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

doi:10.3233/SHTI230059

The Necessity of Multiple Data Sources for ECG-Based Machine Learning Models

Lucas PLAGWITZ[a,1], Tobias VOGELSANG[a], Florian DOLDI[b], Lucas BICKMANN[a], Michael FUJARSKI[a], Lars ECKARDT[b], and Julian VARGHESE[a]
[a] *Institute of Medical Informatics, University of Münster, Germany*
[b] *Department for Cardiology II-Electrophysiology, University Hospital Münster, Germany*

Abstract. Even though the interest in machine learning studies is growing significantly, especially in medicine, the imbalance between study results and clinical relevance is more pronounced than ever. The reasons for this include data quality and interoperability issues. Hence, we aimed at examining site- and study-specific differences in publicly available standard electrocardiogram (ECG) datasets, which in theory should be interoperable by consistent 12-lead definition, sampling rate, and measurement duration. The focus lies upon the question of whether even slight study peculiarities can affect the stability of trained machine learning models. To this end, the performances of modern network architectures as well as unsupervised pattern detection algorithms are investigated across different datasets. Overall, this is intended to examine the generalization of machine learning results of single-site ECG studies.

Keywords. data integration, ECG, machine learning, external validation

1. Introduction

The research trend on decision support systems via machine learning (ML) continues unabated in many disciplines. However, analysis of ML algorithms and pattern recognition for medical problems is subject to strong bias, as it consists mainly of retrospective data that are insufficient for robust clinical application and cannot adequately measure the underlying phenomenon, since they usually consider only one data source. This can be especially problematic when a data collection is not standardized, as in the case of magnetic resonance imaging, where measurements are dependent on the device and sequence [1]. A transfer of trained decision models to other datasets is therefore hardly possible. However, even with standardized data acquisition, which is given in the case of standard electrocardiograms (ECGs) with 12-channel array, 10 seconds duration, and a sampling frequency of 500 HZ, different devices and preprocessing steps potentially alter the outcome. The question arises whether this is sufficient to cause an impact on a trained model. Preliminary work already shows the broad-based data collection and interoperability problems at the level of ECG hardware,

[1] Corresponding Author: Lucas Plagwitz, E-mail: lucas.plagwitz@uni-muenster.de.

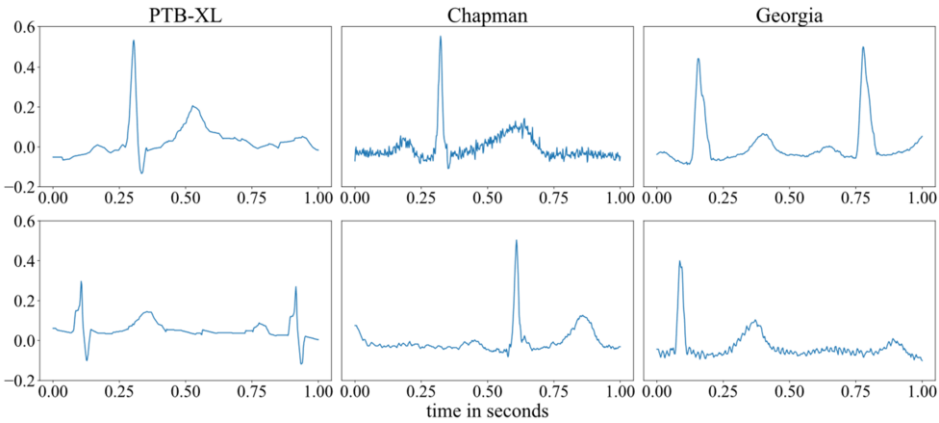

Figure 1. Overview of different ECG measurements depending on the data source. Each column contains two scaled sinus rhythm recordings in lead II of the corresponding data sources (table header).

software, and file formats [2, 3]. Despite this, an in-depth analysis of dataset-specific signal differences is still lacking. To this end, we consider three publicly available data sources that contain quite different ECGs, as shown in Figure 1. Based on these data, we examine the transferability and vulnerability of ML algorithms between these datasets in supervised and unsupervised learning settings.

2. Methods

2.1. Datasets

To investigate comparability between different ECG studies, we investigated the three largest freely available clinical 12-lead ECG datasets (10s length and sampling frequency of 500 HZ) hosted on the PhysioNet online database at the time of this study:

- The PTB-XL dataset contains 21799 ECGs from 18869 patients recorded between years 1989-1996 with Schiller AG equipment [4]. In addition to the raw time signals, information on diagnosis, shape, and rhythm is available.
- Second, a large-scale arrhythmia database from Chapman University is considered, providing ECGs from 45152 patients [5]. These data were collected and stored using devices from General Electric (GE).
- The third dataset, called Georgia, was collected through Emory University, Atlanta, Georgia, and represents a large population of 10344 patients in the southeastern United States. The least is known about this dataset, except that it was provided by Emory University via PhysioNet as part of a 2020 ECG classification challenge [6].

