBBB was comparable to the performance on the original test data set [39] with F1-scores of 1 and 0.919, respectively, indicating good reproducibility. Repeating the experiment with the test data set revealed good method reproducibility as well. [1]

Although similar studies have also indicated that ML algorithms can be reproducible [64, 65], opportunities remain to enhance their reproducibility further when adhering to guidelines such as those proposed by [66] for AI algorithms in healthcare. With these guidelines, the authors found only three out of the eight most commonly cited healthcare algorithms to be reproducible. As demonstrated in Chapter 6, Ribeiro et al. fulfilled the model's *generalizability*. *Collaboration* is also fulfilled, with the publication of the model on Zenodo allowing for reproducibility. Loftus et al. [66] further included external guidelines, such as those for the reproducibility criterion *compliance*, where interventions involving AI algorithms must also adhere to the CONSORT-AI guidelines [67].

To overcome the problem of many AI algorithms not being reproducible [68], criteria for training reproducible DL models were proposed by Chen et al. [69], who successfully reproduced seven different models. Hence, reproducibility needs to be considered at the earliest development stage possible. However, the availability of pre-trained models with published code and test data and exhaustive documentation can contribute largely to the reproducibility of specific models.

7.2. Explainability

Despite their remarkable performance compared with, for example, conventional algorithms adopted in emergency rooms [70], DL models have not yet been adopted in clinical routine. Next to reproducibility issues, as discussed earlier, a critical reason for this is the lack of understanding of how these models work, since a clinician has responsibility in the outcome of a patient when considering model outputs in their decision making.

Whether it is the analysis of street images, where autonomous driving cars could cause casualties, such as if pedestrian detection is complicated by misty weather [71], or models learning shortcuts in radiographic images instead of clinical features [72], explainability plays a critical role whenever an AI-based decision can result in notable damage.

In clinical routine, the application of each algorithm must already adhere to the Medical Device Regulation (EU) 2017/745 (MDR) [73]. Furthermore, for the use of AI algorithms, a new specific EU law is on the way, the so-called AI Act[2], which will regulate requirements especially for high-risk applications, such as those in clinical settings. Reproducibility and explainability will be part of these regulations, so they need to be examined for medical research applications, which have been the focus of this thesis, before they can be used in clinical routine.

> In Chapter 4, quantitative methods for the analysis of XAI output for models analyzing ECGs were proposed and demonstrated as exemplary for the Ribeiro model. The most relevant areas correspond to clinical recommendations regarding which lead to consider, moreover, visible P-waves and concordant T-waves result in clearly negative relevance scores in AF and LBBB classification, respectively. [2]

This approach helps to understand patterns learned by pre-trained models where it is impossible to use ante-hoc methods for a more detailed analysis. Currently, most XAI analyses are demonstrated to be exemplary on single biosignal recordings [74]. As part of my thesis, I was able to demonstrate that quantitative analyses can provide more insights into patterns learned by DNNs, allowing generalized conclusions about the decision making of the model. Once the typical relevance score distributions of a model are identified, it is feasible to return to the qualitative analysis of single outliers. For example, unexpectedly relevant leads might point toward artifacts, while mean relevance scores around zero could be an indicator of difficult decisions, where a clinician might need to check the model output more thoroughly.

In clinical settings, we are still far from the regular use of DL algorithms, especially due to the ongoing research on explainability. As clinicians need to be able to make decisions themselves [75], the focus is currently on clinical research. The proposed analyses can be a first step in the definition of metrics for the trustworthiness of single DNN decisions, highlighting challenging cases where the results should be carefully checked.

Research on post-hoc explainability is constantly yielding new insights. A recent study by Wagner et al. [76] reported results similar to this work about clinically relevant features found in DL models, comparing several post-hoc methods. Most notably, current research, such as that on partial information decomposition by Ehrlich et al. [77] could greatly accelerate the process of making DL trustworthy in terms of explainability.

[2] https://www.artificial-intelligence-act.com/

7.3. Robustness

After finding that the Ribeiro model seems to identify clinically relevant features while exhibiting problems when confronted with artifacts (see Subsection 4.3.3), we took a deeper look at the robustness of the model.

> In Chapter 5, the performance of the model on ECG recordings from PTB-XL annotated with four different types of noise was almost similar to the performance of PTB-XL recordings without these annotations. The presence of noise lead to false positives increased by 0.2% and false negatives by 3.1%, resulting in an F1-score decreased by 0.028. [3]

These results correspond to findings of Venton et al. [26], who reported a decrease in F1-scores of less than 0.05% for noisy data sets. If these findings can be confirmed in further studies, with extensively labeled data sets and other models, DNNs might be especially suited to processing ECG data from clinical settings, where the level of noise is usually higher than in validated data sets such as those from study contexts.

7.4. Generalizability

> In Chapter 6, the analysis of uncurated routine data from the Department of Cardiology at UMG was described. These data were analyzed using the methods described in previous chapters regarding the performance of the Ribeiro model and its explainability, revealing similar results. While the performance is decreased for AF with 26.8% false negatives, DNN features such as the focus on lead V1 and QRS complexes were confirmed.

When extracting the ECG data, multiple standards were considered. These standards for the storage of waveform and especially ECG data are only partly convertible to each other [52]; however, they contribute to reproducibility as given by the Findable, Accessible, Interoperable, and Reproducible (FAIR) principles [78], since they make the data interoperable as well as provide metadata about the signal.

Over the last decades, an increasing number of standards have been developed in the area of biosignals to cover the metadata required for interoperability. Schlögl et al. [79] published an extensive list of biosignal standards in 2010 that could be used for ECGs, promoting the General Data Format (GDF) [80] as the optimal solution.

In 2018, a review of ECG standards named HL7 aECG [81], Standard Communication Protocol for Computer-assisted Electrocardiography (SCP-ECG) [82], DICOM waveforms [10], and International Society for Noninvasive Electrocardiology (ISHNE) [83] as the most popular [52]. All four standards were compared by the authors and found to be almost equal, with only ISHNE falling behind, although it is better suited to long-term measurements.

In this work, the DICOM standard was used for storing ECG data. Since the Schiller devices were already capable of exporting recordings in this format, we did not see the need to change to another standard. Furthermore, all analyses performed were possible with only the data stored in the DICOM files, such as sampling rates. An easy-to-use Python library[3] facilitated the analyses further. Other formats used by the CPSC and PTB-XL databases required more than one file for each ECG recording, and thus, had to be processed and their information integrated before the analyses could start.

The DICOM export from the Schiller servers to an external drive in the secured hospital network was initiated for the first time and executed retrospectively for all available recordings. Unfortunately, the automation of this process has not yet been targeted after sudden personnel changes and shortages in both the MeDIC and the Department of Cardiology, but it was thought possible after initial meetings.

For this reason, the analyses in Chapter 6 could only be performed with the built-in results as ground truth. The integration with electronic health records and other data gathered in the hospital will solve this problem in the near future, as this step is already planned by MeDIC.

However, DL algorithms are currently focusing on ECG waveform data as their only input. Hence, the most important additional information is the label the model should be trained for. In medical scenarios, it is always desirable to have fewer types of data to reach a certain output, especially when invasive procedures are involved, such as the laboratory values used in [12], investigating the relationship between T-wave-based features and serum potassium levels.

Nevertheless, the integration of additional clinical data is still crucial for building cohorts or providing labels for training new models. Therefore, their integration in platforms such as XNAT should be fostered as a basis for further research, thereby providing faster ways to train and evaluate models.

All quantitative XAI analyses performed on this data set obtained similar results as with public datasets. These findings are in conformance with the evaluation of other DNNs trained on clinical ECGs, such as that of Gustafsson et al. [84], who used data

[3] Pydicom, https://pydicom.github.io/

from emergency departments.

Similar to the robustness analyses discussed earlier, XAI analyses exhibit promising results toward explaining DNN decisions on real-world clinical data sets. Although, especially in ante-hoc tasks, it is already possible to make more detailed statements about features learned by DNNs, IG combined with the proposed analyses and visualizations is still feasible in the context of pre-trained models when only post-hoc methods can be applied due to, for example, non-public training data sets.

7.5. Conclusion & Outlook

This work offers insights into the possibilities of using state-of-the-art algorithms with data extracted from clinical settings. On an open source model for 12-lead ECG classification, performance and aspects of trustworthiness were demonstrated with several public data sets and an initial local one. Finally, these findings were confirmed with a large real-world data set acquired in clinical routine.

Along with a rising number of new DL algorithms for ECG classification and prediction tasks each year, the need for post-hoc explainability is also rising, especially in the medical field. To prepare these algorithms for clinical applications, it is not only necessary to include real-world data in training but also to evaluate models on diverse data sets to avoid biases.

As experienced during this work, the process of making routine data available for clinical research poses several challenges. Next to technical decisions such as standards and pseudonymization, regulatory and organizational solutions are required that may slow down the process. Additionally, it could be demonstrated that many aspects of trustworthiness should be considered before using DNNs in clinical settings. Nevertheless, the results regarding performance, reproducibility, explainability, robustness, and generalizability demonstrate that it is possible to foster trust in a DNN for application on routine data from a clinical setting.

In future work, the open source code developed in this work (see Appendix E) will be expanded to enhance the trustworthiness of other neural networks and explore new features learned by DNNs, such as for age prediction on ECGs [85]. Moreover, the code can be integrated into platforms such as XNAT, which allows to analyze ECG recordings in pipelines based on Docker[4] containers. Finally, it is important to include medical researchers as well as clinicians in these next steps toward the integration of DL models into clinical settings.

[4] `https://www.docker.com/`

List of Figures

1.1. Outline of this thesis. 3

2.1. The cardiac cycle and its ECG representation. 6
2.2. Views on the heart with a 12-lead ECG. 7
2.3. Relationship between artificial intelligence, machine learning, and deep learning. 9
2.4. Schema of a perceptron. 9
2.5. The relation between different aspects of AI trustworthiness. 12

3.1. DNN classification results on ECG-AF/LBBB-2021. 22

4.1. Overview of the XAI processing pipeline. 32
4.2. Distribution of Integrated Gradients relevance scores. 33
4.3. Distribution of R_n for each recording as single boxplot. 34
4.4. Distribution of $M_{n,k}$ w.r.t. ECG leads. 35
4.5. Average beats and IG relevance scores for lead V1 in AF classification. . . 36
4.6. Average beats and IG relevance scores for lead aVL in LBBB classification. 37
4.7. SampEn entropy of IG relevance scores for AF classification with regard to lead and label. 37
4.8. Positive and negative relevance scores on correctly classified electrocardiogram. 38
4.9. Positive and negative relevance scores on false negative classified ECG. . . 39
4.10. Integrated Gradients relevance scores for lead V1 averaged over 200 ECGs extracted from CPSC and PTB-XL. 39

4.11. Relevance scores of LBBB recordings from CPSC database compared to PTB-XL data. 40
4.12. Relevance scores calculated with five XAI methods on lead V6 of a correctly classified electrocardiogram. 40
4.13. Average beats and relevance scores for individual leads in a single normal recording. 42

5.1. Distribution of the values of both SNR metrics, with and without noise labels. 47

6.1. DNN classification results on ECG-full-2021. 50
6.2. Distribution of R_n for each LBBB and RBBB recording as single boxplot (ECG-full-2021). 51
6.3. Distribution of $M_{n,k}$ computed with IG w.r.t. ECG leads (ECG-full-2021). 52
6.4. Average beats and IG relevance scores for lead aVL in LBBB classification (ECG-full-2021). 53

List of Tables

3.1. Comparison of DL model results including F1-scores. 22

5.1. Properties of subset extracted from PTB-XL and results of DL-based AF classification. 46

5.2. DNN accuracy w.r.t. four types of noise. 48

6.1. Comparison of DL model results including F1-scores for ECG-full-2021. . 50

List of Listings

3.1. Preprocessing of ECG-AF/LBBB-2021. Each recording *d* in the loaded data set *ds* is resampled and zero-padded. Afterwards, *x* contains all preprocessed recordings, according to the format which is required by the used model. 20

4.1. Preprocessing of CPSC data. Each recording in the index list *idx_x* is reshaped, resampled and zero-padded, and returned in variable *x*. 26

4.2. Loading the Ribeiro model for classification of six ECG abnormalities. The function returns the original model, as well as a copy with a linear instead of sigmoid activation in the last layer. 27

4.3. Calculating relevance scores of data set *x* with both models returned by *loadmodel()* (Listing 4.2). XAI method and output neuron (class) are given as additional variables. 28

4.4. Calculating the sample entropy of IG *relevances* for each recording *i*. . . . 29

4.5. Calculation of the average beat with average relevance scores for a specific *lead* of a single ECG recording (*raw_data*). The first and last beat must further be zero-padded if they would overlap with the full *duration*. 30

5.1. Calculation of the measured SNR in dB, with the signal frequency band defined between 40 and 150 bpm. 46

Glossary

1dAVb first-degree atrioventricular block.

aECG Annotated Electrocardiogram.

AF atrial fibrillation.

AI Artificial Intelligence.

CNN convolutional neural network.

DICOM Digital Imaging and Communications in Medicine.

DL Deep Learning.

DNN deep neural network.

ECG electrocardiogram.

EDF European Data Format.

FAIR Findable, Accessible, Interoperable, and Reproducible.

GDF General Data Format.

HL7 Health Level 7.

IG Integrated Gradients.

ISHNE International Society for Noninvasive Electrocardiology.

LBBB left bundle branch block.

LRP Layer-wise Relevance Propagation.

MDR Medical Device Regulation (EU) 2017/745.

MeDIC medical data integration center.

ML machine learning.

RBBB right bundle branch block.

ReLU rectified linear unit.

SCP-ECG Standard Communication Protocol for Computer-assisted Electrocardiography.

SNR signal-to-noise ratio.

SR sinus rhythm.

SVM support vector machine.

UMG University Medical Center Göttingen.

XAI explainable AI.

References (Own Publications)

[1] T. Bender et al. “Application of Pre-Trained Deep Learning Models for Clinical ECGs”. In: *Studies in health technology and informatics* 283 (2021), pp. 39–45. DOI: 10.3233/SHTI210539.

[2] T. Bender et al. “Analysis of a Deep Learning Model for 12-Lead ECG Classification Reveals Learned Features Similar to Diagnostic Criteria”. In: *IEEE journal of biomedical and health informatics* PP (2023). ISSN: 2168-2194. DOI: 10.1109/JBHI.2023.3271858.

[3] T. Bender et al. “Benchmarking the Impact of Noise on Deep Learning-Based Classification of Atrial Fibrillation in 12-Lead ECG”. In: *Studies in health technology and informatics* 302 (2023), pp. 977–981. DOI: 10.3233/SHTI230321.

References

[4] G. Hindricks et al. “2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS)The Task Force for the diagnosis and management of atrial fibrillation of the European Society of Cardiology (ESC) Developed with the special contribution of the European Heart Rhythm Association (EHRA) of the ESC”. In: *European Heart Journal* 42.5 (2021), pp. 373–498. ISSN: 0195-668X. DOI: 10.1093/eurheartj/ehaa612.

[5] S. Raj and K. C. Ray. “A Personalized Arrhythmia Monitoring Platform”. In: *Scientific reports* 8.1 (2018), p. 11395. DOI: 10.1038/s41598-018-29690-2.

[6] M. Shao et al. “A Wearable Electrocardiogram Telemonitoring System for Atrial Fibrillation Detection”. In: *Sensors (Basel, Switzerland)* 20.3 (2020). DOI: 10.3390/s20030606.

[7] T. Khatibi and N. Rabinezhadsadatmahaleh. “Proposing feature engineering method based on deep learning and K-NNs for ECG beat classification and arrhythmia detection”. In: *Australasian physical & engineering sciences in medicine* (2019). DOI: 10.1007/s13246-019-00814-w.

[8] S. W. E. Baalman et al. “A morphology based deep learning model for atrial fibrillation detection using single cycle electrocardiographic samples”. In: *International journal of cardiology* 316 (2020), pp. 130–136. DOI: 10.1016/j.ijcard.2020.04.046.

[9] B. Kemp and J. Olivan. “European data format ‘plus’ (EDF+), an EDF alike standard format for the exchange of physiological data”. In: *Clinical Neurophysiology* 114.9 (2003), pp. 1755–1761. ISSN: 13882457. DOI: 10.1016/S1388-2457(03)00123-8.

[10] T. Hilbel et al. “Innovation and advantage of the DICOM ECG standard for viewing, interchange and permanent archiving of the diagnostic electrocardiogram”. In: *Computers in cardiology, 2007*. Ed. by A. Murray. Piscataway, NJ: IEEE, 2007, pp. 633–636. ISBN: 978-1-4244-2533-4. DOI: 10.1109/CIC.2007.4745565.

[11] M. Escabí. “Chapter 11 - Biosignal Processing”. In: *Introduction to biomedical engineering*. Ed. by J. D. Enderle and J. D. Bronzino. Academic Press series in biomedical engineering. Amsterdam: Elsevier Acad. Press, 2012, pp. 667–746. ISBN: 978-0-12-374979-6. DOI: 10.1016/B978-0-12-374979-6.00011-3.

[12] D. Yoon et al. “Quantitative Evaluation of the Relationship between T-Wave-Based Features and Serum Potassium Level in Real-World Clinical Practice”. In: *BioMed research international* 2018 (2018), p. 3054316. DOI: 10.1155/2018/3054316.

[13] S. Hong et al. “Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review”. In: *Computers in biology and medicine* 122 (2020), p. 103801. DOI: 10.1016/j.compbiomed.2020.103801.

