Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240483

Benchmarking Approaches: Time Series Versus Feature-Based Machine Learning in ECG Analysis on the PTB-XL Dataset

Lucas BICKMANN[a,1], Lucas PLAGWITZ[a] and Julian VARGHESE[a]
[a]*Institute of Medical Informatics, University of Münster, Germany*

Abstract. Extensive research has been conducted on time series and tabular data in the context of classification tasks, considering their distinct data domains. While feature extraction enables the transformation of series into tabular data, direct comparative comparisons between these data types remain scarce. Especially in the domain of medical data, such as electrocardiograms (ECGs), deep learning faces challenges due to its lack of easy and fast interpretability and explainability. However, these are crucial aspects for a wide and reliable adoption in the field. In our study, we assess the performance of XGBoost and InceptionTime on ECG features and time series data respectively. Our findings reveal that features extracted from ECG signals not only achieve competitive performance but also retain advantages during training and inference. These advantages encompass accuracy, resource efficiency, stability, and a high level of explainability.

Keywords. Machine Learning, Explainable AI, Electrocardiogram, Explainability

1. Introduction

In recent advancements, deep learning, especially convolutional neural networks (CNNs), have witnessed significant adoption for ECG classification [1], while traditional machine learning techniques have seen a general decline in focus. Previous research shows that traditional tree ensemble methods are often superior to deep learning approaches on tabular data [2,3], however, these analyses only cover the same domain. A recent study [4] shows that a CNN outperforms classification built on top of automated Feature Extraction, in the field of ECG data. However, our study aims to assess the performance of time-series data of ECGs against specialized extracted features based upon UniG [5]. Compared to time-series data, feature extraction is less susceptible to batch effects, reduces dimension and therefore speeds up training [6]. We show that specialized feature extraction combined with an explainable machine learning classifier outperforms deep learning methods. Additionally, these features provide direct insights, without the need for intrinsic waveform analysis. This enables the use of explainable clinical decision support systems, with benefits of machine learning based methods, while being transparent in the decision process.

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

2. Methods

PTB-XL v1.0.2 [7–9] and the associated PTB-XL+ v1.0.1 [10,11] database, which cover over 21801 ECGs with annotations and features, are used to train and evaluate machine learning models. Each ECG is 10 seconds long, with 12-leads and a sampling rate of 500hz. We split the database according to the supplied stratified folds, utilizing fold ten for the computation and generation of results for each method. This allows fair and comparable evaluation between all methods. We conduct a multiclass-classification using the annotated superclasses of the dataset: Myocardial Infection, Conduction Disturbance, Hypertrophy, ST/T-Changes and normal.

For XGBoost [12], we use tabular based data, which includes all ECGs with extracted features from the University of Glasgow (UniG) ECG analysis program [5], based on PTB-XL+ (Table 1). An extensive Halving Grid Search was conducted using the following parameter-space with a 5-fold cross-validation using sample-weights and balanced accuracy as optimization parameter:

Table 1. XGBoost optimization parameters applied in a cross-validation Halving Grid Search.

parameter	max_depth	n_estimators	min_child_weight	gamma	colsample_bytree
values	[1, 2, 3, 4, 5, 6]	[50, 100, 150, 500]	[1, 5, 10, 15]	[0.5, 1, 1.5]	[0.4, 0.6, 0.8]

InceptionTime [13] is trained on an 80/20-split for a train- and evaluation-dataset of the raw 12-lead ECG time series data. As not all time series have extracted features in the extended dataset, we show the performance on the complete and the matched subset of UniG-feature data. The ECGs are normalised with a global mean/std scaling. The used optimizer is AdamW with weighted CrossEntropyLoss and a ReduceLROnPlateu-Scheduler based on the validation loss. We train InceptionTime with different parameters, such as the number of filters (32, 48, 64), batch-sizes (8, 16) and initial learning rate (1e-3, 1.5e-3, 1e-4), three times each for up to 20 Epochs. The best performing model is presented in this work (n_filters=48, batch_size=8, lr=1.5e-3).

Table 2. Number of available data points in the combined training and evaluation dataset for each model, along the distribution across different classes. The feature chance level indicates the percentage of data points in the extended dataset with extracted features.

Model	Myocardial Infarction	Conduction Disturbance	Hypertrophy	ST/T-Change	Normal
XGBoost (UniG)	796	576	215	871	3692
InceptionTime (Fair)					
InceptionTime (All)	1673	1276	455	1893	8159
Available annotations (average 46.25%)	47.58%	45.14%	47.25%	46.01%	45.25%

We compute a Receiver-under-Operator-Curve (ROC), outlining the performance of sensitivity vs. specificity under a certain threshold. Due to conducting multiclass-classification, we compute the micro-weighted Area under Curve (AUC) to account for class imbalances, and also outline the macro-AUC and balanced accuracy (BACC). We also compute combined approaches for the whole dataset, as well as for subsets of features only (Table 2). The whole dataset is computed by using UniG predictions where applicable, and otherwise InceptionTime (both variants). For the combined subset model, we use the average of both, XGBoost and InceptionTime. To generate explainability-plots

for XGBoost, we use SHAP [15] to compute example plots for individual assessments with the UniG-feature based model.

