Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)
© 2023 European Federation for Medical Informatics (EFMI) and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI230099

Post Hoc Sample Size Estimation for Deep Learning Architectures for ECG-Classification

Lucas BICKMANN[a,1], Lucas PLAGWITZ[a] and Julian VARGHESE[a]
[a] *Institute of Medical Informatics, University of Münster, Münster, Germany*
ORCiD ID: Lucas Bickmann https://orcid.org/0000-0001-9043-7844, Lucas Plagwitz https://orcid.org/0000-0001-7626-8853, Julian Varghese https://orcid.org/0000-0002-7206-3719

Abstract. Deep Learning architectures for time series require a large number of training samples, however traditional sample size estimation for sufficient model performance is not applicable for machine learning, especially in the field of electrocardiograms (ECGs). This paper outlines a sample size estimation strategy for binary classification problems on ECGs using different deep learning architectures and the large publicly available PTB-XL dataset, which includes 21801 ECG samples. This work evaluates binary classification tasks for Myocardial Infarction (MI), Conduction Disturbance (CD), ST/T Change (STTC), and Sex. All estimations are benchmarked across different architectures, including XResNet, Inception-, XceptionTime and a fully convolutional network (FCN). The results indicate trends for required sample sizes for given tasks and architectures, which can be used as orientation for future ECG studies or feasibility aspects.

Keywords. machine learning, ecg, sample size, estimation, deep learning

1. Introduction

Twelve-lead electrocardiograms (ECGs) are complex time-series which require a large amount of manual expertise and time for annotation. Still, they give insights for many heart malfunctions and diseases. Automating and increasing the classification of these tasks are an important part for the future of ECG-based precision medicine. With the advancement of machine learning in the recent years, many possibilities arise in life-sciences. However, medical datasets are scarce, but machine learning and especially deep learning preferably require large datasets. Additionally, traditional sample size estimations are hardly applicable for machine learning, and it is arduous for the large variety of architectures and different tasks [1][2]. Pre hoc estimates are based on model parameters, but are challenging to compute, especially for the increasing complexity of architectures. This paper aims to introduce a large-scale post hoc sample size estimation on different architectures for binary classification tasks on ECGs, by computing and fitting the learning curve. This gives researchers insights and guidelines for required samples, as well as the ability to estimate a sample size for new studies of automatic ECG-classification.

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

2. Methods

2.1. Dataset

The training is conducted on the PTB-XL v1.0.2 [3][4][5] with a sampling rate of 100Hz, including 21801 samples. The tasks are binary classification for the PTB-XL diagnostic super-classes Myocardial Infarction (*MI*), Conduction Disturbance (*CD*), ST/T Change (*STTC*), and *sex*. For each classification task, the dataset consists of single labeled *NORM* (healthy controls) and the respective diagnostic superclass. For *sex*-classification, only *NORM* annotated ECGs are selected. The training was conducted on predefined folds 1-8 using shuffled stratified sampling, with a variable number of samples from 100 to a maximum of 4000. To complement the increasing training-samples size, the validation-set increases linearly alongside up to the complete fold at 4000 train-samples. Yet, a minimum of 7.5% of the validation-fold is defined for those train-sample splits which would result in a lower fraction. This estimates dataset splits in real world conditions and constricts the minimal number of samples for a quality validation estimate. The validation- and test-datasets are fold 9 and 10 respectively, as suggested by the authors of PTB-XL, as these include manual curation. The label distribution for the validation and test set are given in Table 1.

Table 1. Normal/anomaly and male/female distribution in the datasets (rounded).

	CD	MI	STTC	Sex
Validation	.187	.255	.278	.972
Test	.202	.281	.265	.880

2.2. Training

The training is conducted with python3.9 using tsai [6]. The chosen architectures are partly based on highest benchmark scores [7]. These include *XResNet1d101*, *InceptionTime*, *XceptionTime* and *FCN*. Each training is conducted 25 times for each number of samples and each architecture, resulting in 225 datapoints for each individual architecture and classification task. The initial learning rate is estimated via an initial run of a learning rate finder [8]. The *1cylce* policy [9] with a maximum number of 500 epochs, early stopping callback with a $\delta = 5e^{-3}$ and a patience of 50 with validation loss as monitoring metric is utilized. The chosen loss function is weighted Cross Entropy. Each sample is standardized independently using batch transformations. The training is conducted on a NVIDIA A40 GPU with a train-batchsize of 1024. Testing and validation are conducted using an equally large batchsize of 1024.

2.3. Evaluation

The half standard deviation around the mean, means (dot-markers) and a logarithmic trendline with $f(x) = a + log_{10}(x) * b$ for each architecture is plotted. A combined average plot is computed via the equally weighted average of all binary classifications. The optimal-threshold-point (x-markers) is given by the highest deviation between the trendline and a linear function, which is computed via the origin and trendline value at 4000 samples, for each architecture respectively. This illustrates the point of maximum score and diminishing returns of the gained performance in relation to extra samples.