[14] D. Yoon et al. “Discovering hidden information in biosignals from patients using artificial intelligence”. In: *Korean journal of anesthesiology* 73.4 (2020), pp. 275–284. DOI: 10.4097/kja.19475.

[15] B. Li et al. “Trustworthy AI: From Principles to Practices”. In: *ACM Computing Surveys* 55.9 (2023), pp. 1–46. ISSN: 0360-0300. DOI: 10.1145/3555803.

[16] A. L. Beam, A. K. Manrai, and M. Ghassemi. “Challenges to the Reproducibility of Machine Learning Models in Health Care”. In: *JAMA* 323.4 (2020), p. 305. DOI: 10.1001/jama.2019.20866.

[17] S. N. Goodman, D. Fanelli, and J. P. A. Ioannidis. “What does research reproducibility mean?” In: *Science translational medicine* 8.341 (2016), 341ps12. DOI: 10.1126/scitranslmed.aaf5027.

[18] R. Guidotti et al. “A Survey of Methods for Explaining Black Box Models”. In: *ACM Computing Surveys* 51.5 (2019), pp. 1–42. ISSN: 0360-0300. DOI: 10.1145/3236009.

[19] H. Taniguchi et al. “Explainable Artificial Intelligence Model for Diagnosis of Atrial Fibrillation Using Holter Electrocardiogram Waveforms”. In: *International heart journal* 62.3 (2021), pp. 534–539. DOI: 10.1536/ihj.21-094.

[20] M. Bodini, M. W. Rivolta, and R. Sassi. "Opening the black box: interpretability of machine learning algorithms in electrocardiography". In: *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* 379.2212 (2021), p. 20200253. DOI: `10.1098/rsta.2020.0253`.

[21] I. Sturm et al. "Interpretable deep neural networks for single-trial EEG classification". In: *Journal of neuroscience methods* 274 (2016), pp. 141–145. DOI: `10.1016/j.jneumeth.2016.10.008`.

[22] C. J. Kelly et al. "Key challenges for delivering clinical impact with artificial intelligence". In: *BMC medicine* 17.1 (2019), p. 195. DOI: `\url{10.1186/s12916-019-1426-2}`.

[23] S. Festag and C. Spreckelsen. "Semantic Anomaly Detection in Medical Time Series". In: *Studies in health technology and informatics* 278 (2021), pp. 118–125. DOI: `10.3233/SHTI210059`.

[24] Z. F. M. Apandi et al. "An Analysis of the Effects of Noisy Electrocardiogram Signal on Heartbeat Detection Performance". In: *Bioengineering (Basel, Switzerland)* 7.2 (2020). ISSN: 2306-5354. DOI: `10.3390/bioengineering7020053`.

[25] P. Kumar and V. K. Sharma. "Detection and classification of ECG noises using decomposition on mixed codebook for quality analysis". In: *Healthcare technology letters* 7.1 (2020), pp. 18–24. ISSN: 2053-3713. DOI: `10.1049/htl.2019.0096`.

[26] J. Venton et al. "Robustness of convolutional neural networks to physiological electrocardiogram noise". In: *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* 379.2212 (2021), p. 20200262. DOI: `10.1098/rsta.2020.0262`.

[27] H.-J. Trappe and H.-P. Schuster. *EKG-Kurs für Isabel.* Stuttgart: Georg Thieme Verlag, 2017. ISBN: 9783132407992. DOI: `10.1055/b-005-143650`.

[28] K. Harris, D. Edwards, and J. Mant. "How can we best detect atrial fibrillation?" In: *The journal of the Royal College of Physicians of Edinburgh* 42 Suppl 18 (2012), pp. 5–22. DOI: `10.4997/JRCPE.2012.S02`.

[29] A. Bollmann et al. "Analysis of surface electrocardiograms in atrial fibrillation: techniques, research, and clinical applications". In: *Europace : European pacing, arrhythmias, and cardiac electrophysiology : journal of the working groups on cardiac pacing, arrhythmias, and cardiac cellular electrophysiology of the European Society of Cardiology* 8.11 (2006), pp. 911–926. ISSN: 1099-5129. DOI: `10.1093/europace/eul113`.

[30] N. Y. Tan et al. "Left Bundle Branch Block: Current and Future Perspectives". In: *Circulation. Arrhythmia and electrophysiology* 13.4 (2020), e008239. DOI: `10.1161/CIRCEP.119.008239`.

[31] D. G. Strauss, R. H. Selvester, and G. S. Wagner. "Defining left bundle branch block in the era of cardiac resynchronization therapy". In: *The American journal of cardiology* 107.6 (2011), pp. 927–934. DOI: `10.1016/j.amjcard.2010.11.010`.

[32] Y. LeCun, Y. Bengio, and G. Hinton. "Deep learning". In: *Nature* 521.7553 (2015), pp. 436–444. DOI: `10.1038/nature14539`.

[33] F. Rosenblatt. "The perceptron: A probabilistic model for information storage and organization in the brain". In: *Psychological Review* 65.6 (1958), pp. 386–408. ISSN: 0033-295X. DOI: `10.1037/h0042519`.

[34] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri. "Activation functions in deep learning: A comprehensive survey and benchmark". In: *Neurocomputing* 503 (2022), pp. 92–108. ISSN: 09252312. DOI: `10.1016/j.neucom.2022.06.111`.

[35] K. He et al. "Deep Residual Learning for Image Recognition". In: *29th IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE, 2016, pp. 770–778. ISBN: 978-1-4673-8851-1. DOI: `10.1109/CVPR.2016.90`.

[36] G. Petmezas et al. "State-of-the-Art Deep Learning Methods on Electrocardiogram Data: Systematic Review". In: *JMIR medical informatics* 10.8 (2022), e38454. ISSN: 2291-9694. DOI: `10.2196/38454`.

[37] A. H. Ribeiro et al. "Automatic diagnosis of the 12-lead ECG using a deep neural network". In: *Nature communications* 11.1 (2020), p. 1760. DOI: `10.1038/s41467-020-15432-4`.

[38] A. H. Ribeiro et al. *Pre-trained deep neural network models for ECG automatic abnormality detection*. 2020. DOI: `10.5281/zenodo.3765717`.

[39] A. H. Ribeiro et al. *Annotated 12-lead ECG dataset*. 2020. DOI: `10.5281/zenodo.3765780`.

[40] S. Bach et al. "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation". In: *PloS one* 10.7 (2015), e0130140. DOI: `10.1371/journal.pone.0130140`.

[41] M. Sundararajan, A. Taly, and Q. Yan. "Axiomatic Attribution for Deep Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 3319–3328.

[42] Q. Hu et al. *X-MyoNET: Biometric Identification using Deep Processing of Transient Surface Electromyography.* 2021. DOI: 10.1101/2021.11.30.470688.

[43] M. Doborjeh et al. “Deep Learning of Explainable EEG Patterns as Dynamic Spatiotemporal Clusters and Rules in a Brain-Inspired Spiking Neural Network”. In: *Sensors (Basel, Switzerland)* 21.14 (2021). DOI: 10.3390/s21144900.

[44] Y. Elul et al. “Meeting the unmet needs of clinicians from AI systems showcased for cardiology with deep-learning-based ECG analysis”. In: *Proceedings of the National Academy of Sciences of the United States of America* 118.24 (2021). DOI: 10.1073/pnas.2020620118.

[45] M. D. Zeiler and R. Fergus. “Visualizing and Understanding Convolutional Networks”. In: *Computer vision - ECCV 2014.* Ed. by D. Fleet et al. Vol. 8689. Lecture Notes in Computer Science. Cham: Springer, 2014, pp. 818–833. ISBN: 978-3-319-10589-5. DOI: 10.1007/978-3-319-10590-1_53.

[46] L. M. Zintgraf et al. *Visualizing Deep Neural Network Decisions: Prediction Difference Analysis.* 2017. DOI: 10.48550/arXiv.1702.04595.

[47] W. Samek et al. *Evaluating the visualization of what a Deep Neural Network has learned.* 2015. DOI: 10.48550/arXiv.1509.06321.

[48] G. Montavon et al. “Explaining nonlinear classification decisions with deep Taylor decomposition”. In: *Pattern Recognition* 65 (2017), pp. 211–222. ISSN: 00313203. DOI: \url{10.1016/j.patcog.2016.11.008}.

[49] F. Liu et al. “An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection”. In: *Journal of Medical Imaging and Health Informatics* 8.7 (2018), pp. 1368–1373. ISSN: 2156-7018. DOI: 10.1166/jmihi.2018.2442.

[50] P. Wagner et al. “PTB-XL, a large publicly available electrocardiography dataset”. In: *Scientific data* 7.1 (2020), p. 154. DOI: 10.1038/s41597-020-0495-6.

[51] S. M. Meystre et al. “Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress”. In: *Yearbook of medical informatics* 26.1 (2017), pp. 38–52. DOI: 10.15265/IY-2017-007.

[52] D. Stamenov, M. Gusev, and G. Armenski. “Interoperability of ECG standards”. In: *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO).* Piscataway, NJ: IEEE, 2018, pp. 0319–0323. ISBN: 978-953-233-095-3. DOI: 10.23919/MIPRO.2018.8400061.

[53] M. Parciak et al. "FAIRness through automation: development of an automated medical data integration infrastructure for FAIR health data in a maximum care university hospital". In: *BMC medical informatics and decision making* 23.1 (2023), p. 94. DOI: `10.1186/s12911-023-02195-3`.

[54] M. Beier et al. "Multicenter data sharing for collaboration in sleep medicine". In: *Future Generation Computer Systems* 67 (2017), pp. 466–480. ISSN: 0167739X. DOI: `10.1016/j.future.2016.03.025`.

[55] D. Krefting et al. "Reliability of quantitative neuroimage analysis using FreeSurfer in distributed environments". In: *HP-MICCAI/MICCAI-DCI 2011 workshop*. 2011, p. 10.

[56] R. J. LeVeque, I. M. Mitchell, and V. Stodden. "Reproducible research for scientific computing: Tools and strategies for changing the culture". In: *Computing in science & engineering* 14.4 (2012), pp. 13–17. ISSN: 1521-9615. DOI: `10.1109/MCSE.2012.38`.

[57] C. Jansen et al. "Curious Containers: A framework for computational reproducibility in life sciences with support for Deep Learning applications". In: *Future Generation Computer Systems* 112 (2020), pp. 209–227. ISSN: 0167739X. DOI: `10.1016/j.future.2020.05.007`.

[58] M. Alber et al. "iNNvestigate Neural Networks!" In: *Journal of Machine Learning Research* 20.93 (2019), pp. 1–8.

[59] M. Sundararajan, A. Taly, and Q. Yan. "Axiomatic Attribution for Deep Networks". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. JMLR.org, 2017, pp. 3319–3328.

[60] P. S. Hamilton and W. J. Tompkins. "Compression of the ambulatory ECG by average beat subtraction and residual differencing". In: *IEEE transactions on biomedical engineering* 38.3 (1991), pp. 253–259. DOI: `10.1109/10.133206`.

[61] P. Langley, J. P. Bourke, and A. Murray. "Frequency analysis of atrial fibrillation". In: *2000 Computers in Cardiology*. Piscataway: I E E E, Nov. 2000, pp. 65–68. ISBN: 0-7803-6557-7. DOI: `10.1109/CIC.2000.898456`.

[62] N. Strodthoff et al. *Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL*. 2020. DOI: `10.48550/arXiv.2004.13701`.

[63] G. de Haan and V. Jeanne. "Robust pulse rate from chrominance-based rPPG". In: *IEEE transactions on bio-medical engineering* 60.10 (2013), pp. 2878–2886. DOI: `10.1109/TBME.2013.2266196`.

[64] C. A. Ellis et al. “Examining Reproducibility of EEG Schizophrenia Biomarkers Across Explainable Machine Learning Models”. In: *2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering (BIBE).* IEEE, 2022, pp. 305–308. ISBN: 978-1-6654-8487-9. DOI: 10.1109/BIBE55377.2022.00069.

[65] A. Nebli et al. “Quantifying the reproducibility of graph neural networks using multigraph data representation”. In: *Neural networks : the official journal of the International Neural Network Society* 148 (2022), pp. 254–265. DOI: 10.1016/j.neunet.2022.01.018.

[66] T. J. Loftus et al. “Ideal algorithms in healthcare: Explainable, dynamic, precise, autonomous, fair, and reproducible”. In: *PLOS digital health* 1.1 (2022). DOI: 10.1371/journal.pdig.0000006.

[67] X. Liu et al. “Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension”. In: *Nature medicine* 26.9 (2020), pp. 1364–1374. DOI: 10.1038/s41591-020-1034-x.

[68] M. Hutson. “Artificial intelligence faces reproducibility crisis”. In: *Science (New York, N.Y.)* 359.6377 (2018), pp. 725–726. DOI: \url{10.1126/science.359.6377.725}.

[69] B. Chen et al. “Towards training reproducible deep learning models”. In: *Proceedings of the 44th International Conference on Software Engineering.* Ed. by M. B. Dwyer, D. Damian, and A. Zeller. New York, NY, USA: ACM, 2022, pp. 2202–2214. ISBN: 9781450392211. DOI: 10.1145/3510003.3510163.

[70] S. W. Smith et al. “A deep neural network learning algorithm outperforms a conventional algorithm for emergency department electrocardiogram interpretation”. In: *Journal of Electrocardiology* 52 (2019), pp. 88–95. ISSN: 00220736. DOI: 10.1016/j.jelectrocard.2018.11.013.

[71] D. Parekh et al. “A Review on Autonomous Vehicles: Progress, Methods and Challenges”. In: *Electronics* 11.14 (2022), p. 2162. DOI: 10.3390/electronics11142162.

[72] A. J. DeGrave, J. D. Janizek, and S.-I. Lee. “AI for radiographic COVID-19 detection selects shortcuts over signal”. In: *Nature Machine Intelligence* 3.7 (2021), pp. 610–619. DOI: \url{10.1038/s42256-021-00338-7}.

[73] *Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on Medical Devices, Amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and Repealing Council Directives 90/385/EEC and 93/42/EEC.*

[74] T. A. Abdullah et al. “Explainable Deep Learning Model for Cardiac Arrhythmia Classification”. In: *2022 International Conference on Future Trends in Smart Communities (ICFTSC)*. IEEE, 2022, pp. 87–92. ISBN: 979-8-3503-3454-8. DOI: 10.1109/ICFTSC57269.2022.10039860.

[75] S. van Baalen, M. Boon, and P. Verhoef. “From clinical decision support to clinical reasoning support systems”. In: *Journal of evaluation in clinical practice* (2021). DOI: 10.1111/jep.13541.

[76] P. Wagner et al. *Explaining Deep Learning for ECG Analysis: Building Blocks for Auditing and Knowledge Discovery.* 2023. DOI: 10.48550/arXiv.2305.17043.

[77] D. A. Ehrlich et al. “A Measure of the Complexity of Neural Representations based on Partial Information Decomposition”. In: (2022). DOI: 10.48550/arXiv.2209.10438.

[78] M. D. Wilkinson et al. “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific data* 3 (2016), p. 160018. DOI: 10.1038/sdata.2016.18.

[79] A. Schlögl. “An overview on data formats for biomedical signals”. In: *World Congress on Medical Physics and Biomedical Engineering, September 7 - 12, 2009, Munich, Germany.* Ed. by O. Dössel and W. C. Schlegel. IFMBE Proceedings. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg, 2010, pp. 1557–1560. ISBN: 9783642038815.

[80] A. Schlögl. *GDF - A general dataformat for biosignals.* 2006. DOI: 10.48550/arXiv.cs/0608052.

[81] B. D. Brown and F. Badilini. *aECG Implementation Guide.* 2005. URL: https://www.amps-llc.com/uploads/2017-12-7/aECG_Implementation_Guide(1).pdf (visited on 09/21/2023).

[82] P. Rubel et al. “SCP-ECG V3.0: An enhanced standard communication protocol for computer-assisted electrocardiography”. In: *2016 Computing in Cardiology Conference (CinC).* 2016, pp. 309–312. ISBN: 2325-887X.

[83] F. Badilini. “The ISHNE Holter Standard Output File Format”. In: *Annals of Noninvasive Electrocardiology* 3.3 (1998), pp. 263–266. ISSN: 1082-720X. DOI: 10.1111/j.1542-474X.1998.tb00353.x.

[84] S. Gustafsson et al. “Development and validation of deep learning ECG-based prediction of myocardial infarction in emergency department patients”. In: *Scientific reports* 12.1 (2022), p. 19615. DOI: 10.1038/s41598-022-24254-x.

[85] E. M. Lima et al. "Deep neural network-estimated electrocardiographic age as a mortality predictor". In: *Nature communications* 12.1 (2021), p. 5117. DOI: `\url{10.1038/s41467-021-25351-7}`.