Since the three datasets were not scaled consistently, we applied sample-based abs-max scaling: The entire 12-lead ECG was divided by its highest absolute value. This results in a loss of important information, therefore the performance results are no longer comparable with other studies, but it still enabled us to exclude scaling effects between different data sources.

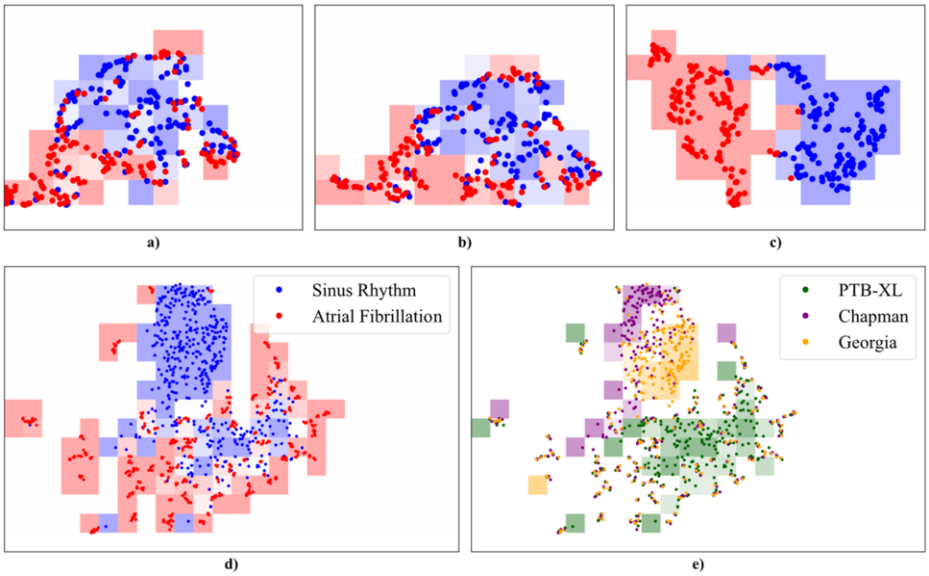

Figure 2. Results of clustering analysis reveal source-specific differences. a) Separation results for PTB-XL samples, labeled according to SR (blue) and AF (red). b) Same analytical procedure as a) based on the Chapman dataset. c) Result of the separation analysis for samples from the Georgia dataset. d) Clustering analysis performed on all datasets simultaneously labeled according to SR and AF. e) Same distribution as in d) relabeled according to the data source. The color of the squares indicates the distribution of the labels within this space.

2.2. Unsupervised analysis

To visualize the influence of possible dataset-specific patterns, we considered two diagnostic subclasses of all datasets: For each dataset, 150 ECGs with labeled atrial fibrillation (AF) and 150 sinus rhythm (SR) ECGs were selected. Measurements were matched in two ways: between diagnoses in a dataset and between data sources in terms of age and sex. Four different tests with the same procedure were calculated based on different data bases: three times for the individual sets and once for all datasets. To compare the scaled ECG data in a meaningful way, we applied a word representation based on the Bag of Symbolic Fourier Approximation Symbols (BOSS) method [7]. This representation was then applied as input to the Uniform Manifold Approximation and Projection (UMAP) algorithm to create a two-dimensional visualization of the data [8].

2.3. Supervised learning

To investigate the model validity of predictions across dataset boundaries, we systematically trained models with a stratified 5-fold cross-validation on one of the datasets, followed by a test of this model on the other two sets. Four different binary classifications were considered: sex, age (>50), AF vs. SR, and first-degree AV block (1AVB) vs. SR. These are all matched as best as possible for age and sex. For the train and test procedure we utilized two convolutional neural networks (CNNs) from PyTorch and tsai projects: a fully convolutional network (FCN) and XceptionTime (XcTime) architecture [9]. The performances were compared using the balanced accuracy score (BACC) based on an iterative scheme whereby training was performed three times on each set and the trained model is additionally tested on the remaining two data sources (known: 3, reference: 6).

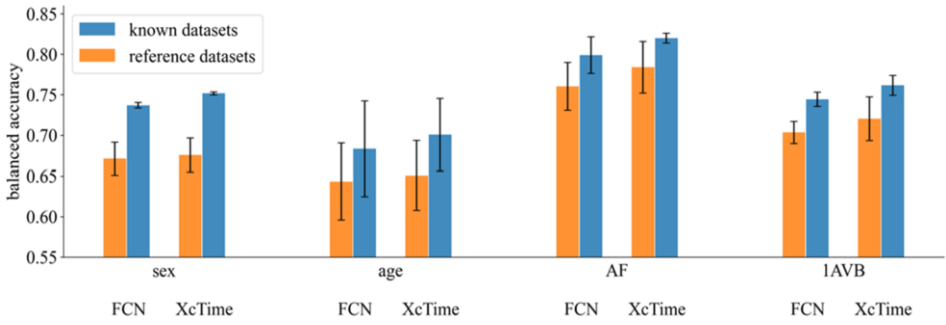

Figure 3. Transferability of classification models across datasets per architecture and label. The blue bar represents the average three test performances based on the known dataset. The orange bar illustrates the average six test performances based on the reference datasets, which source differs from the training set.

