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.

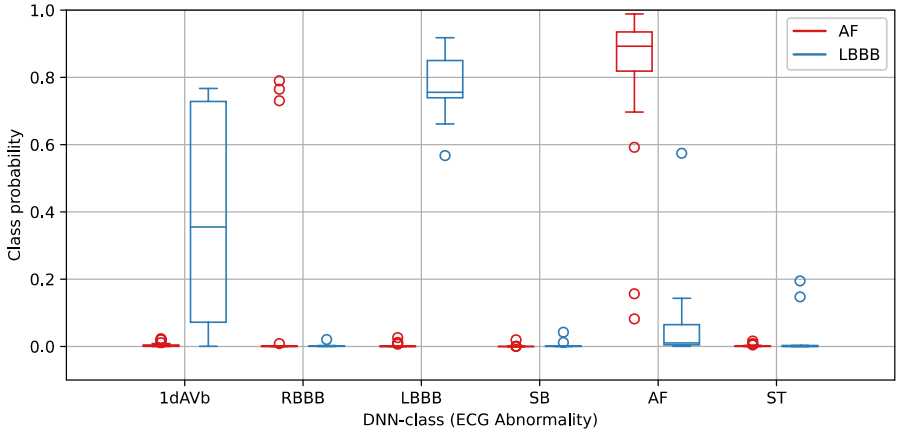

Figure 1. DNN-classification on local data set. Separated into cohorts diagnosed with either atrial fibrillation (red, n=19) or left bundle branch block (blue, n=10), Considered abnormalities are: 1st degree AV block (1dAVb), right bundle branch block (RBBB), left bundle branch block (LBBB), sinus bradycardia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