Appendices

APPENDIX A

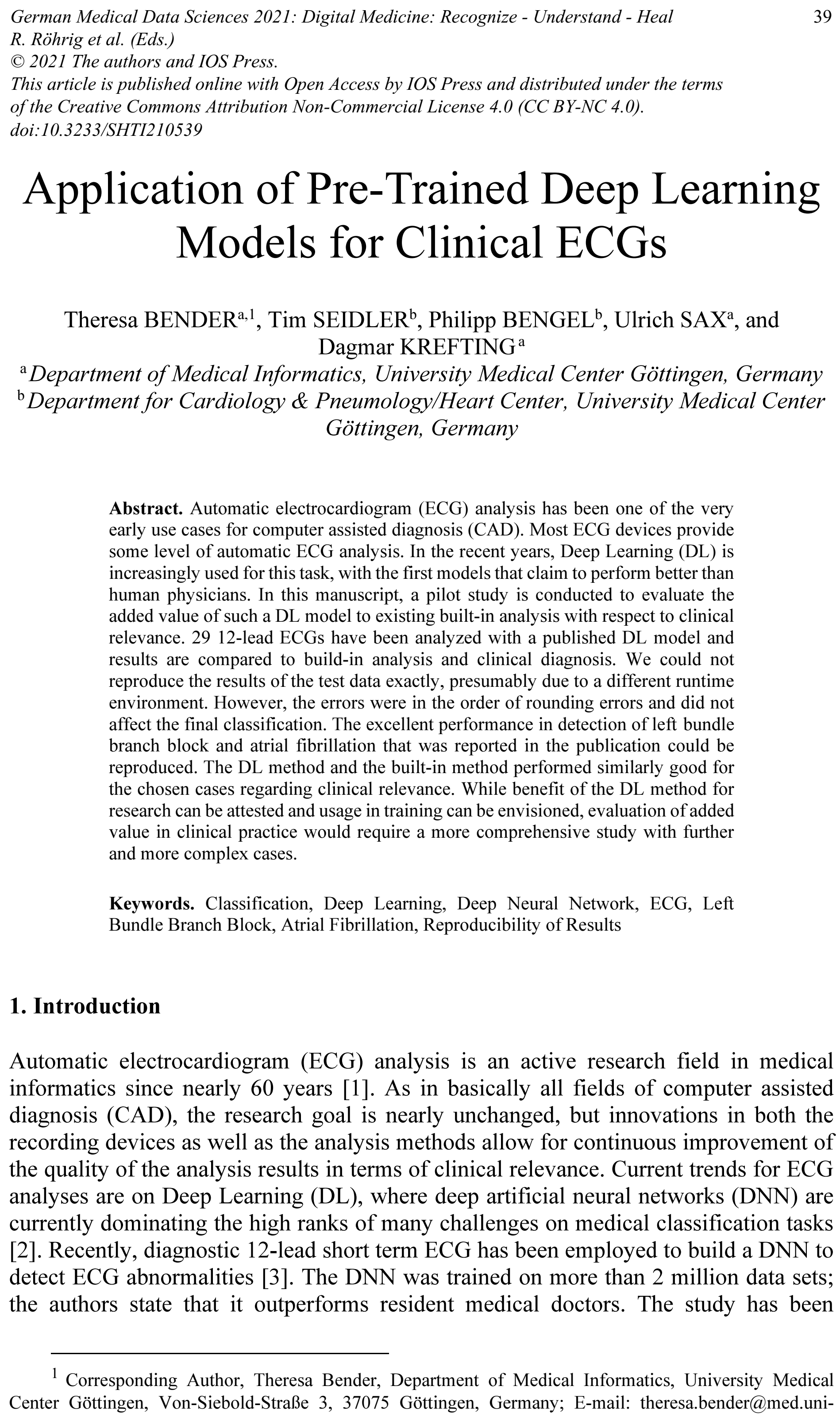
Article A

Bender T, Seidler T, Bengel P, Sax U, Krefting D. Application of Pre-Trained Deep Learning Models for Clinical ECGs. Stud Health Technol Inform 2021; 283:39–45, doi: 10.3233/SHTI210539. CC BY-NC[1]

Author Contributions

I conceptualized and implemented the software, resulting in Figure 1 and Table 1 and all sections. I prepared the original draft and worked in the reviewer feedback. I prepared and held the online conference talk at the 66th Annual Meeting of the German Association of Medical Informatics, Biometry, and Epidemiology e.V. (GMDS) 2021.

[1] `https://creativecommons.org/licenses/by-nc/4.0/deed.en_US`



German Medical Data Sciences 2021: Digital Medicine: Recognize - Understand - Heal
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI210539

Application of Pre-Trained Deep Learning Models for Clinical ECGs

Theresa BENDER[a,1], Tim SEIDLER[b], Philipp BENGEL[b], Ulrich SAX[a], and
Dagmar KREFTING[a]
[a]*Department of Medical Informatics, University Medical Center Göttingen, Germany*
[b]*Department for Cardiology & Pneumology/Heart Center, University Medical Center Göttingen, Germany*

Abstract. Automatic electrocardiogram (ECG) analysis has been one of the very early use cases for computer assisted diagnosis (CAD). Most ECG devices provide some level of automatic ECG analysis. In the recent years, Deep Learning (DL) is increasingly used for this task, with the first models that claim to perform better than human physicians. In this manuscript, a pilot study is conducted to evaluate the added value of such a DL model to existing built-in analysis with respect to clinical relevance. 29 12-lead ECGs have been analyzed with a published DL model and results are compared to build-in analysis and clinical diagnosis. We could not reproduce the results of the test data exactly, presumably due to a different runtime environment. However, the errors were in the order of rounding errors and did not affect the final classification. The excellent performance in detection of left bundle branch block and atrial fibrillation that was reported in the publication could be reproduced. The DL method and the built-in method performed similarly good for the chosen cases regarding clinical relevance. While benefit of the DL method for research can be attested and usage in training can be envisioned, evaluation of added value in clinical practice would require a more comprehensive study with further and more complex cases.

Keywords. Classification, Deep Learning, Deep Neural Network, ECG, Left Bundle Branch Block, Atrial Fibrillation, Reproducibility of Results

1. Introduction

Automatic electrocardiogram (ECG) analysis is an active research field in medical informatics since nearly 60 years [1]. As in basically all fields of computer assisted diagnosis (CAD), the research goal is nearly unchanged, but innovations in both the recording devices as well as the analysis methods allow for continuous improvement of the quality of the analysis results in terms of clinical relevance. Current trends for ECG analyses are on Deep Learning (DL), where deep artificial neural networks (DNN) are currently dominating the high ranks of many challenges on medical classification tasks [2]. Recently, diagnostic 12-lead short term ECG has been employed to build a DNN to detect ECG abnormalities [3]. The DNN was trained on more than 2 million data sets; the authors state that it outperforms resident medical doctors. The study has been

[1] Corresponding Author, Theresa Bender, Department of Medical Informatics, University Medical Center Göttingen, Von-Siebold-Straße 3, 37075 Göttingen, Germany; E-mail: theresa.bender@med.uni-goettingen.de.

conducted in Brazil within a large telemedicine network. However, the question remains how these results translate to other countries and to other settings such as inpatient care. In particular, we were wondering if the application of this model would bring added value to our University Medical Center by supporting research, training and health care.

Closely connected to new methods of data driven analysis, good scientific practice is focusing more and more towards open science to allow reproducibility and transparency. In this sense, trained models and test data should be published alongside with the description of the model architectures in publications. This facilitates reproducibility studies, but still the correct usage of the described methods might be difficult if the runtime environment differs due to critical deviation in any of the components of the employed software or hardware stack [4].

Therefore, for successful implementation of the CAD we need to address both aspects of reproducibility, following the definition of Goodman et al. [5]:

a) *methods reproducibility:* will the same data sets result in the same output of the DL model?
b) *results reproducibility:* will the DL model bring added value in another environment (other data, other physicians and another healthcare system)?

In this paper, both aspects of reproducibility are addressed. The before mentioned DNN-model for the automatic detection of certain cardiovascular diseases on 12-lead electrocardiogram data were applied to pseudonymized ECGs of patients of the department of cardiology of the University Medical Center. The classification results are compared with the clinical diagnosis and the automatic built-in analysis of the ECG device. While methods reproducibility can easily be assessed quantitatively, the evaluation of added value is much more complex and will only be assessed superficially within this manuscript.

2. Methods

Ribeiro et al. developed a Deep Learning model for automatic classification of six cardiac disorders, among them left bundle branch block (LBBB) and atrial fibrillation (AF), for details c.f. [3]. The pre-trained model is archived and published, as well as the used test data [6,7]. For *methods reproducibility*, we checked for metadata on the runtime environment settings in the original paper, the Zenodo repository and the corresponding source-code repository [2]. The model has been implemented in the university's JupyterHub and has been executed on the provided test data. As the model outputs probabilities, the used thresholds are required to reproduce the results and evaluate possible differences. As they are not explicitly given in the paper, threshold values found in a current version of the code are used (generate_figures_and_tables.py).

For assessing the *results reproducibility*, 15 patients with diagnosed LBBB and 20 patients with diagnosed AF have been selected by a clinical expert based on the printed ECG reports. The two disorders have been selected based on clinical relevance and the fact that the respective ECG abnormalities are characteristic and present in the ECG when clinically diagnosed. From these data, five patients with LBBB and one patient with AF have been removed, as the digital ECG was no longer available. For the remaining patients, ECGs have been pseudonymized using the diagnosis and a

[2] https://github.com/antonior92/automatic-ecg-diagnosis

consecutive number as code, and have been exported into DICOM format. The DICOM headers have been checked for possible identifying data in private tags. The data was loaded with the program library *pydicom* in version 2.1.2[3].

The data have been resampled using the Scify function *resample_poly*[4] to fulfill requirements on sampling rate (400 Hz) and have been padded to the sample number of 4096 by appending zeros. Furthermore, the data was rescaled from microVolt to milliVolt by division by 1000.

The classification results from the DNN are compared with the actual diagnosis as well as with the corresponding built-in automatic annotation. To get an impression about the overall classification results in all six categories on these patients, the distribution of the class probabilities are shown for the two patient groups. The model performance is assessed by sensitivity (recall) and specificity, precision and F1-score, following the original publication of the DNN. Here the subjects suffering from the respective other disorder have been used as negative samples.

Diverging results in either DNN classification, built-in classification or diagnosis results were finally evaluated by a cardiologist regarding clinical soundness and relevance.

3. Results

3.1. Methods reproducibility

There is no meta data description in the journal article [3], but it refers to the open source repository on GitHub, where library-versions are given in a specific requirements file, containing all Python libraries used. There is no marked release that would indicate the actual version used for the paper, and there have been seven updates meanwhile. But, as supplementary material has been uploaded on May 1, 2020, the first submission is assumed to be the environment settings for the published results. The supplements itself also refer to the GitHub repository and do not contain any further meta data. We could not find any information about the employed operating system or hardware environment. There is some confusion about the employed TensorFlow version. It seems that the authors have used version 2.2, but have downgraded to version 1.15. However, it is not discernible whether the version switch has been performed before or after model training.

In our environment that uses the latest versions by default, basically all Python libraries have been updated since original publication of the model.

Applying the model on the original test data of 827 ECGs produced similar, but not equal results. The comparison with the reported abnormality probabilities by the authors showed differences in about 88% of the values (4381). Interestingly, a switch from TensorFlow 2.3.1 to 2.2 resulted in one more value differing. However, differences are in the order of rounding errors in floating point values, i.e. ~1e-7. When compared to class-thresholds, none of these differences resulted in a different classification.

[3] https://pydicom.github.io

[4] https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.resample_poly.html

3.2. Results reproducibility

The classification results from the DL method are summarized in Figure 1. For an overview of F1-scores elaborated in this section cf. Table 1.

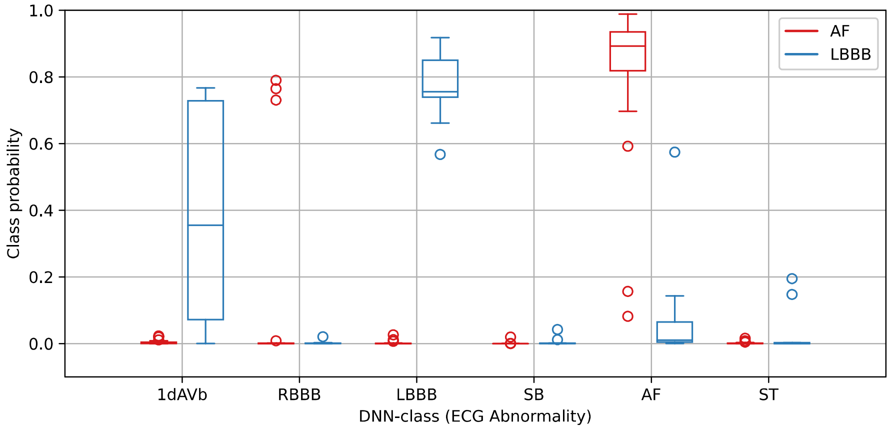


Figure 1. DNN-classification on local data set. Separated into cohorts diagnosed with either atrial fibrillation (red, n=19) or left bundle branch block (blue, n=10), Considered abnormalities are: 1st degree AV block (1dAVb), right bundle branch block (RBBB), left bundle branch block (LBBB), sinus bradycardia (SB), atrial fibrillation (AF) and sinus tachycardia (ST).

Table 1. Comparison of F1-scores. Calculated for both left bundle branch block (LBBB) and atrial fibrillation (AF). Considered analyses: Built-In algorithm of ECG devices as well as DNN-classification on the same data set, in addition to scores published by Ribeiro et al. [3].

	Built-In				**DNN (local data set)**				**DNN (Ribeiro et al.)**			
	P	**FP**	**FN**	**F1-Score**	**P**	**FP**	**FN**	**F1-Score**	**P**	**FP**	**FN**	**F1-Score**
LBBB	10	0	1	0.947	10	0	0	1.000	30	0	0	1.000
AF	19	0	0	1.000	19	1	2	0.919	13	0	3	0.870

For the 19 AF patients, 17 were classified correctly with AF with probabilities highly above the classification threshold of 0.390. Two patients have not been classified, with probabilities below 0.2 and therefore not close to the threshold. For the AF patients, three subjects have been additionally classified with RBBB abnormality with a probability of almost 0.8. 1dAVb has been detected for seven of the LBBB patients, with a minimum probability of 0.27 considerably higher than the threshold of 0.124. These findings have been confirmed.

One LBBB patient has also been clearly classified with AF - with a probability of ~0.6. This finding has been identified as false positive. For LBBB, we could reproduce the “perfect detection” with an F1-score of 1, for AF the F1-score of 0.919 is even higher than the published score of 0.870.

Comparison with the built-in method shows large agreement in the findings: For LBBB the built-in method detected eight cases and one subject was annotated as “unspecific interventricular block“, which has been confirmed as clinically equivalent to LBBB. One patient however has not been detected as LBBB. AF has been detected for all 19 ECGs from AF patients. One ECG has been annotated with “irregular rhythm, no p-wave detected”, which has also been confirmed to be equivalent to AF diagnosis. Here, the built-in method showed a better performance for AF-detection (F1-Score: 1), but a lower F1-score for LBBB (0.947) due to the one missed LBBB diagnosis.

The three cases of RBBB have also been detected by the built-in method, and additionally an “incomplete RBBB”. The 1dAVb classification has also been annotated for five patients by the built-in method. However, as the built-in method has more fine-grained annotations (in total, about 50 different annotations were found in the data set, including different probability levels of an abnormality), quantitative comparison is not straightforward but would require mapping of the larger value set to the 6 classes.

4. Discussion

The DL method proposed by [3] could not be fully reproduced numerically, differences in the class probabilities were found in the order of rounding errors. Different rounding methods are known to affect reproducibility of numerical methods, when mathematical system libraries are used rather than static libraries [8]. Interestingly, different TensorFlow versions produced slightly different results in the otherwise identical runtime environment. This might be due to different mathematical optimization procedures. These issues can be avoided by container-based provision of the method, or at least full description of the meta data [9,10]. We would like to state that many important information was provided in the source code by the authors, only information about the Python version and the used operating system would have been required for full *methods reproducibility*. However, in our data, the predicted class probability was always clearly above or below the threshold, so the numerical differences did not have any influence on the classification results.

To improve methods reproducibility, better handling of research results is required. While the FAIR guiding principles are widely recognized and are increasingly required to be addressed in grant applications, they are mainly applied only for the data. In the original publication by Wilkinson et al. it is explicitly stated that the principles apply not only to ‘data’ in the conventional sense, but also to the algorithms, tools, and workflows that led to that data [11]. We strongly support this statement. However few metadata standards - such as the common workflow language - are yet available for the description of code and processing, and to our knowledge there is no common standard for the description of the runtime environment [10,12]. But simple measures such as tagged releases of source code versions and build-files for containers (Docker files) used for a publication can easily increase the FAIRness of a research result. We suggest that aspects of code and processing handling should also be an integral part of a study’s data management plans.

The *results reproducibility* for our data set is excellent, the performance parameters could be reproduced with our data, although data had to be downsampled, padded and rescaled. While the DL method did not perform better on our data than the built-in method, due to its free availability and applicability to data from different devices, it provides definitive added value at least for multi-center studies where heterogeneous ECGs are required to be analysed consistently. Furthermore, due to its good performance, it could be implemented in self-training modules on ECG analysis for medical students. Added value to the clinical routine is not so clear, as the built-in method was comparable, while offering more classes like left anterior fascicular block or annotations about changes caused by ischemia as well as passed infarcts.

Limitations of our pilot study are a relative low number of samples and missing healthy controls. It should be noted that the built-in method typically has a high sensitivity for AF, so all selected ECGs also had been annotated accordingly by the built-in method, which might be a bias. Therefore, results should be taken with care and should be seen as a first step in a closer evaluation of the method.

In conclusion, benefit of the DL method for research can be attested and usage in training can be envisioned. But an evaluation of added value in clinical practice would require a more comprehensive study with further and more complex cases.

Declarations

Ethical vote: Ethik-Kommission der Universitätsmedizin Göttingen, Prof. Dr. Jürgen Brockmöller, vote-no:29/4/21, 21.04.2021.