3. Results

Figure 2 summarizes the results of the cluster analysis. It was obtainable in the first three subfigures that AF separated well from SR ECGs. Especially on the Georgia dataset, the two labels split almost homogeneously, while PTB-XL and Chapman still had AF measurements within SR regions. In addition to these plots, the cluster distribution was calculated across the three datasets using the same procedure. On the one hand, the AF-SR dependent areas were still recognizable to some extent, but on the other hand, there were separable regions due to the source.

Table 1. List of all investigated ECG datasets with their class distribution per dataset. The matching was done across all data sources.

	sex	**age (> 50)**	**AF**	**1AVB**
dataset distribution	793 / 700	948 / 552	700 / 568	679 / 277
subset	SR (age matching)	SR (sex matching)	SR + AF (age, sex matching)	SR + 1AVB (age, sex matching)

For every label, a sub-dataset was created based on different subsets and properties. Table 1 lists the underlying criteria and class distribution per data source. Two CNN architectures were trained on each of these problems. Figure 3 presents the average model performance based on the unknown and familiar datasets. First, the performances differed between classification label between age (known: 0.7, reference: 0.65) to AF (known: 0.8, reference: 0.76). Furthermore, it was noticeable on every instance that the performance for the known datasets was consistently better compared to the non-trained reference datasets. The difference of known and reference performance varied depending on the label. For example, in the case of sex recognition, the difference in BACC was 0.08, whereas it was 0.04 in the case of AF and 1AVB. Besides this effect, a continuously higher BACC was shown for the XceptionTime model compared to the FCN by about 0.025.

4. Discussion and Conclusion

The results of our experiments showed a clear impact of dataset-specific features on ML algorithms. While a diagnosis-based characterization clustering was observable, delineations depending on the data source clearly emerged. The distinction of AF and SR was possible, but minor diagnosis-based effect sizes could be completely obscured by such artifacts. However, unsupervised methods on ECG data account for only a small fraction of ML applications. According to our research, supervised approaches were influenced by source-specific characteristics as well. We have shown that for four different binary classification tasks, transferability of the model to other datasets was associated with lower predictive performance. This varied between 0.04 - 0.08 BACC depending on the label and architecture. We used a matching procedure to align demographic characteristics between datasets as much as possible. However, annotation- or cohort-specific characteristics (apart from age and sex) could not be eliminated.

Because this effect has significant implications for the application of ML studies to clinical practice, we advocate observing study-specific effects (such as device type or preprocessing steps) when constructing a predictive model. As a by-product, comparable to previous work, we endorse more complex CNN models when it comes to pure predictive performance - even on relatively small sample sizes [10].

Overall, the impact of dataset-specific characteristics on ML algorithms was evaluated for ECG data. Both unsupervised and supervised analyses were revealed to be affected by these side effects. To get one step closer to the underlying medical phenomenon and thus to clinical relevance, we plead for increasingly frequent external validation of study results or models that were already trained across multiple data sources.

References

[1] Biberacher V, Schmidt P, Keshavan A, Boucard C, Righart R, Sämann P, Preibisch C, Fröbel D, Aly L, Hemmer B, Zimmer C, Henry RG, Mühlau M. Intra- and interscanner variability of magnetic resonance imaging based volumetry in multiple sclerosis. Neuroimage. 2016; 142.

[2] Husain K, Zahid MSM, Hassan SU, Hasbullah S, Mandala S. Advances of ECG Sensors from Hardware, Software and Format Interoperability Perspectives. Electronics. 2021; 10, 105.

[3] Stamenov D, Gusev M, Armenski G. Interoperability of ECG standards. 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). 2018.

[4] Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, Schaeffter T. PTB-XL: A Large Publicly Available ECG Dataset. Scientific Data. 2020: 7, 154.

[5] Zheng J, Guo H, Chu H. A large scale 12-lead electrocardiogram database for arrhythmia study (version 1.0.0). PhysioNet. 2022.

[6] Perez Alday EA, Gu A, Shah A, Liu C, Sharma A, Seyedi S, Bahrami Rad A, Reyna M, Clifford G. Classification of 12-lead ECGs: The PhysioNet/Computing in Cardiology Challenge 2020 (version 1.0.2). PhysioNet. 2022.

[7] Schäfer P. The BOSS is concerned with time series classification in the presence of noise. Data Mining Knowledge. 2015; 29.

[8] McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint. 2018 arXiv:1802.03426.

[9] Oguiza I. tsai - A state-of-the-art deep learning library for time series and sequential data. GitHub. 2022; https://github.com/timeseriesAI/tsai.

[10] Doldi F, Plagwitz L, Hoffmann LP, Rath B, Frommeyer G, Reinke F, Leitz P, Büscher A, Güner F, Brix T, Wegner FK, Willy K, Hanel Y, Dittmann S, Haverkamp W, Schulze-Bahr E, Varghese J, Eckardt L. Detection of Patients with Congenital and Often Concealed Long-QT Syndrome by Novel Deep Learning Models. Journal of Personalized Medicine. 2022; 12, 1135.