3. Results

Figure 1 shows the balanced accuracy score (BACC) for each of the individual selected targets. It stands out that *XceptionTime* is in the two top performers with *InceptionTime*, except for *MI*-classification, for which the latter performs the worst. *XResNet* performs very close, but slightly better to a *FCN* in *CD*, *MI* and *STTC*, while falling behind in *sex*-classification by a small margin. The peak performance is mostly reached for all models at 4000 samples, with an exception. *MI*-classification of *InceptionTime* is stagnating very early at ≈ 1000 samples. The optimal-threshold-point is distributed differently for all architectures and all targets between ≈ 370-1030 train-samples. *XceptionTime* requires most train-samples for all targets, while *InceptionTime* requires the least, except for *CD*-classification.

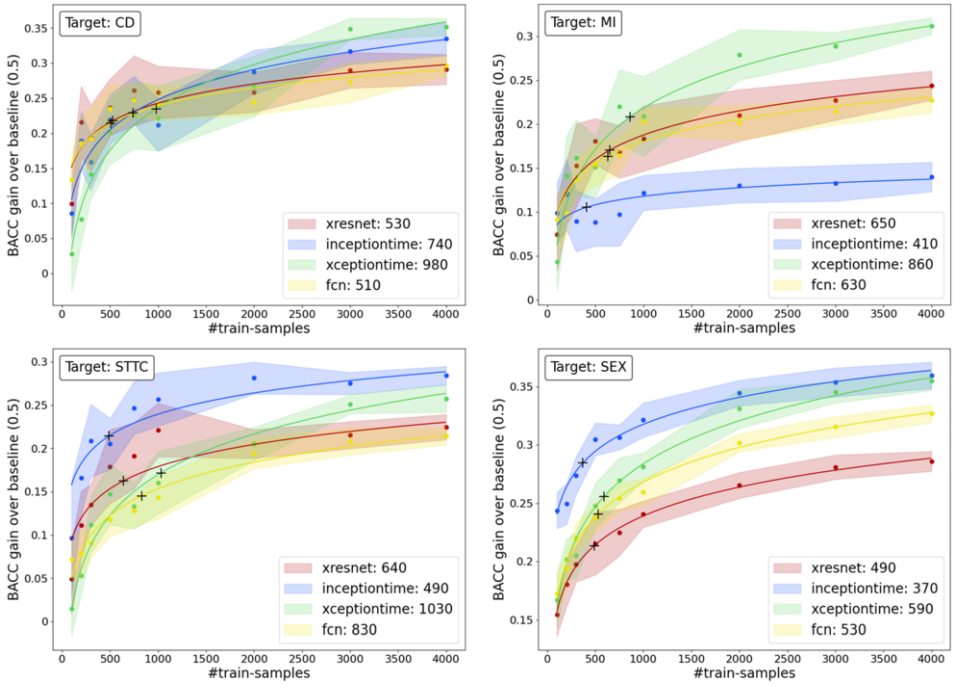

Figure 1. BACC-score gain over baseline of binary classification for different architectures and classification targets: Conduction Disturbance (CD), Myocardial Infarction (MI), ST/T Change (STTC), Sex.

Figure 2 visualizes the combined average performance. It clearly outlines the highest average performing *XceptionTime*. However, the trendline clearly state the necessity of slightly >1000 samples to outperform all other architectures in mean. This fact is underlined with the highest optimal-threshold-point of ≈ 840 train-samples. Yet, *InceptionTime* performs the best in average for ≤ 1000 number of train-samples. At ≈

530-610 samples, all architectures, except *XceptionTime*, perform with the highest BACC/sample-ratio, but still having a rather high standard deviation.