Conflict of Interest: The authors declare that there is no conflict of interest.

Author contributions: TB, TS, PB, DK: conception of the work, data acquisition and interpretation; DK, US, TS: study design, ethics; TB, DK data analysis and interpretation; TB, DK: writing the manuscript. All authors approved the manuscript in the submitted version and take responsibility for the scientific integrity of the work.

Acknowledgement: The work has been supported by the German Federal Ministry of Education and Research (HiGHmed, grant no. 01ZZ1802B) and the Lower Saxony Ministry of Science and Culture (ZDIN/Zukunftslabor Gesundheit, grant no. ZN349).

References

[1] Levine HD. Clinical interpretation of electrocardiogram by means of electronic computers. American Heart Journal 1965; 69(2):147–9. doi: 10.1016/0002-8703(65)90030-X.

[2] Hong S, Zhou Y, Shang J, Xiao C, Sun J. Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. Comput Biol Med 2020; 122:103801. doi: 10.1016/j.compbiomed.2020.103801.

[3] Ribeiro AH, Ribeiro MH, Paixão GMM, Oliveira DM, Gomes PR, Canazart JA et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat Commun 2020; 11(1):1760. doi: 10.1038/s41467-020-15432-4.

[4] Jansen C, Krefting D. Reproduzierbarkeit eines Deep Learning Verfahrens zur Bestimmung von Schlafphasen, in (German Medical Science GMS Publishing House, 2019), p. DocAbstr. 300. doi: 10.3205/19GMDS068.

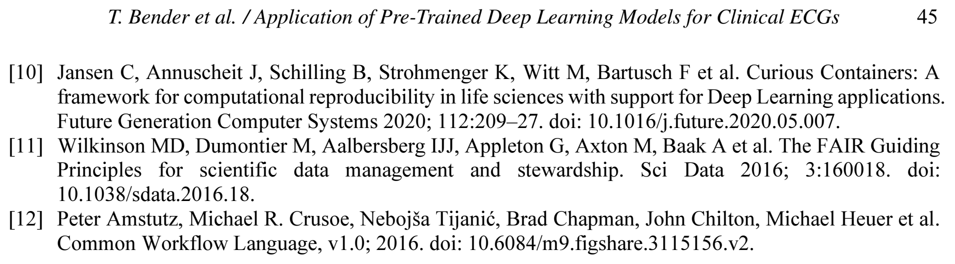
[5] Goodman SN, Fanelli D, Ioannidis JPA. What does research reproducibility mean? Sci Transl Med 2016; 8(341):341ps12. doi: 10.1126/scitranslmed.aaf5027.

[6] Ribeiro AH, Ribeiro MH, Paixão GM, Oliveira DM, Gomes PR, Canazart JA et al. Pre-trained deep neural network models for ECG automatic abnormality detection; 2020. doi: 10.5281/zenodo.3765717.

[7] Ribeiro AH, Ribeiro MH, Paixão GM, Oliveira DM, Gomes PR, Canazart JA et al. Annotated 12-lead ECG dataset; 2020. doi: 10.5281/zenodo.3765780.

[8] D. Krefting, M. Scheel, A. Freing, S. Specovius, F. Paul, and A. Brandt, Reliability of Quantitative Neuroimage Analysis Using FreeSurfer in Distributed Environments, in HP-MICCAI/MICCAI-DCI 2011 Workshop (Toronto, 2011), p. 10.

[9] LeVeque RJ, Mitchell IM, Stodden V. Reproducible research for scientific computing: Tools and strategies for changing the culture. Comput Sci Eng 2012; 14(4):13–7. doi: 10.1109/MCSE.2012.38.



[10] Jansen C, Annuscheit J, Schilling B, Strohmenger K, Witt M, Bartusch F et al. Curious Containers: A framework for computational reproducibility in life sciences with support for Deep Learning applications. Future Generation Computer Systems 2020; 112:209–27. doi: 10.1016/j.future.2020.05.007.

[11] Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016; 3:160018. doi: 10.1038/sdata.2016.18.

[12] Peter Amstutz, Michael R. Crusoe, Nebojša Tijanić, Brad Chapman, John Chilton, Michael Heuer et al. Common Workflow Language, v1.0; 2016. doi: 10.6084/m9.figshare.3115156.v2.

APPENDIX B

Article B

Author Contributions

I partly conceptualized the software. I implemented and published the software for XAI analysis, and jointly evaluated the analyses on all data sets, resulting in all figures and tables and all sections except II.B.1) and II.B.2). I prepared the original draft, communicated with the journal and worked in the reviewer feedback.

Analysis of a Deep Learning Model for 12-Lead ECG Classification Reveals Learned Features Similar to Diagnostic Criteria

Theresa Bender, *Student Member, IEEE*, Jacqueline M. Beinecke, Dagmar Krefting, Carolin Müller, Henning Dathe, Tim Seidler, Nicolai Spicher, *Member, IEEE*, and Anne-Christin Hauschild

***Abstract*— Despite their remarkable performance, deep neural networks remain unadopted in clinical practice, which is considered to be partially due to their lack of explainability. In this work, we apply explainable attribution methods to a pre-trained deep neural network for abnormality classification in 12-lead electrocardiography to open this ”black box” and understand the relationship between model prediction and learned features. We classify data from two public databases (CPSC 2018, PTB-XL) and the attribution methods assign a ”relevance score” to each sample of the classified signals. This allows analyzing what the network learned during training, for which we propose quantitative methods: average relevance scores over a) classes, b) leads, and c) average beats. The analyses of relevance scores for atrial fibrillation and left bundle branch block compared to healthy controls show that their mean values a) increase with higher classification probability and correspond to false classifications when around zero, and b) correspond to clinical recommendations regarding which lead to consider. Furthermore, c) visible P-waves and concordant T-waves result in clearly negative relevance scores in atrial fibrillation and left bundle branch block classification, respectively. Results are similar across both databases despite differences in study population and hardware. In summary, our analysis suggests that the DNN learned features similar to cardiology textbook knowledge.**

***Index Terms*—atrial fibrillation, electrocardiogram, explainable artificial intelligence, integrated gradients, layer-wise relevance propagation, left bundle branch block**

I. Introduction

The development and evaluation of algorithms for automatic interpretation of biosignals has attracted great interest in the last decade. Biosignals are time series, i.e. they are ordered sequences of measurements, which are usually acquired in successive and equally-spaced time intervals. Typical examples are the electrocardiogram (ECG) representing the electrical activity of the heart or the electroencephalogram (EEG) representing brain activity. The temporal ordering discriminates biosignals from many other types of biomedical data without any order, such as lab tests or sequencing, and introduces challenges in their interpretation by humans and algorithms alike. Next to measurement artefacts including loss of electrode contact, signals are influenced by other physiological processes, for example ECG by respiration, and (in)voluntary movement of the patient.

Traditionally, the field of ECG signal processing was dominated by methods based on mathematical or physical models recreating human physiology. Human experts defined semantic models or features which were used for different tasks, e.g. for generating synthetic waveforms [1], waveform delineation [2], or even human identification [3]. Evidently, this led to a plethora of proposed features and the question of which feature set is optimal for a specific task, e.g. for ECG classification [4]. Regarding this application, the aim is to either assign a label to individual heart beats or to a whole recording. As an example for the latter use case, the PhysioNet/CinC Challenge 2020 posed the task to automatically assign one or multiple of 27 classes to a large, multi-institutional database of 12-lead ECGs [5]. More than 200 teams took part with the most common algorithms being deep neural networks (DNNs).

In recent years, data-driven methods from the field of machine learning (ML) became popular, a significant percentage accounted for by DNNs [6]. At first many works used DNNs as classifiers and used traditional, semantic features as their input. However, recently there has been a trend towards ”end-to-end” pipelines where the raw signal is processed and DNNs extract relevant features themselves [7]–[11]. Although these methods are able to produce outstanding results and outperform conventional methods in many areas [12], [13], a pitfall lies in the fact that they are black box models and often based on agnostic features. While they bear the theoretical potential to aid in diagnostics or treatment decisions, clinicians need to be able to comprehend their reasoning as a ”Clever Hans” prediction [14], based on spurious or artifactual correlations, might lead to wrong decisions and adverse consequences for patients. Hence, next to issues such as inadequate performance metrics [15] and data leakage [16], one of the main reasons

Submitted on 14.11.2022. This research was funded by the German Federal Ministry of Education and Research (grant no. 16TTP073 11, and HiGHmed, grant no. 01ZZ1802B) and the Lower Saxony “Vorab” of the Volkswagen Foundation and the Ministry for Science and Culture of Lower Saxony (grant no. 76211-12-1/21).

T. Bender, J. M. Beinecke, D. Krefting, H. Dathe, N. Spicher and A.-C. Hauschild are with the Department of Medical Informatics, University Medical Center Göttingen, Göttingen, 37075 Germany. (e-mail: theresa.bender@med.uni-goettingen.de).

C. Müller and T. Seidler are with the Department for Cardiology & Pneumology/Heart Center, University Medical Center Göttingen, Göttingen, 37075 Germany.

(N. Spicher and A.-C. Hauschild are co-last authors.)

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3271858

for DNNs remaining unadopted in clinical practice is missing explainability [17], [18].

To address this need, frameworks and methods from the field of Explainable Artificial Intelligence (XAI) are developed and evaluated [19]. While XAI for text and tabular input data is advancing, XAI for time series data such as biosignals is still in the need for further research [20]. XAI methods for DNNs include layer-wise relevance propagation (LRP) [21], integrated gradients (IG) [22], and GRAD-Cam [23]. However, with regard to ECG classification, these methods are usually applied qualitatively [24]–[26] by showing individual recordings and corresponding XAI information, e.g. as pseudo-colored overlays. This qualitative evaluation of single recordings is rather anecdotal evidence and does not suffice the requirements for integrating DNNs in clinical practice, which needs a comprehensive characterization of models and their limitations.

Hence, in this work, we address the unmet clinical need of missing explainability by proposing a quantitative analysis pipeline (Fig. 1) enabling an objective justification of a DNN's decision. We use a state-of-the-art, pre-trained DNN proposed by Ribeiro et al. for abnormality classification in 12-lead ECGs [27] and apply attribution XAI methods to public ECG databases. In order to analyze the generalizability of this approach, we evaluate the explanatory power of different XAI methods and evaluate results on two different databases.

The XAI methods assign to each sample of the ECG time series a relevance score reflecting how much it influenced the DNN's decision. The main contribution of this work are novel analysis methods for processing these scores. These analyses allow to gain insight into the DNN's reasoning when classifying unseen ECG signals. By mapping the results to clinical knowledge, we investigate in how far the DNN's features align with clinical knowledge. By doing so, we also propose novel visualization methods of relevance scores, allowing an intuitive and quick assessment of DNN classifications.

II. Methods

A. Physiological Introduction

An ECG measures electrical activity on a patient's skin to monitor his/her cardiac cycle. It is a routine measurement in clinical settings, especially in emergency care as it allows a fast, accurate and comfortable assessment of key clinical parameters. Standard parameters derived from ECGs include heart rate, lengths between different peaks and waves, as well as the heart's electrical axis. Differences of these parameters to normal values can be interpreted as abnormalities, substantiating diagnoses. The acquisition of ECGs differs in length, e.g. 10 s in acute care or 24 h for Holter measurements, as well as circumstances, such as resting or exercise.

Raw ECG data is measured at equally-spaced points in time (samples) in units millivolt (mV) from multiple directions (leads) which are computed from differences in electrical potentials measured in two distinct electrodes. A standard resting ECG uses 10 electrodes, resulting in 12 leads, including six chest leads and six limb leads derived from electrodes on each arm and the left leg.

The stages of the cardiac cycle, a single heart beat, are represented by characteristic waves and peaks in a P-QRS-T sequence. The P-wave represents the depolarization before the contraction of the atria which is initiated by the sinus node. The QRS-complex consists of the Q-, R-, and S-waves and corresponds to the ventricular systole, and the T-wave represents the ventricular relaxation.

The morphology of the different waves, such as amplitude or width, as well as the intervals in between are clinically relevant. For example, atrial fibrillation (AF) is an arrhythmia based on uncoordinated electrical impulses in the atrium of the heart and a non-functioning sinus node [28] that can be diagnosed from ECGs. Criteria for diagnosis are absence of P-waves, as they are initiated by the sinus node, and irregular RR intervals [28]. However, repeating fibrillatory waves (f-waves) mimic P-waves and can usually be observed best in leads V1-6, especially V1 [29]. Another abnormality is left bundle branch block (LBBB), where the cardiac conduction through the left bundle branch is compromised downstream from lesions of the His bundle or its derivatives. LBBB criteria for ECGs include unusually wide QRS-complexes with the ST-segment and T-waves pointing in opposite direction [30]. I, aVL, V5 and V6 are left-sided leads, where broad notched or slurred R-waves can be observed, while Q waves are absent [30]. Both, AF and LBBB, can be diagnosed by ECG acquisition with a reduced number of leads, but the gold standard for diagnosis is 12-lead ECG [31].

B. Technical Background

Ribeiro et al. published a residual network (ResNet) trained on more than two million ECGs from a Brazilian telehealth network, showing F1-scores of more than 80 % for classification of six ECG abnormalities. The output from convolutional layers in each of four residual blocks are fed into a fully connected layer with sigmoid activation function, yielding independent probabilities for six classes of ECG abnormalities [32]. Thresholds calculated for the final classifications are available on GitHub[1]. In previous work, we demonstrated methods and results reproducibility with local data [33].

The model accepts a matrix with dimensions $N \times 4096 \times 12$ with 4096 and 12 defining the number of samples and leads, respectively. N denotes the number of recordings to be processed. The model outputs a matrix with dimensions $N \times 6$ assigning probabilities for six ECG abnormalities, namely first degree AV block, right bundle branch block, LBBB, sinus bradycardia, AF and sinus tachycardia.

In medical applications such as ECG diagnostics it is important for clinicians to understand the reasoning of a DNN. XAI methods build a wrapper around the black box model, giving insight into possible features that led to the DNN's output. In this paper, we focus on two state-of-the-art attribution methods, IG and LRP.

[1] `https://github.com/antonior92/automatic-ecg-diagnosis/blob/master/generate_figures_and_tables.py`, commit 89f929d, line 121

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3271858

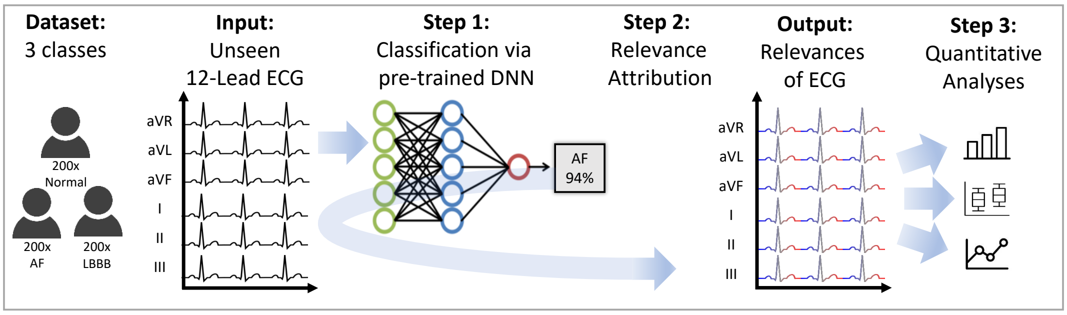


Fig. 1: Overview of the processing pipeline which is applied separately to data stemming from two different databases (CPSC/PTB-XL): For each database, the data set consists of 200 healthy controls (Normal) that are compared to patients showing AF and LBBB. Each (unseen) 12-lead ECG is fed into the pre-trained DNN and subsequently results are explored with the XAI methods, yielding a relevance score for each input sample, indicated here by blue (negative relevance score), grey (neutral), and red values (positive relevance score). We propose novel analysis methods for these scores, allowing to gain insight into the DNN's reasoning.

1) Integrated Gradients: IG attribute the prediction of a neural network on unseen data to its input features. However, IG use a baseline input for attribution calculation. The authors [22] motivate this by noting that if we assign blame to something, we implicitly consider the absence of it as a baseline for comparing outcomes.

IG are calculated as follows: Let f be a function that represents a neural network, x the input at hand, and $\tilde{x}$ the baseline input. The IG are defined as the path integral of the gradients along the straight-line path from the baseline $\tilde{x}$ and input x. The straight-line path can easily be written down as $\tilde{x} + \alpha(x - \tilde{x})$ for $\alpha \in [0, 1]$. The integrated gradient for the i-th input dimension is defined as

$$\mathrm{IG}_i(x) := (x_i - \tilde{x}_i) \cdot \int_0^1 \frac{\partial f(\tilde{x} + \alpha(x - \tilde{x}))}{\partial x_i} d\alpha, \qquad (1)$$

where $\frac{\partial f(x)}{\partial x_i}$ is the gradient of $f(x)$ along the i-th dimension.

The property of the LRP methods that the relevance scores of the input can be summed up and approximate the prediction score (see (4)) can also be proven for IG by using the fundamental theorem of calculus for path integrals. This states that if $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable almost everywhere[2] then

$$\sum_{i=1}^{n} \mathrm{IG}_i(x) = f(x) - f(\tilde{x}). \qquad (2)$$

For a baseline $\tilde{x}$ with prediction $f(\tilde{x})$ near zero, we can see that the sum over the IG in (2) also approximates the prediction score $f(x)$ similar to how the sum over the relevance scores calculated by LRP approximates the prediction score $f(x)$ in (4). This property is termed *completeness* in [22].

For computing IG the integration is replaced by a sum over sufficiently small intervals along the straight-line path

$$\mathrm{IG}_i^{approx}(x) := (x_i - \tilde{x}_i) \cdot \sum_{k=1}^{m} \frac{\partial f(\tilde{x} + \frac{k}{m}(x - \tilde{x}))}{\partial x_i} \frac{1}{m}. \qquad (3)$$

2) Layer-wise Relevance Propagation: LRP tries to explain the output $f(x)$ made by a classifier f with respect to an input x by decomposing the output $f(x)$ in such a way that

$$f(x) \approx \sum_{d=1}^{V} R_d, \qquad (4)$$

where V is the input dimension. $R_d > 0$ would then indicate the presence of the structure which is to be classified and $R_d < 0$ would indicate its absence.

Propagation of relevance scores works as follows: Let $R_j^{(\ell+1)}$ be a known relevance score of a certain neuron j in the $\ell+1$-th layer of a neural network, for a classification decision $f(x)$. The decomposition of the relevance score $R_j^{(\ell+1)}$ in terms of messages $R_{i \leftarrow j}$ sent to neurons of the previous layer ℓ must hold the conservation property

$$\sum_i R_{i \leftarrow j}^{(\ell,\ell+1)} = R_j^{(l+1)}, \qquad (5)$$

where $\sum_i$ describes the sum over all neurons in the ℓ-th layer of the neural network.

One possible relevance decomposition that satisfies (5) would be to use the ratio of local and global activations:

$$\begin{aligned} R_{i \leftarrow j}^{(\ell,\ell+1)} &= \frac{x_i^{(\ell)} \omega_{ij}^{(\ell,\ell+1)}}{\sum_k x_k^{(\ell)} \omega_{kj}^{(\ell,\ell+1)} + b_j^{(\ell)}} R_j^{(\ell+1)} \\ &= \frac{x_i^{(\ell)} \omega_{ij}^{(\ell,\ell+1)}}{z_j} R_j^{(\ell+1)}, \qquad (6) \end{aligned}$$

[2] This means f is continuous everywhere and the partial derivative of f along each input dimension is Lebesgue integrable. This holds for most neural networks using Sigmoid, ReLU, or Pooling functions.

where x_i is the activation (calculated by a non-linear activation function) of the i-th neuron in the ℓ-th layer, $w_{ij}^{(\ell,\ell+1)}$ is the weight connecting neuron i in the ℓ-th layer to neuron j in the $\ell+1$-th layer, $b_j^{(\ell)}$ is a bias term, and $\sum_k$ describes the sum over all neurons in the ℓ-th layer.

A problem with (6) is that if z_j gets very small, the relevance scores $R_{i\leftarrow j}$ can get infinitely large. To overcome this problem, the authors of [21] introduced a stabilizer $\epsilon \geq 0$:

$$R_{i\leftarrow j}^{(\ell,\ell+1)} = \begin{cases} \frac{x_i^{(\ell)}\omega_{ij}^{(\ell,\ell+1)}}{z_j+\epsilon} R_j^{(\ell+1)}, & \text{if } z_j \geq 0, \\ \frac{x_i^{(\ell)}\omega_{ij}^{(\ell,\ell+1)}}{z_j-\epsilon} R_j^{(\ell+1)}, & \text{if } z_j < 0. \end{cases} \tag{7}$$

As we can see in (7), if ϵ becomes very large, the relevance scores will tend to zero which poses another problem. To counteract this, a different treatment of positive and negative activations x_i is proposed in [21]. Let z_j^+ and z_j^- denote the positive and negative part of z_j such that $z_j^+ + z_j^- = z_j$. The same notation will be used for the positive and negative parts of $x_i^{(\ell)}\omega_{ij}^{(\ell,\ell+1)}$. Relevance decomposition can now be defined by

$$R_{i\leftarrow j}^{(\ell,\ell+1)} = R_j^{(\ell+1)} \cdot \left(\alpha \cdot \frac{\left(x_i^{(\ell)}\omega_{ij}^{(\ell,\ell+1)}\right)^+}{z_j^+} \right. \\ \left. + \beta \cdot \frac{\left(x_i^{(\ell)}\omega_{ij}^{(\ell,\ell+1)}\right)^-}{z_j^-} \right), \tag{8}$$

where $\alpha + \beta = 1$.

A different propagation rule has been proposed by [34] for real valued inputs that redistributes relevance scores according to the square magnitude of the weights:

$$R_{i\leftarrow j}^{(\ell,\ell+1)} = \frac{\left(\omega_{ij}^{(\ell,\ell+1)}\right)^2}{\sum_k \left(\omega_{kj}^{(\ell,\ell+1)}\right)^2} R_j^{(\ell+1)} \tag{9}$$

Other papers such as [35] and [36] propose a combination of different decomposition rules for different layer types, like (7) for fully connected layers to truthfully represent the decisions made via the layers' linear mapping and (8) for convolutional layers with ReLU activation functions to separately handle the positive and negative parts of $x_i^{(\ell)}\omega_{ij}^{(\ell,\ell+1)}$.

C. Experimental Design

Fig. 1 shows an overview of our DNN and XAI pipeline applied in this work. This pipeline is run separately on data stemming from two different databases.

1) Databases: The data set for our main analysis stems from the CPSC2018 database[3] acquired in eleven Chinese hospitals containing 12-lead ECGs with a ground truth provided by human experts [37]. Additionally, we validate the generalizability of our results using the PTB-XL database [38] as described in sec. II-C.7. An overview of the properties of both databases is shown in Tbl. I.

TABLE I: Properties of CPSC [37] and PTB-XL [38]

	# ECGs	Duration	Sampling	# Patients	Country
CPSC	$9,831$	$6-60$s	500 Hz	$9,458$	PRC
PTB-XL	$21,799$	10s	500 Hz	$18,869$	GER

For our main analysis on the CPSC database, we use a subset of 200 each for AF, LBBB and healthy subjects showing normal signals, resulting in $N = 600$ recordings. We investigate these two classes as AF is defined by an abnormal heart rhythm, i.e. irregular distances between heart beats, and therefore it can only be diagnosed by analyzing multiple heart beats. In contrast, LBBB can be diagnosed by a single heart beat as it is characterized by distinct morphological features, e.g. a notched QRS-complex.

2) Processing pipeline: All recordings were resampled to 400 Hz and trimmed or zero-padded to 4096 samples. In the remainder of this work, we denote a single ECG sample as $E_{n,j,k}$ with $n = \{0, 1, \ldots 599\}$ representing the recording index, $j = \{0, 1, \ldots 4095\}$ representing samples, and $k = \{0, 1, \ldots 11\}$ representing leads. Regarding data processing[4], each ECG signal is fed to the model by Ribeiro et al. [39] for classification, resulting in a matrix with dimensions $N \times 6$ assigning probabilities for six ECG abnormalities. In the following, we define $\{C_n \in \mathbb{R} \mid 0 \leq C_n \leq 1\}$ indicating the prediction score of the model with sigmoid activation, representing the classification probability. We utilize the package iNNvestigate [40], which implements multiple XAI methods, to compute relevance scores for each sample of the input ECGs. We use the XAI methods IG and LRP with the IG implementation being with baseline input zero and interval size $m = 64$, after changing the activation of the DNN's last layer to linear. Sigmoid activation does not change the ranking order of the predicted classes, but might obfuscate the true confidence of the model's individual class predictions[5].

The XAI methods assign a relevance score $R_{j,k} \in \mathbb{R}$ to each input sample of a classified ECG recording. By computing this for all N recordings we obtain $R_{n,j,k}$ with the same dimensions as our input ECG data $E_{n,j,k}$. Both are the basis for our analysis to compare features embedded in the DNN model to clinically-relevant criteria. We analyze the obtained relevance scores $R_{n,j,k}$ with three novel quantitative methods and one qualitative method as described in the following sections. With each new analysis, we take more details into account. While in the first analysis relevance scores are binned to each class, in the second analysis we split relevance scores w.r.t. their lead and in the third analysis w.r.t. lead and heart beats.

3) Binned and Average Relevance Scores Over Class: We first analyze relevance scores for all 200 normal, 200 LBBB, and 200 AF recordings separately and bin the values for their respective class, allowing us to compare the overall distribution

[3] `https://storage.cloud.google.com/physionet-challenge-2020-12-lead-ecg-public/PhysioNetChallenge2020_Training_CPSC.tar.gz`

[4] All computations are implemented using Python v3.6.8 and the libraries iNNvestigate v1.0.9, Tensorflow v1.12.0, neurokit2 V0.1.7, and h5py v2.10.0.

[5] `https://github.com/albermax/innvestigate/issues/84`, accessed: October 14, 2022

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3271858

of $R_{n,j,k}$ for the different classes.

We then aggregate all leads of each recording n into

$$M_n := \frac{1}{J\ K} \sum_{k=0}^{K-1} \sum_{j=0}^{J-1} R_{n,j,k}, \tag{10}$$

with $K = 12$ and $J = 4096$. $R_{n,j,k}$ takes positive or negative values, hence a higher M_n is associated with a higher prediction score, termed *completeness* in [22]. Here, the prediction score is the output of the model with linear activation.

4) Average Relevance Scores Over Class and Lead: We aggregate relevance scores for each lead k and recording n in

$$M_{n,k} := \frac{1}{J} \sum_{j=0}^{J-1} R_{n,j,k}, \tag{11}$$

with $J = 4096$. This allows for comparing the distribution of $R_{n,j,k}$ w.r.t. class and ECG leads and thus the importance of the individual ECG leads for the DNN. This is required as the different leads show different morphologies and signal shapes that might cancel out in the first analysis.

5) Average Relevance Scores Over Class, Lead, and Beats: In the first two analysis methods, time information is lost. However, for explaining the DNNs decision this is crucial as we need to compare whether the agnostic features trained by the DNN reflect the clinical features described in section II-A such as missing P-waves, unusually wide QRS-complexes etc.

Analyzing individual ECG records gives only anecdotal evidence. Therefore, we perform a two-step averaging procedure which averages the information over several recordings while preserving time information. First, for each ECG record and lead, we use the concept of ”average beats” [41] by splitting the whole signal into individual heart beats with the *ecg_segment()* function of neurokit2. We average them into a single, time-aligned representative beat for each lead. Then we use the exact same indices of the heart beats and perform the same steps on the relevance scores $R_{n,j,k}$, yielding an ”average relevance score”. All average beats and average relevance scores are then averaged for a given class. All segments are of equal size for one recording, hence we fill segments overlapping start or end of the recording with zeros. Finally, amplitudes are normalized to $[-1, 1]$. For scatter plot visualizations, relevance scores are upsampled by a factor of 5.

6) Qualitative Analysis of XAI Relevance Scores: The results of all processed ECG signals were visualized as heatmap-colored scatter plots for each lead, after a normalization of the output to $[-1, 1]$, keeping the center of the values at zero. Furthermore, these relevance score plots were evaluated by an experienced cardiologist.

7) Comparison Between Databases: To evaluate the the generalizability of our processing pipeline, we evaluate results on another publicly-available dataset. For this task we use PTB-XL [38] which is which is an older public database acquired between October 1989 and June 1996 in Germany. Therefore, the ECG measurement equipment and subject's origin are completely different to the CPSC database and additionally there is the chance of different clinical guidelines being in practice for the annotation by cardiologists.

8) Comparison Between XAI Methods: Since both methods, IG and LRP, differ substantially in their approach on how to calculate relevance scores for the input, we believe that using both methods will help uncover important information about why the DNN made certain decisions. Hence, we compare IG results to LRP using the following LRP decomposition rules implemented in the iNNvestigate [40] package:

a) The ϵ-LRP decomposition (see (7)) with $\epsilon = 1e - 07$.
b) The $\alpha\beta$-LRP decomposition (see (8)) with $\alpha = 1$ and $\beta = 0$.
c) The ω^2-LRP decomposition (see 9)).
d) The combination of $\alpha\beta$-LRP decomposition (see (8)) with $\alpha = 1$ and $\beta = 0$ for convolutional layers and ϵ-LRP decomposition (see (7)) with $\epsilon = 0.1$ for fully connected layers.

The sigmoid function (used in the output layer) maps from $\mathbb{R}$ to $\mathbb{R}^+$ and thus inverts the signs of all negative values, as well as scales all values into the interval of $[0, 1]$. This results in only small and positive values being backpropagated by the LRP method possibly resulting in small and only positive relevance scores. Thus we compared these relevance scores to those obtained by using a linear output in the last layer. Since both activations yield similar results when compared visually in heatmaps, we decided to continue with linear activation, to avoid the possible sign flip.

D. Ethics approval

Human subject research: This work only makes use of public data and does not contain any additional information involving human participants obtained by the authors.

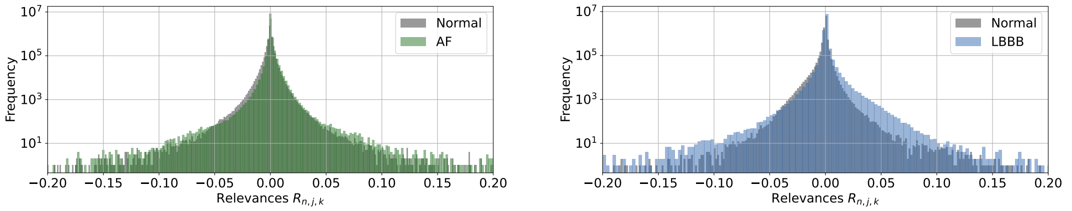
III. Results

After processing recordings with the DNN, C_n is the probability that a recording n shows the interrogated abnormality. The recording is classified as this abnormality if C_n is higher than a threshold defined by Ribeiro et al., which is 0.39 for AF and 0.05 for LBBB. Applying an XAI method results in a relevance score $R_{n,j,k} \in \mathbb{R}$ for each input sample of a classified ECG with $j = \{0, 1, \ldots 4095\}$ representing sample index, and $k = \{0, 1, \ldots 11\}$ representing the lead.

A. Average Relevance Scores Over Class

The mean of the distributions of IG relevance scores $R_{n,j,k}$ for each class (Fig. 2) is close to zero, representing that the majority of ECG samples is not relevant for the DNN's decision. Distributions for both abnormalities are almost similar to normal recordings, although they are slightly broader and shifted to positive values. For LBBB, in the range $[0.0, 0.10]$ there is a large number of more positive relevance scores compared to normal recordings (Fig. 2b).

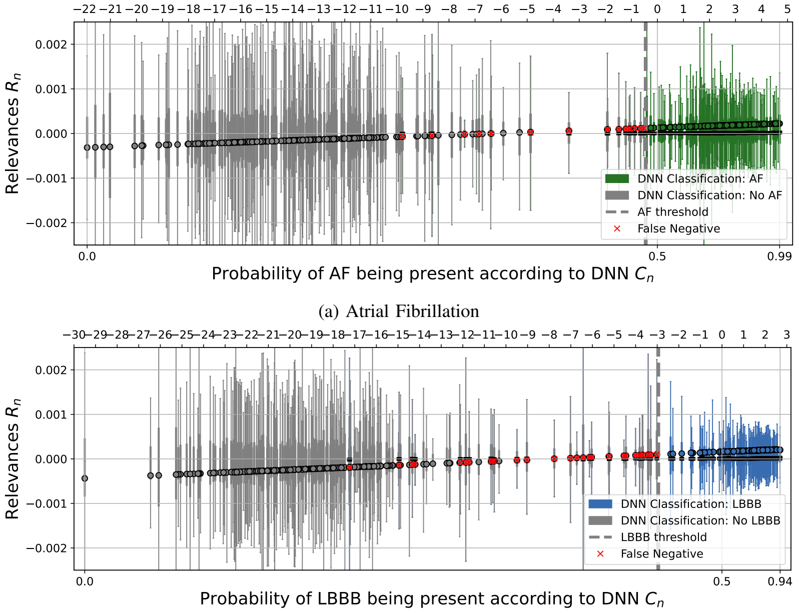
The relevance scores of individual recordings are again centered close to zero and rather equally-distributed (Fig. 3). In general, AF shows larger values in positive and negative direction compared to LBBB. While the median value is



(a) Normal and AF recordings. Colors denote ground truth label of data set. Values for AF range from $[-0.5, 0.5]$ and values for normal recordings from $[-0.3, 0.4]$.

(b) Normal and LBBB recordings. Colors denote ground truth label of data set. Values for LBBB range from $[-0.6, 0.9]$ and values for normal recordings from $[-0.4, 0.5]$.

Fig. 2: Distribution of IG relevance scores $R_{n,j,k}$. To increase visibility, x-axes are limited to $[-0.20, 0.20]$.