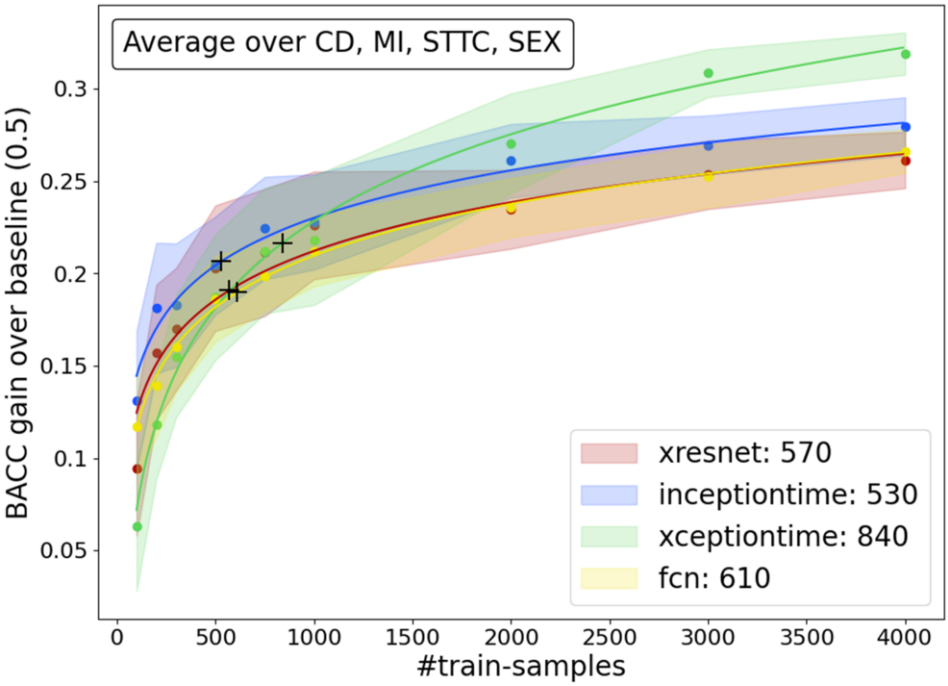

Figure 2. Average BACC-score gain over baseline of binary classification for different architectures.

Table 2 shows the size and label distribution in the validation-set for a train-set with 530 samples. Considering the growing validation-dataset, these additional samples have to be included in the perspective. Therefore, a total of ≥ 650 samples, depending on architecture and target, are required for achieving comparable results, as shown previously.

Table 2. Label distribution in the validation dataset for 530 train-samples (interpolated).

Type	CD	MI	STTC	Sex	Type
Normal	120	120	120	61	Male
Anomaly	22	30	33	59	Female
Total	142	150	153	120	Total

4. Discussion

The highest BACC/sample-ratio is naturally dependent on the classification target, but also on the chosen architecture. Whereas *XceptionTime* excels with a larger number of samples and consistently outperforms other models, *InceptionTime* performs the best in the averaged mean over all targets in the range of 100-1000 train-samples. Yet, it performed worst in *MI*-classification. Therefore, we advise to conduct training with at least two architectures for a specific target, to double-check inconsistencies for the specific task.

5. Conclusion

This paper shows results and a guideline for a preferred minimum number of samples, which yield the highest per-samples-scores. For a lower number of samples (< 1000), *InceptionTime* performs the best, otherwise *XceptionTime* excels. We suggest a minimal train-size of ≥ 530 samples for most applications, to exploit the highest BACC/sample-ratios. The train- and validation-set should therefore contain at least ≈ 650 samples combined. Test-samples are not included in this approximation. We only evaluated binary classification as a first step, and it remains interesting which outline can be drawn for multi-class, -label and regression tasks in future research. Additional apprehension could be drawn for other targets and additional datasets as well. We suggest another study with additional tasks, such as regression, using the same environment and parameters. To summarize, this sample size estimation for binary classification tasks on electrocardiograms indicate helpful guidelines for further research.

References

[1] Balki I, Amirabadi A, Levman J, Martel AL, Emersic Z, Meden B, et al. Sample-Size Determination Methodologies for Machine Learning in Medical Imaging Research: A Systematic Review. Canadian Association of Radiologists Journal. 2019;70(4):344-53. Available from: https://doi.org/10.1016/j.carj.2019.06.002.

[2] Du SS, Wang Y, Zhai X, Balakrishnan S, Salakhutdinov RR, Singh A. How Many Samples are Needed to Estimate a Convolutional Neural Network? In: Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc.; 2018. Available from: https://proceedings.neurips.cc/paper/2018/file/03c6b06952c750899bb03d998e631860-Paper.pdf.

[3] Wagner P, Strodthoff N, Bousseljot R, Samek W, Schaeffter T. PTB-XL, a large publicly available electrocardiography dataset (version 1.0.2). PhysioNet. 2022 Aug. Available from: https://doi.org/10.13026/zx4k-te85.

[4] Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze F, Samek W, et al. PTB-XL, a large publicly available electrocardiography dataset. Sci Data [Internet]. 2020 [cited 2022 Nov 30];7((1):154). Available from: https://www.nature.com/articles/s41597-020-0495-6.

[5] Goldberger A, Amaral L, Glass L, Hausdorff J, Ivanov P, Mark R, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation [Online]. 2000 Jun;101(23):e215–e220.

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

[7] Strodthoff N, Wagner P, Schaeffter T, Samek W. "Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL,". IEEE Journal of Biomedical and Health Informatics. 2021 May;25(5):1519-28. Available from: https://doi.org/10.1109/JBHI.2020.3022989.

[8] Smith LN. Cyclical Learning Rates for Training Neural Networks. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV); 2017. p. 464-72.

[9] Smith LN, Topin N. Super-Convergence: Very Fast Training of Residual Networks Using Large Learning Rates. CoRR. 2018 May;abs/1708.07120. Available from: http://arxiv.org/abs/1708.07120.