(a) Atrial Fibrillation

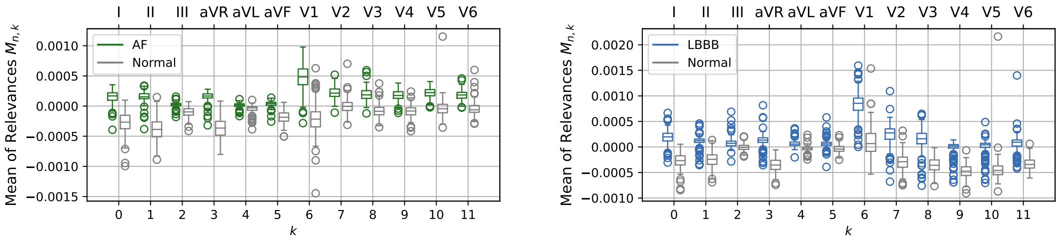
(b) Left Bundle Branch Block

Fig. 3: Distribution of R_n computed with IG for each recording as single boxplot. The bottom x-axis represents sigmoid activation output of the DNN, while the upper x-axis represents the output with linear activation. Boxplot colors denote DNN classification results and red crosses indicate false negatives.

always very close to zero, the mean value of relevance scores is increasing with increasing C_n. For AF classification (Fig. 3a) a large amount of normal recordings correctly classified as not showing AF have a C_n near 0 and correctly classified AF recordings are near 1. In between is a ”transition area” with nine false negative classifications in $[0.1, 0.39[$. The remaining seven false negatives show M_n values close to zero. LBBB has similar properties to AF, although there is no visible transition area and the values are not as close to 1 (Fig. 3b).

B. Average Relevance Scores Over Class and Lead

Analyzing model results of each lead k for AF classification (Fig. 4a), mean relevance scores showed medians of $0.0002, -0.0001$ and ranges of $[-0.0002, 0.0010]$ and $[-0.0014, 0.0012]$ for AF and normal recordings, respectively. For LBBB classification (Fig. 4b), medians were $0.0001, -0.0002$ and ranges were $[-0.0008,\ 0.0016]$ and $[-0.0009, 0.0022]$ for LBBB and normal recordings, respectively. For each lead, the mean relevance scores were significantly higher for both abnormalities compared to normal recordings, with a Wilcoxon-Rank-Sum-Test and p-value $<$ 0.01. Particularly, lead V1 shows the highest difference in



(a) AF classification

(b) LBBB classification

Fig. 4: Distribution of $M_{n,k}$ computed with IG w.r.t. ECG leads, colors denoting ground truth label. For AF classification (a) and LBBB classification (b) boxplots show that the abnormal mean is higher for each lead with the highest difference in V1.

median values.

C. Average Relevance Scores Over Class, Lead, and Beats

Average beats over 200 recordings show mostly positive relevance scores for both abnormalities, and mostly negative relevance scores for normal recordings for both classifications (Fig. 5).

When classifying AF, QRS-complexes are the most relevant areas, especially R-peaks. For normal recordings, we observed high negative values for the area of P-waves as well. Negative values of normal recordings are higher compared to positive values of AF recordings. For LBBB classification, QRS-complexes are most relevant as well (Fig. 6). Furthermore, the concentration of high absolute relevance scores on specific waves or peaks is clearer, such as the negative T-wave in LBBB, assigned with negative relevance scores when positive in normal recordings. In contrast, for AF many smaller relevance scores with higher variance are distributed on the whole beat.

D. Qualitative Analysis

We observed clusters of high absolute relevance scores in the area of QRS-complexes during visual inspection of single recordings visualized as heatmap (Fig. 7). For LBBB, IG seems to focus on negative S-waves and prolonged ST-segments in lead V1. Occasionally, broad and notched R-waves were also marked relevant. On the contrary, for AF recordings, the relevant parts were usually R-waves and in rare instances areas with missing P-waves.

When looking at individual recordings we also observed that in cases of artefacts, such as baseline drifts or noise, IG relevance scores are usually accumulated mainly in these areas. This can be seen on multiple false negative classifications, such as recordings A1017 (lead V1, Fig. 8), A0745 (V6), and A0205, A0502 (both multiple leads, mainly: V1-6). In some cases the classification was still correct despite the focus on artefacts, e.g. A0639 (V1) classified as AF with ≈ 0.904.

E. Comparison of Databases

We repeated all experiments conducted on the CPSC database using data from PTB-XL instead. All quantitative methods show similar results for PTB-XL data, exemplarily shown for average beats in AF classification in Fig. 9. Especially the distribution of relevance scores for LBBB recordings is narrower and shifted closer to positive values than for CPSC data (Fig. 10).

F. Comparison of XAI Methods

IG and all considered LRP methods yield diverging results for the given data set. As can be seen in Fig. 11 as an example, LRP methods ϵ and $\alpha\beta$ distribute high absolute relevance scores especially around R-peaks, while ω^2 shows higher absolute values on waves in between as well as artefacts. IG can also concentrate high absolute relevance scores around artefacts, but generally shows more high absolute values, especially on R peaks, when comparing leads of single patients to each other.

IV. Discussion

Results of the first analysis show that IG relevance scores follow a reasonable distribution (Fig. 2) with the majority of values being close to zero. This is expected as the majority of samples in an ECG is at baseline, e.g. the interval between two heart beats from the end of the T-wave to the beginning of the P-wave, and carry little clinically-relevant information. Comparing AF and LBBB classification shows that the AF relevance scores are more evenly spread around zero while the LBBB relevance scores tend to more positive relevance scores which can also be seen clearly in Fig. 2b with a distinct gap for positive relevance scores between LBBB and normal recordings. We conclude that the DNN trained a larger inter-class distance for LBBB classification.

Analyzing individual recordings (Fig. 3) shows similar distributions for both classifications. Additionally, a distinct relationship between the averaged relevance scores M_n and the probability of the DNN C_n can be observed. An optimal DNN classifier would show a cluster nearby $C_n = 0$ and $M_n \ll 0$ for normal recordings as well as a cluster nearby $C_n = 1$ and $M_n \gg 0$ for AF/LBBB. The analyzed DNN shows a sub-optimal relationship that can generally be expected with a transition area between both clusters in which the DNN does not have high certainty in its decisions (e.g. Fig. 3a:

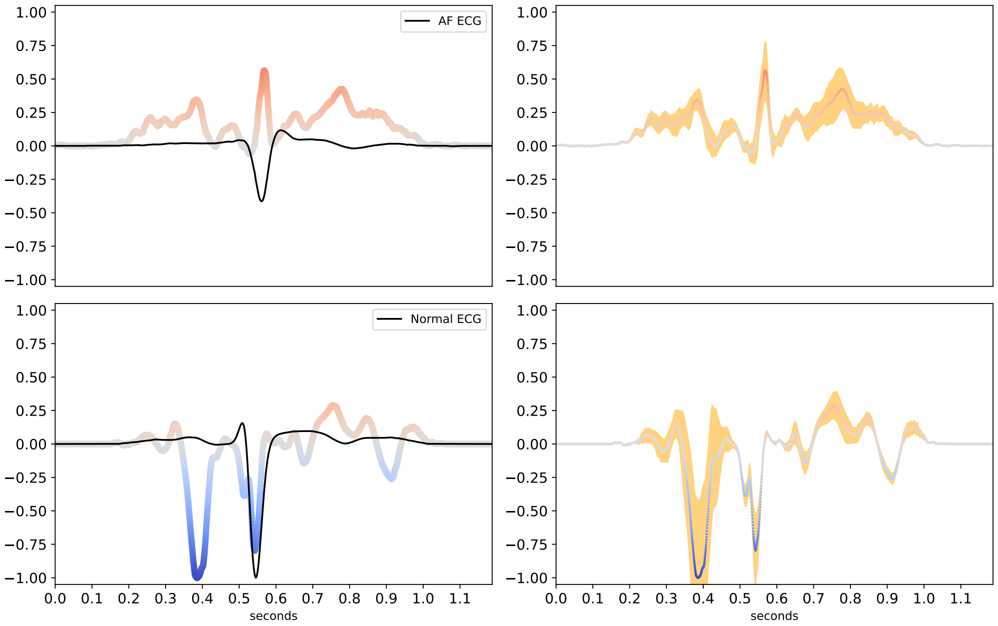


Fig. 5: Left column: Average beats (black curves) and IG relevance scores for lead V1 in AF classification. Abnormal ECGs show positive relevance scores (red) distributed over the whole P-QRS-T-cycle, negative relevance scores (blue) on normal recordings cover QRS-complexes and especially P-waves. Right column: Instead of average beats, the variance of relevance scores across recordings is shown (orange).

$C_n \in [0.1, 0.4]$). Furthermore, we observed many of the false negative classifications slightly below the threshold, indicating that the thresholds might not be optimal for the CPSC data set.

When analyzing individual leads, significant differences in relevance score distributions between abnormal and normal recordings were revealed (Fig. 4). This indicates which leads are most relevant for the DNNs decision. In general, for AF, the limb leads show lower relevance scores compared to the chest leads [29]. For AF as well as LBBB classifications, lead V1 shows clear positive relevance scores, indicating that the DNN trained clinically-relevant features: For AF, f-waves can often be observed in V1 [42] and for LBBB a negative terminal deflection in V1, e.g. a rS-complex with a tiny R-wave and a huge S-wave, is a clear diagnostic marker [43]. Interestingly, there is a large difference in the distributions of the precordial leads V4-V6. While in AF it shows a clear tendency towards positive relevance scores, for LBBB the median is close to zero. Another sign for LBBB are prolonged R-waves and absence of Q-waves in left-sided leads [44] which might not have been learned.

For these first analyses, we used averaged mean values of relevance scores, which have been used for explanations of models that take feature based input instead of raw data [45], [46]. However, this is a rather coarse measure. As the relevance scores are signed, values can be composed of rather low relevance scores or competing strong relevance scores for and against the respective class. Still, outliers in overall means or means of leads could be an indicator for false classification due to artefacts, for example if a lead not typically being relevant for this abnormality has the highest mean, such as in lead V6 in Fig. 4b.

As time information is lost in average means, we proposed the third analysis. As can be seen in Figs. 5 and 6, the ”average beat” and ”average relevance scores” of a single lead can give an even more detailed idea of the model’s features. Although it is still not possible to uniquely identify the actual features learned by the DNN, positively relevant areas in case of missing P-waves for AF classification indicate a good fit to clinical criteria [42]. Additionally, for the healthy controls, there are very pronounced negatively relevant areas nearby P-waves, demonstrating that the DNN learned that existence of P-waves is a counter-sign for AF. As IG does not allow to gain insight into the time scale, we cannot quantify to what extent RR-interval variations impact relevance scores. However, as the QRS-complex has similar shapes in AF and normal recordings, we assume that the DNN took the arrhythmic RR-intervals of AF recordings into account.

Moreover, when analyzing the shape of an average relevance signal, which is continuously averaged over more and more recordings in Fig. 5 (see Supplemental Material for a video), it

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3271858

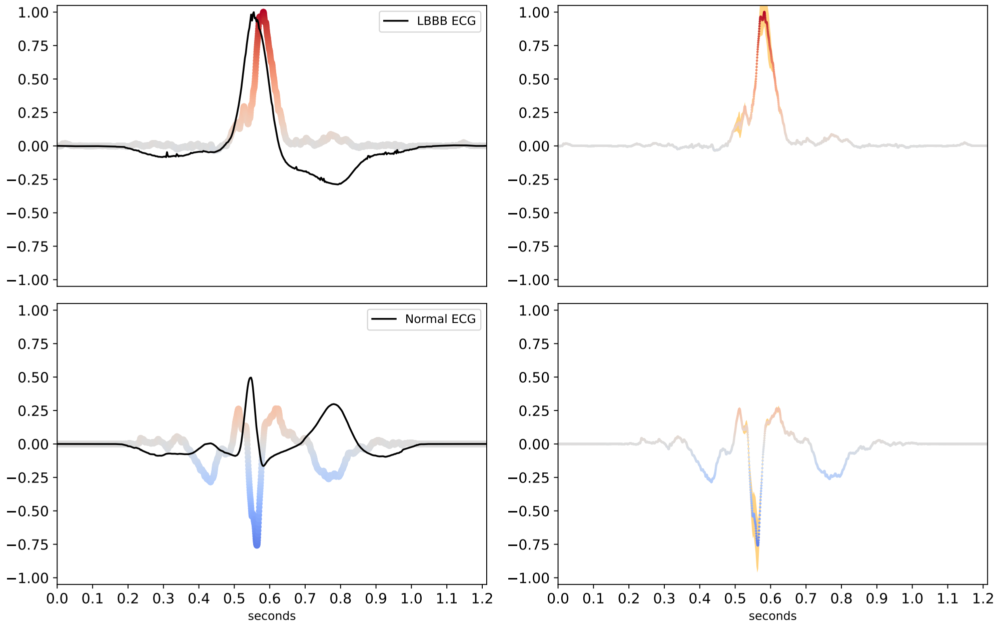


Fig. 6: Left column: Average beats and IG relevance scores for lead aVL in LBBB classification. Abnormal ECGs show positive relevance scores (red) on QRS-complexes; negative scores (blue) on normal recordings can be seen on P- and T-waves. Right column: Instead of average beats, the variance of relevance scores across recordings is shown (orange).

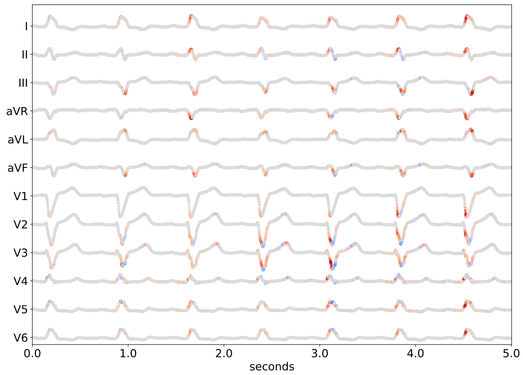


Fig. 7: Positive (red) and negative (blue) relevance scores calculated with IG on correctly classified electrocardiogram (LBBB: ~ 0.871) from CPSC data set (ID A0977). Relevance scores normed to $[-1, 1]$ per lead.

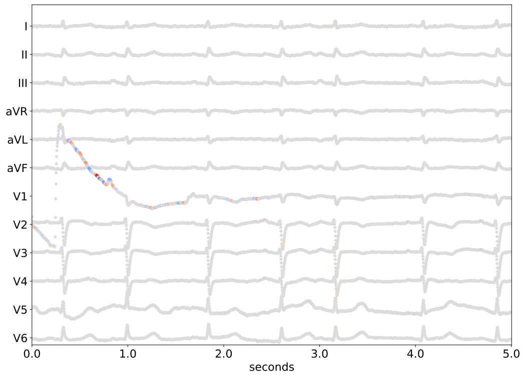


Fig. 8: Positive (red) and negative (blue) relevance scores calculated with IG on false negative classified ECG (AF: ~ 0.008) from CPSC data set (ID A1017). Relevance scores are clustered around the artefact in lead V1.

can be seen that, for AF as well as normal ECGs, the variance of relevance scores is quite low. This indicates a robustness of the DNN as it generates similar relevance scores despite the natural inter-patient variability in abnormal ECGs. Regarding LBBB classification (Fig. 6), high relevance scores around broadened QRS-complexes indicate a good fit to clinical criteria [30]. The criterion of a T-wave displacement opposite to the major deflection of the QRS-complex [30] can also be observed very well, although it results in small positive relevance scores only. In contrast, for healthy controls, T-

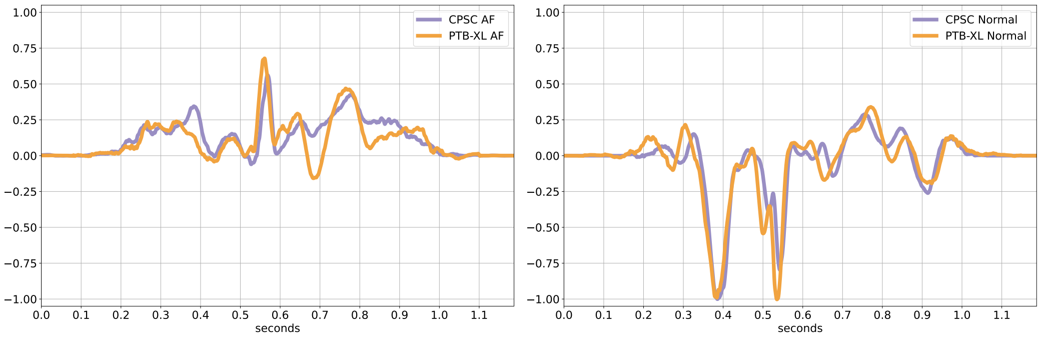


Fig. 9: IG relevance scores for lead V1 averaged over 200 ECGs extracted from CPSC (blue) and PTB-XL (orange). Figures depict AF recordings (left) and normal recordings (right), respectively.

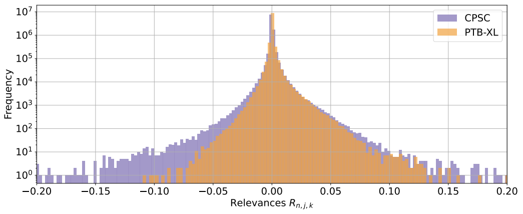


Fig. 10: Relevance scores of LBBB recordings from CPSC database (blue) compared to PTB-XL data (orange). To increase visibility, the x-axis is limited to $[-0.20, 0.20]$. Values for LBBB range from $[-0.11, 0.21]$ and values for normal recordings from $[-0.64, 0.56]$.

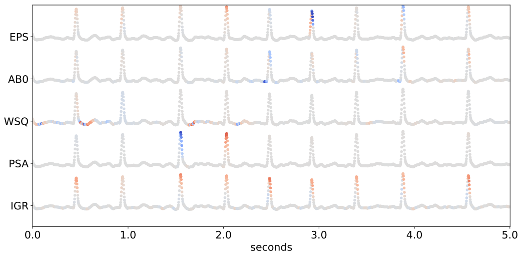


Fig. 11: Relevance scores calculated with five XAI methods normed to $[-1, 1]$ each on lead V6 of a correctly classified electrocardiogram (AF: $\sim$ 0.987) from CPSC data set (ID A0086). EPS: LRPEpsilon, AB0: LRPAlpha1Beta0, WSQ: LRPWSquare, PSA: LRPSequentialPresetA, IGR: IntegratedGradients.

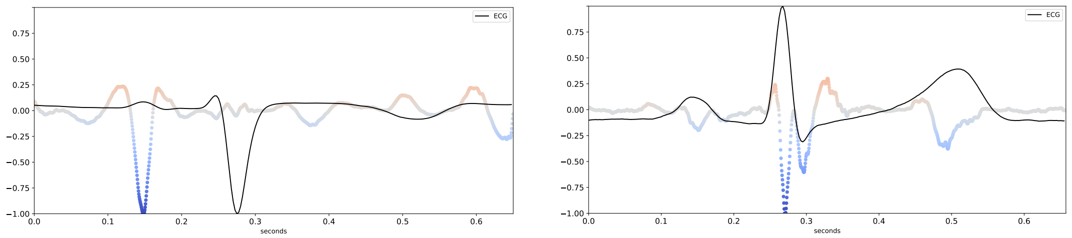
waves result in very pronounced negatively relevant areas (e.g. Fig. 12b). Similarly, for AF classifications, P-waves are learned as a feature that indicates the absence of AF (e.g. Fig. 12a). Furthermore, the robustness of the relevance scores in terms of variance is even higher than for AF.

1) Comparison of XAI methods: In this work we applied the XAI attribution methods IG and LRP. There are other approaches available for explaining models for biosignal data using ante-hoc methods as in [18], [47], [48], but these methods are not suitable for pre-trained DNNs where no adaption to the model itself is possible. Other methods, such as perturbation methods [49], [50], focus on occluding different parts of images and then analyzing the resulting changes in activations. These methods can also be used to calculate relevance scores for every input feature, but as shown by [51] they produce noisier heatmaps compared to LRP methods. Our results indicate that both methods, IG and LRP, are well suited for gaining insight into reasoning of DNNs applied to biosignals. Additionally, we conducted a comparison of IG and LRP methods (Fig. 11) and came to the conclusion that IG gives most distinct results.

2) Comparison of databases: To account for a change in the underlying data set, we validated our results on the CPSC database using PTB-XL instead and obtained similar results. One noticeable difference was observed in the relevance score distribution of LBBB recordings, where less negative values for PTB-XL could be explained by the more specific label ”Complete LBBB”, which might be easier to classify. These more differentiated labels bear the potential for comparison of model performance on complete and incomplete LBBBs.

3) Artifacts: We observed that the DNN tends to produce wrong classifications when artefacts are present as can be seen exemplarily in Fig. 8. This effect has been observed by others as well [24]. Although we have not attempted it in this work, artefact detection based on our approach could be a promising avenue for future work. Additionally, we observed that the relevance scores result in certain temporal patterns that might allow the application of analysis methods from nonlinear signal processing [52] which we will analyze in future work.

4) Key findings: In summary, our analysis suggests that the model by Ribeiro et al. learned features similar to cardiology textbook knowledge. IG relevance scores indicate that it learned features pointing towards a disease, such as the



(a) Lead V1 used for AF classification.

(b) Lead aVL used for LBBB classification.

Fig. 12: Average beats (black curve) and relevance scores for individual leads in a single normal recording correctly classified by the DNN: a) Highly negative relevance scores (blue) are found during the occurrence of the P-wave. b) Negative relevance scores (blue) are found during the P-/T-waves, and especially during occurrence of the P-wave of the QRS-complex.

abnormal QRS-complex in LBBB, while other features, such as the T-wave pointing in opposite direction, are not used for LBBB detection. Instead, the opposite of the feature, a T-wave pointing in expected direction, is used as a feature for detecting healthy ECGs. Our proposed analysis and visualization methods for relevance scores facilitate a rapid and effective assessment of the DNN's learned features and were confirmed by cardiologists.

5) Limitations: However, a limitation of our analysis based on IG is that we cannot infer any time-dependent information of the relevance scores. Especially for AF it is not clear whether e.g. the R-peaks are marked as relevant because of their morphology or their distance to one another. Therefore, we rate our results as more robust for LBBB as a morphological abnormality compared to AF as an arrhythmic and therefore time-dependent abnormality. Another limitation of our work is that we used public ECG databases which might introduce a certain bias. Therefore, using a data set from actual clinical practice on a cardiology ward or in emergency care might show different results. Thus, in future work, we will verify our results with more diverse data sources.

V. Conclusion

Missing explainability of ML methods for ECG analysis is a pressing issue preventing the dissemination of these methods in clinical practice. In this work we aimed enabling an objective justification of a DNN's decision by analyzing a state-of-the-art DNN for ECG classification with different XAI methods and data from different databases. Although this approach does not provide absolute certainty about the features learned by the DNN, it allows for inferring assumptions about its decision process. For example, our results reveal that the DNN learned that clearly-visible P-waves are a countersign for AF and T-waves pointing in same direction as the QRS-complex in particular leads are counter-signs for LBBB. Furthermore, decisions of the DNN for LBBB classification are based on unusual QRS-complexes. We conclude that the DNN learned cardiology textbook knowledge covering the whole cardiac cycle including P-wave, QRS-complex and T-wave. Moreover, we were able to explain false classifications due to transient noise which attracts the DNN's relevance scores, leading to relevant features being ignored.

In future work, we will use the methods proposed in this work for developing an interactive tool for clinical practice which offers cardiologists an intuitive overview of the DNN's reasoning, supporting them in their decision whether to trust the DNN's classification, or not.

Competing Interests

The authors declare no competing interests.

Code Availability

All source code developed in this work is publicly available on GitLab: `https://gitlab.gwdg.de/medinfpub/biosignal-processing-group/xai-ecg`, commit #aed722d8.

Supplementary Material

We provide plots of all quantitative analyses on PTB-XL in *PTB_analyses.pdf* and videos showing average beats and relevance scores for all CPSC data: *beats_V1_AF.mp4* (AF classification, lead V1), *beats_AVL_LBBB.mp4* (LBBB classification, lead aVL).

References

[1] P. E. McSharry, G. D. Clifford, L. Tarassenko, and L. A. Smith, "A dynamical model for generating synthetic electrocardiogram signals," *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 3, pp. 289–294, 2003.

[2] C. Böck, P. Kovács, P. Laguna, J. Meier, and M. Huemer, "Ecg beat representation and delineation by means of variable projection," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 10, pp. 2997–3008, 2021.

[3] S. A. Israel, J. M. Irvine, A. Cheng, M. D. Wiederhold, and B. K. Wiederhold, "Ecg to identify individuals," *Pattern Recognition*, vol. 38, no. 1, pp. 133–142, 2005.

[4] T. Mar, S. Zaunseder, J. P. Martínez, M. Llamedo, and R. Poll, "Optimization of ecg classification by means of feature selection," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 8, pp. 2168–2177, 2011.

[5] E. A. Perez Alday *et al.*, "Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020," *Physiological Measurement*, vol. 41, no. 12, p. 124003, Dec. 2020.

[6] F. Piccialli, V. Di Somma, F. Giampaolo, S. Cuomo, and G. Fortino, "A survey on deep learning in medicine: Why, how and when?" *Information Fusion*, vol. 66, no. 1, pp. 111–137, 2021.

[7] S. Yang *et al.*, "A multi-view multi-scale neural network for multi-label ecg classification," *IEEE Transactions on Emerging Topics in Computational Intelligence*, pp. 1–13, 2023.

[8] T. Pokaprakarn *et al.*, "Sequence to sequence ecg cardiac rhythm classification using convolutional recurrent neural networks," *IEEE journal of biomedical and health informatics*, vol. 26, no. 2, pp. 572–580, 2022.

[9] F. Liu *et al.*, "Automatic classification of arrhythmias using multi-branch convolutional neural networks based on channel-based attention and bidirectional lstm," *ISA Transactions*, 2023.

[10] D. Le, S. Truong, P. Brijesh, D. Adjeroh, and N. Le, "scl-st: Supervised contrastive learning with semantic transformations for multiple lead ecg arrhythmia classification," *IEEE journal of biomedical and health informatics*, pp. 1–10, 2023.

[11] Z. Yu *et al.*, "Ddcnn: A deep learning model for af detection from a single-lead short ecg signal," *IEEE journal of biomedical and health informatics*, vol. 26, no. 10, pp. 4987–4995, 2022.

[12] A. Y. Hannun *et al.*, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Medicine*, vol. 25, no. 1, pp. 65–69, Jan. 2019.

[13] S. W. Smith *et al.*, "A deep neural network learning algorithm outperforms a conventional algorithm for emergency department electrocardiogram interpretation," *Journal of Electrocardiology*, vol. 52, pp. 88–95, Jan. 2019.

[14] S. Lapuschkin *et al.*, "Unmasking Clever Hans predictors and assessing what machines really learn," *Nature Communications*, vol. 10, no. 1, p. 1096, Dec. 2019.

[15] M. A. Reyna, E. O. Nsoesie, and G. D. Clifford, "Rethinking Algorithm Performance Metrics for Artificial Intelligence in Diagnostic Medicine," *JAMA*, vol. 328, no. 4, pp. 329–330, 07 2022.

[16] S. Kapoor and A. Narayanan, "Leakage and the Reproducibility Crisis in ML-based Science," 2022, publisher: arXiv Version Number: 1.

[17] D. Yoon, J.-H. Jang, B. J. Choi, T. Y. Kim, and C. H. Han, "Discovering hidden information in biosignals from patients using artificial intelligence," *Korean journal of anesthesiology*, vol. 73, no. 4, pp. 275–284, 2020.

[18] Y. Elul, A. A. Rosenberg, A. Schuster, A. M. Bronstein, and Y. Yaniv, "Meeting the unmet needs of clinicians from AI systems showcased for cardiology with deep-learning–based ECG analysis," *Proceedings of the National Academy of Sciences*, vol. 118, no. 24, p. e2020620118, Jun. 2021.

[19] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, "Explaining deep neural networks and beyond: A review of methods and applications," *Proceedings of the IEEE*, vol. 109, no. 3, pp. 247–278, 2021.

[20] R. Guidotti *et al.*, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, 2019.

[21] S. Bach *et al.*, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[22] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17. JMLR.org, 2017, pp. 3319–3328.

[23] R. R. Selvaraju *et al.*, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020.

[24] H. Taniguchi *et al.*, "Explainable artificial intelligence model for diagnosis of atrial fibrillation using holter electrocardiogram waveforms," *International heart journal*, vol. 62, no. 3, pp. 534–539, 2021.

[25] M. Bodini, M. W. Rivolta, and R. Sassi, "Opening the black box: interpretability of machine learning algorithms in electrocardiography," *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, vol. 379, no. 2212, p. 20200253, 2021.

[26] I. Sturm, S. Lapuschkin, W. Samek, and K.-R. Müller, "Interpretable deep neural networks for single-trial eeg classification," *Journal of neuroscience methods*, vol. 274, pp. 141–145, 2016.

[27] A. H. Ribeiro *et al.*, "Automatic diagnosis of the 12-lead ecg using a deep neural network," *Nature communications*, vol. 11, no. 1, p. 1760, 2020.

[28] G. Hindricks *et al.*, "2020 esc guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the european association for cardio-thoracic surgery (eacts)the task force for the diagnosis and management of atrial fibrillation of the european society of cardiology (esc) developed with the special contribution of the european heart rhythm association (ehra) of the esc," *European Heart Journal*, vol. 42, no. 5, pp. 373–498, 2021.

[29] A. Bollmann *et al.*, "Analysis of surface electrocardiograms in atrial fibrillation: techniques, research, and clinical applications," *Europace : European pacing, arrhythmias, and cardiac electrophysiology : journal of the working groups on cardiac pacing, arrhythmias, and cardiac cellular electrophysiology of the European Society of Cardiology*, vol. 8, no. 11, pp. 911–926, 2006.

[30] N. Y. Tan, C. M. Witt, J. K. Oh, and Y.-M. Cha, "Left bundle branch block: Current and future perspectives," *Circulation. Arrhythmia and electrophysiology*, vol. 13, no. 4, p. e008239, 2020.

[31] K. Harris, D. Edwards, and J. Mant, "How can we best detect atrial fibrillation?" *The Journal of the Royal College of Physicians of Edinburgh*, vol. 42 Suppl 18, pp. 5–22, 2012.

[32] A. H. Ribeiro *et al.*, "Annotated 12-lead ecg dataset," 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3765780

[33] T. Bender, T. Seidler, P. Bengel, U. Sax, and D. Krefting, "Application of pre-trained deep learning models for clinical ecgs," *Studies in health technology and informatics*, vol. 283, pp. 39–45, 2021.

[34] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern Recognition*, vol. 65, p. 211–222, May 2017.

[35] W. Samek, A. Binder, S. Lapuschkin, and K.-R. Müller, "Understanding and comparing deep neural networks for age and gender classification," *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 1629–1638, 2017.

[36] M. Kohlbrenner *et al.*, "Towards best practice in explaining neural network decisions with lrp," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–7.

[37] F. Liu *et al.*, "An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection," *Journal of Medical Imaging and Health Informatics*, vol. 8, no. 7, pp. 1368–1373, 2018.

[38] P. Wagner *et al.*, "Ptb-xl, a large publicly available electrocardiography dataset," *Scientific data*, vol. 7, no. 1, p. 154, 2020.

[39] A. H. Ribeiro *et al.*, "Pre-trained deep neural network models for ecg automatic abnormality detection," 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3765717

[40] M. Alber *et al.*, "innvestigate neural networks!" *Journal of Machine Learning Research*, vol. 20, no. 93, pp. 1–8, 2019.

[41] P. Hamilton and W. Tompkins, "Compression of the ambulatory ecg by average beat subtraction and residual differencing," *IEEE Transactions on Biomedical Engineering*, vol. 38, no. 3, pp. 253–259, 1991.

[42] P. Langley, J. Bourke, and A. Murray, "Frequency analysis of atrial fibrillation," in *Computers in Cardiology 2000. Vol.27 (Cat. 00CH37163)*, 2000, pp. 65–68.

[43] D. G. Strauss, R. H. Selvester, and G. S. Wagner, "Defining Left Bundle Branch Block in the Era of Cardiac Resynchronization Therapy," *The American Journal of Cardiology*, vol. 107, no. 6, pp. 927–934, Mar. 2011.

[44] P. W. Macfarlane, "New ECG Criteria for Acute Myocardial Infarction in Patients With Left Bundle Branch Block," *Journal of the American Heart Association*, vol. 9, no. 14, p. e017119, Jul. 2020.

[45] S. M. Lauritsen *et al.*, "Explainable artificial intelligence model to predict acute critical illness from electronic health records," *Nature communications*, vol. 11, no. 1, p. 3852, 2020.

[46] C. Jansen *et al.*, "Network physiology in insomnia patients: Assessment of relevant changes in network topology with interpretable machine learning models," *Chaos (Woodbury, N.Y.)*, vol. 29, no. 12, p. 123129, 2019.

[47] Q. Hu *et al.*, *X-MyoNET: Biometric Identification using Deep Processing of Transient Surface Electromyography*, 2021.

[48] M. Doborjeh, Z. Doborjeh, N. Kasabov, M. Barati, and G. Y. Wang, "Deep learning of explainable eeg patterns as dynamic spatiotemporal clusters and rules in a brain-inspired spiking neural network," *Sensors (Basel, Switzerland)*, vol. 21, no. 14, 2021.

[49] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer vision - ECCV 2014*, ser. Lecture Notes in Computer Science, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer, 2014, vol. 8689, pp. 818–833.

[50] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," in *International Conference on Learning Representations*, 2017.

[51] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Muller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 11, pp. 2660–2673, 2017.

[52] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 278, no. 6, pp. H2039–H2049, 2000, pMID: 10843903.

APPENDIX C

Article C

Bender T, Gemke P, Idrobo-Avila E, Dathe H, Krefting D, Spicher N. Benchmarking the Impact of Noise on Deep Learning-Based Classification of Atrial Fibrillation in 12-Lead ECG. Stud Health Technol Inform 2023; 302:977–81, doi: 10.3233/SHTI230321. CC BY-NC[1]

Author Contributions

I partly processed the noise metadata and conceptualized the software. I implemented and published the software for noise analysis, and evaluated the results, resulting in Figure 1, both tables and all sections. I prepared the original draft and worked in the reviewer feedback. I prepared and held the conference talk at Medical Informatics Europe 2023 in Gothenburg.

[1] `https://creativecommons.org/licenses/by-nc/4.0/deed.en_US`

Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230321

Benchmarking the Impact of Noise on Deep Learning-Based Classification of Atrial Fibrillation in 12-Lead ECG

Theresa BENDER[a,b,1], Philip GEMKE[a], Ennio IDROBO-AVILA[a], Henning DATHE[a], Dagmar KREFTING[a,b] and Nicolai SPICHER[a,b]
[a] *Department of Medical Informatics, University Medical Center Göttingen, Göttingen, Germany*
[b] *DZHK (German Centre for Cardiovascular Research), partner site Göttingen, Göttingen, Germany*

Abstract. Electrocardiography analysis is widely used in various clinical applications and Deep Learning models for classification tasks are currently in the focus of research. Due to their data-driven character, they bear the potential to handle signal noise efficiently, but its influence on the accuracy of these methods is still unclear. Therefore, we benchmark the influence of four types of noise on the accuracy of a Deep Learning-based method for atrial fibrillation detection in 12-lead electrocardiograms. We use a subset of a publicly available dataset (PTB-XL) and use the metadata provided by human experts regarding noise for assigning a signal quality to each electrocardiogram. Furthermore, we compute a quantitative signal-to-noise ratio for each electrocardiogram. We analyze the accuracy of the Deep Learning model with respect to both metrics and observe that the method can robustly identify atrial fibrillation, even in cases signals are labelled by human experts as being noisy on multiple leads. False positive and false negative rates are slightly worse for data being labelled as noisy. Interestingly, data annotated as showing baseline drift noise results in an accuracy very similar to data without. We conclude that the issue of processing noisy electrocardiography data can be addressed successfully by Deep Learning methods that might not need preprocessing as many conventional methods do.

Keywords. Deep Learning, Electrocardiogram, Atrial Fibrillation, Noise

1. Introduction

Electrocardiograms (ECGs) are recordings of the electrical activity of the heart and are frequently used in emergency and in-patient care. However, different types of noise, either stemming from the patient's behaviour (e.g. motion) or the devices (e.g. power line interference), can be introduced during measurement. The presence of noise leads to a twofold problem: It impedes detection of anomalies leading to false findings and alarms [1] and, if the signal-to-noise ratio (SNR) reaches a certain level, detecting diagnostically relevant features becomes impossible [2].

[1] Corresponding Author: Theresa Bender, University Medical Center Göttingen, Robert-Koch-Str. 40, 37075 Göttingen, Germany, E-mail: theresa.bender@med.uni-goettingen.de.

One class of features with high clinical importance are the so-called “fiducial points”, i.e. the center, on- and offsets of ECG waves such as the QRS complex and the P-/T-wave. They are used for segmenting heartbeats into meaningful intervals [3] and by doing so allow for arrhythmia detection. Atrial fibrillation (AF) is the most prevalent arrhythmia which is characterized by uncoordinated electrical impulses in the atrium and might lead to severe cardiovascular issues, such as stroke or heart failure. Analyzing the interval in a heartbeat where a P-wave is expected is crucial for AF classification as its absence indicates a lack of sinoatrial node activity and is thereby a sign for AF [4]. However, so-called fibrillatory waves might occur, mimicking P-waves, impeding the assessment of sinoatrial node activity.

Many state-of-the-art algorithms for ECG classification are based on extracting semantic features derived from human expert knowledge, such as fiducial points. However, as these algorithms tend to wrong results in case of noise [5], various denoising strategies [6] have been proposed. In contrast, algorithms from the field of deep learning (DL) were explored for ECG classification tasks recently [7,8]. Instead of semantic features, they are based on agnostic features derived from fully-automatic correlation analysis between input ECGs and output classes in an end-to-end fashion. These models are based on the underlying premise that training and test datasets are stemming from the same distribution, which is often their pitfall in case of dataset shifts (variant devices, users, noise). Although initial studies indicate a better robustness to high SNRs [9,10], it remains unclear to which extend it affects these models.

Thereby, in this work we benchmark the accuracy of a state-of-the-art pre-trained DL model for 12-lead ECG classification regarding its susceptibility to different types of noise. We use the publicly available PTB-XL dataset which contains annotations for several categories of noise made by human technical experts and compare the model’s accuracy w.r.t. type of noise.

2. Methods

We analyze a subset of the PTB-XL dataset containing 12-lead ECGs of 10 second length with a sampling rate of 500 Hz that were acquired between 1989 and 1996 [11]. The subset contains all 1,514 ECGs annotated as showing AF (label in PTB-XL: *AFIB*) and we add the first 2,000 normal ECGs (*NORM*) as healthy controls. For each signal, we use a qualitative and a quantitative method to estimate SNR.

2.1. SNR Based on Annotations (SNR$_a$)

For each ECG we determine the number of noisy leads using the columns *baseline_drift*, *static_noise*, *burst_noise* and *electrodes_problems* provided in the PTB-XL metadata. In the majority of cases, they contain the name of a single lead (e.g. “aVL”), multiple leads (“I,aVR”) or ranges (e.g. “I-III”). Using a custom script, we convert this information to numeric values ranging from 0 to 12 for each type of noise. The labels “alles” (all) and “noisy recording” are converted to 12. We remove ECGs associated with other labels as they are of a more qualitative nature (e.g. “leicht” (light)). In this way, for each signal a qualitative, unit-less, linear SNR measure is computed, ranging from 0 (no noise reported) to 12*4=48 (all leads are affected by all types of noise). As shown in Tbl. 1, we use this information to split the dataset in ECGs without (“w/o”) a noise label and ECGs with (“w/”) a noise label.

Table 1. Properties of subset extracted from PTB-XL (left) and results of DL-based AF classification (right). ECGs are grouped according to annotations: In case there is one or more noise label in the metadata, an ECG is assigned to "w/", else to "w/o". FP and FN denote False Positive and False Negative, respectively.

Noise Label	AF	Healthy controls
w/o	1,097	1,581
w/	417	419

Noise Label	DL: FP	DL: FN
w/o	0.04 %	3.96 %
w/	0.24 %	7.06 %

It has to be underlined that a value of zero does not have to mean that there is no noise, it just reflects that there is a potential for a noise-free ECG. The authors of PTB-XL also indicated that missing annotations in case of artifacts or false annotations in case of noise-free signals might occur. However, they concluded that the metadata bears the potential for ECG quality assessment [12].

2.2. Measured SNR (SNR_m)

Due to the limitations of the manual annotations and as they are only available for 22 % of the PTB-XL database [12], we additionally use a quantitative SNR measure for each signal. We compute the Fourier Transform of the signals as well as the ratio of energies in two frequency bands as proposed in [13]. Based on the expected heart rates during AF, we define the "signal" frequency band ranging from 40 to 150 beats-per-minute (0.66 to 2.5 Hz) and define the "noise" frequency band as < 40 and > 150 beats-per-minute. By scaling with 10 log 10, we arrive at an SNR expressed in logarithmic decibel scale (dB).

2.3. DL Classification

ECG data is classified with a pre-trained model by Ribeiro et al. [7]. The model is a residual network and was trained on more than two million ECGs that were acquired within a Brazilian telehealth network. It outputs independent probabilities for six abnormalities, but we limit our analysis to AF. We use a threshold defined by the authors[2].

2.4. Data Analysis

We analyze the subset regarding differences between ECGs with and without noise labels for i) their distribution of SNR_m and SNR_a as well as ii) the accuracy of DL classification of each noise category. For ii) we compared the noisy recordings ($SNR_a > 0$) with randomly drawn signals from equally sized control groups ($SNR_a = 0$).

3. Results

Fig. 1 shows the distribution of SNR_a and SNR_m values on the left and right side. The majority of ECGs with noise labels has less than 15 with the maximum being 29. This shows that even in the duration of 10 seconds, different data quality issues per lead may occur. SNR_m values are occurring in the range of [-33.03,-7.78] dB with no clear difference between ECGs with and without noise labels.

[2] https://github.com/antonior92/automatic-ecg-diagnosis/blob/master/generate_figures_and_tables.py}, commit 89f929d, line 121

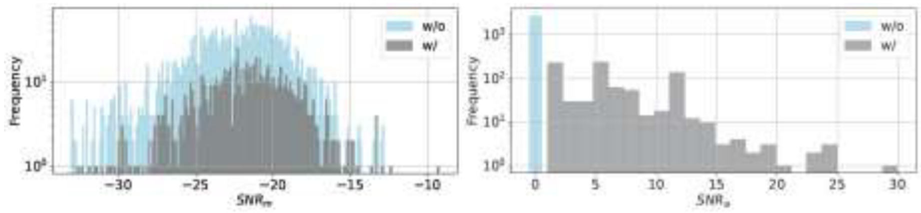


Figure 1. Distribution of values of both SNR metrics with (grey) and without (blue) noise labels.

Table 2. DL accuracy w.r.t. the four types of noise. The variable *n* represents the number of signals with the given label (w/). For comparison to signals without a label (w/o), *n* ECGs are randomly drawn 100 times and accuracy is given as mean ± standard deviation.

Type **Label**	**Baseline Drift** (*n* = 305)	**Static Noise** (*n* = 478)	**Burst Noise** (*n* = 156)	**Electrode** **Problems** (*n* = 6)
w/o	96.8 % ± 0.9 %	96.8 % ± 0.7 %	96.9 % ± 1.3 %	96.3% ± 8.0 %
w/	97.7 %	94.6 %	94.9 %	100.0 %

Tbl. 1 (right) shows FP and FN rates of AF classification w.r.t. the existence of noise labels. FP is worsened by 0.2 % and FN by 3.1 % in case ECGs are annotated with noise labels. Tbl. 2 shows the DL accuracy for each type of noise compared to the same number of ECGs but randomly drawn 100 times from data without noise labels. ECGs with baseline drift or electrode problems are classified more accurately in comparison to random ECG signals without noise annotations, whereas ECGs with annotated burst and static noise reveal worse performance.

4. Discussion

In general, the DL model robustly classifies AF, even in case ECGs are labelled by human experts as having multiple leads influenced by noise. Interestingly, in presence of baseline drift or electrode problems, accuracy is not deteriorated, but within one standard deviation compared to signals without noise labels. As a limitation, it has to be underlined that annotations are non-complete [12] and the subset contains only six signals annotated with electrode problems.

As the DL model can be assumed as a “black box”, we can only speculate about the reasons for this behaviour. It could be explained by partial misinterpretation of baseline drift or static noise as P-waves. As we could show in previous work [14], the DL model was trained such that P-waves and R-peaks have a high relevance, similar to human perception, while numerous other features influence its decision. This multi-factor decision process could be robust to different kinds of noise, but this requires its presence during training. A shift between training and test datasets is always an issue for DL models. To mitigate this effect is has been suggested to intentionally include noise during training [9]. The model used in this work was trained on two million non-public ECGs.

However, since the distribution of SNR_m looks visually similar with or without noise labels, SNR_a might not be optimal for quality assessment on its own. A “no noise” label, explicitly identifying ECGs without data quality issues, and more labels in general would be a valuable addition for future experiments.

5. Conclusion

Results show that the DL model is able to detect AF in 12-lead ECGs with high accuracy, even in the presence of data quality issues according to human experts. We conclude that end-to-end DL models based on agnostic features can address the difficulty of processing noisy ECGs. In contrast to conventional methods based on semantic features, they might not require preprocessing methods for achieving high accuracy. However, more experiments with larger and more diverse datasets should be the subject of future work.

References

[1] Festag S, Spreckelsen C. Semantic Anomaly Detection in Medical Time Series. Stud Health Technol Inform 2021; 278:118–25, doi: 10.3233/SHTI210059.

[2] Apandi ZFM, Ikeura R, Hayakawa S, Tsutsumi S. An Analysis of the Effects of Noisy Electrocardiogram Signal on Heartbeat Detection Performance. Bioengineering 2020; 7(2), doi: 10.3390/bioengineering7020053.

[3] Spicher N, Kukuk M. Delineation of Electrocardiograms Using Multiscale Parameter Estimation. IEEE J Biomed Health Inform 2020; 24(8):2216–29, doi: 10.1109/JBHI.2019.2963786.

[4] Kreimer F, Aweimer A, Pflaumbaum A, Mügge A, Gotzmann M. Impact of P-wave indices in prediction of atrial fibrillation-Insight from loop recorder analysis. Ann Noninv Electrocard 2021; 26(5):e12854, doi: 10.1111/anec.12854.

[5] Kumar P, Sharma VK. Detection and classification of ECG noises using decomposition on mixed codebook for quality analysis. Healthc Technol Lett 2020; 7(1):18–24, 10.1049/htl.2019.0096.

[6] Mir HY, Singh O. ECG denoising and feature extraction techniques – a review. Journal of Medical Engineering \& Technology 2021; 45(8):672–84, doi: 10.1080/03091902.2021.1955032.

[7] Ribeiro AH, Ribeiro MH, Paixão GMM, Oliveira DM, Gomes PR, Canazart JA et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat Commun 2020; 11(1):1760, doi: 10.1038/s41467-020-15432-4.

[8] Attia ZI, Harmon DM, Behr ER, Friedman PA. Application of artificial intelligence to the electrocardiogram. Eur Heart J 2021; 42(46):4717–30, doi: 10.1093/eurheartj/ehab649.

[9] Venton J, Harris PM, Sundar A, Smith NAS, Aston PJ. Robustness of convolutional neural networks to physiological electrocardiogram noise. Philos Trans A Math Phys Eng Sci 2021; 379(2212):20200262, doi: 10.1098/rsta.2020.0262.

[10] Sraitih M, Jabrane Y, Hajjam El Hassani A. A Robustness Evaluation of Machine Learning Algorithms for ECG Myocardial Infarction Detection. JCM 2022; 11(17):4935, doi: 10.3390/jcm11174935.

[11] Wagner P, Strodthoff N, Bousseljot R-D, Kreiseler D, Lunze FI, Samek W et al. PTB-XL, a large publicly available electrocardiography dataset. Sci Data 2020; 7(1):154, doi: 10.1038/s41597-020-0495-6.

[12] Strodthoff N, Wagner P, Schaeffter T, Samek W. Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL; 2020 Apr 28, doi: 10.48550/arXiv.2004.13701.

[13] Haan G de, Jeanne V. Robust pulse rate from chrominance-based rPPG. IEEE Trans Biomed Eng 2013; 60(10):2878–86, doi: 10.1109/TBME.2013.2266196.

[14] Bender T, Beinecke JM, Krefting D, Müller C, Dathe H, Seidler T et al. Analysis of a Deep Learning Model for 12-Lead ECG Classification Reveals Learned Features Similar to Diagnostic Criteria; 2022, doi: 10.48550/arXiv.2211.01738.

APPENDIX D

Abstract D

Bender T, Beinecke J, Hauschild AC, Krefting D, Spicher N. Towards Explaining Decisions of a Deep Learning Model for AF Detection in 12-lead ECGs. Joint Annual Conference of the Austrian, German and Swiss Societies for Biomedical Engineering (BMT) 2022, Innsbruck.

Author Contributions

I partly conceptualized the software. I implemented and published the software resulting in this abstract, prepared the original draft and worked in the reviewer feedback. I prepared and held the conference talk.

Towards Explaining Decisions of a Deep Learning Model for AF Detection in 12-lead ECGs

T Bender[1], J Beinecke[1], AC Hauschild[1], D Krefting[1], N Spicher[1]

[1] Department of Medical Informatics, University Medical Center Göttingen, Göttingen

Introduction

Currently, an increasing number of deep neural networks (DNNs) for biosignal classification are developed, often outperforming conventional methods based on handcrafted features. Understanding the reasoning of DNNs is a challenge, making their application difficult in clinical settings. In previous work, we applied a pre-trained DNN by Ribeiro et al. for 12-lead ECG classification to our local clinical data and reproduced the reported performance. In this work, we evaluate the feasibility of the attribution method “Integrated Gradients” (IG) for explaining the DNN’s classifications by means of a qualitative visual inspection and a quantitative analysis based on complexity.

Methods

We apply the Ribeiro model to a subset of the China Physiological Signal Challenge 2018 dataset. The model assigns probabilities for six ECG abnormalities, but we limit our analysis to atrial fibrillation (AF). We change the activation of the last layer to linear and apply IG to 200 AF and 200 sinus rhythm (SR) signals (10s duration) by using iNNvestigate. It assigns a positive or negative relevance value to each ECG sample. Subsequently, we calculate the sample entropy (SampEn, m=2, r=0.2std, N=4096) of these relevances w.r.t. lead and label and aggregate results as boxplots.

Results

Analyzing the relevances of model probabilities for AF classification with SampEn showed similar ranges of [0.03,0.80] and [0.06,0.82] for AF and SR patients, respectively. 11 out of 12 leads showed lower median values for AF patients, with the highest difference being 0.15 (lead V5). Regarding visual inspection, we observed a clustering of relevances in the area of QRS complexes. During measurement noise (e.g. inadequate skin-electrode contact) clusters with high absolute values and interchanging signs agglomerate.

Conclusion

We observed a tendency of IG relevances showing a higher complexity in SR than AF patients during AF detection, suggesting a higher uncertainty of the DNN when tending towards low probabilities.

Keywords: Deep Learning, Electrocardiogram, Atrial Fibrillation, Explainable Artificial Intelligence

Information about submission 252:

- Last change: 12 Apr 2022 18:07
- Theme: (1) BMT 2022 - Joint Annual Conference of the Austrian, German and Swiss Societies for Biomedical Engineering
- Topic: Focus Session: Signal-Based Risk Prediction in Cardiovascular Diseases (Joint Session GMDS-DGBMT)
- Submission type: Abstract Only

Submission decision

- Generic data:

  Please indicate in which form you would like to present your contribution. The decision on this lies with the scientific program committee: **Oral presentation**

  Click here to upload the Student Competition Application Form. This is mandatory only for participants of the student competition.: **No**

  To participate in the student competition, please make sure to also select Student Competition (Abstract and Conference Paper) as Submission type.: **-**
- Objective submission decision: Accepted (Accept, Only Abstract Submitted) as Oral Presentation
- Formal submission decision: Accepted (Accept, Only Abstract Submitted) as Oral Presentation
- Submission status: Formal decision

APPENDIX E

Source Code

All source code developed in this work is publicly available on GitLab.

- Reproducibility/Explainability:
  `https://gitlab.gwdg.de/medinfpub/biosignal-processing-group/xai-ecg`, commit #aed722d8.
- Robustness:
  `https://gitlab.gwdg.de/medinfpub/biosignal-processing-group/xai-ecg/-/tree/noise`, commit #0b456adf