3. Results

3.1. Classification Results

Table 3 outlines the balanced accuracy, weighted-macro and -micro One-vs-Rest ROC for XGBoost trained on different features. Figure 1 plots the ROC for all three models. It can be clearly outlined that UniG features result in higher performance than InceptionTime on the same data. InceptionTime performs worse in BACC as well as in both AUC metrics. As shown in previous work [16], this model achieves comparable multiclass performance without extensive finetuning optimizations. We also see a marginal better performance of combining the feature-based predictions, if available, with a time-series-based classification model for the remainder data.

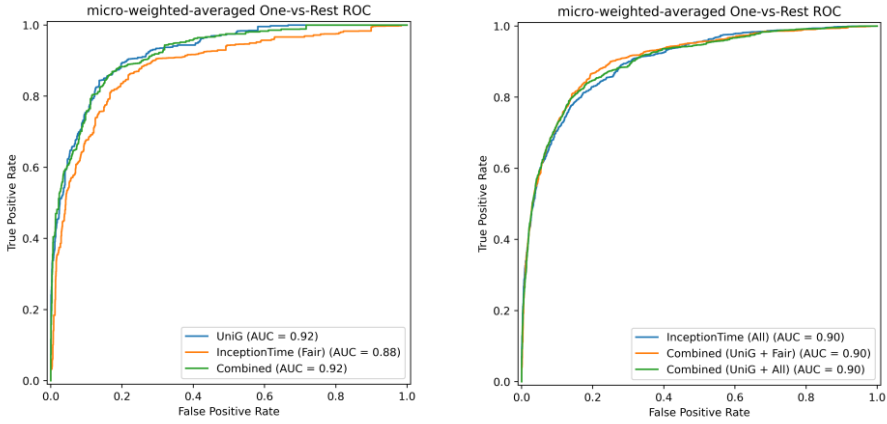

Figure 1. Micro-weighted-averaged One-vs-Rest Receiver-under-Operator-Curve (ROC) outlining the performance of all models and their respective Area-Under-Curve (AUC). UniG-feature subset (left) and complete test set performance (left) and with their corresponding machine learning models.

Table 3. Performance metrics for XGBoost and InceptionTime using the corresponding features. Highest performing metrics are bold with the follow-up underlined. The subset indicates if they were trained and evaluated on the same data points which are in the UniG-feature subset.

			weighted OvR AUC	
Features	Subset	BACC	macro	micro
UniG	x	**0.709**	**0.926**	**0.922**
InceptionTime (Fair)	x	0.644	0.882	0.882
Combined (UniG + Fair)	x	<u>0.691</u>	<u>0.921</u>	<u>0.919</u>
InceptionTime (All)		0.664	**0.907**	0.897
Combined (UniG + Fair)		**0.682**	<u>0.902</u>	**0.903**
Combined (UniG + All)		<u>0.681</u>	<u>0.902</u>	<u>0.898</u>

3.2. Explainability

While InceptionTime explainability can solely rely on graphical representation of the time-series, XGBoost can be well analysed using feature importance such as SHAP. Figure 2 represents the local feature importance for two distinct classes based on two samples. This analysis can be conducted for each class, and each prediction separately, which allows high transparency and explainability.

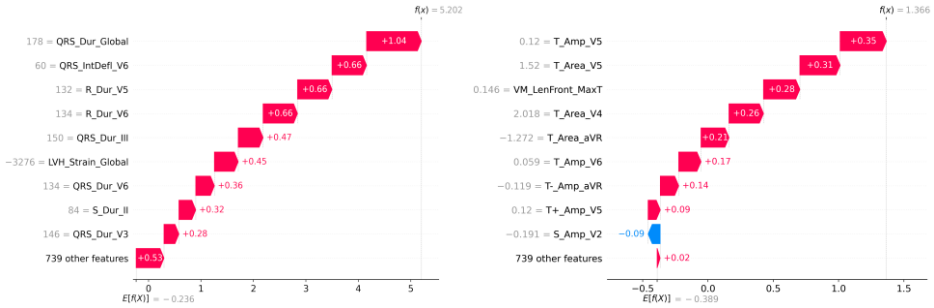

Figure 2. SHAP waterfall plot for two example data points. The features outline their corresponding impact on the models decision for their respective classes Conduction Disturbance (left) and ST/T-Change (right).

4. Discussion

Another aspect, not numerically evaluated, is the training-time, -resources and -stability. While XGBoost can be optimised comparably fast, even on a CPU, deep learning requires much more hardware sophistication. Additionally, deep-learning, which inherently relies on random initialization and numerical stability, varies a models performance based on the number of features and complexity. It also outlines its stability, as our findings assessed performance differences of more than 6% for InceptionTime using the same training data, while XGBoost remained more stable and reproducible using the same parameters. Additionally, studies have demonstrated the occurrence of batch effects related to the capturing device and calibration [17]. Computing features which are less prone to these effects, could improve the inter-dataset classification performance, addressing a problem common in other methods.

However, as stated, not all ECGs have computed features, which can limit the applicability of this approach. This could be the case, as no important markers, such as QRS-points, could be detected based on the used algorithms. If they can be computed, they exhibit better performance in our findings, while deep-learning is more feasible on a wider variety of abnormal ECGs which hinder feature extraction. Nonetheless, a combined approach shows better performance than a model trained on a larger number of samples, as its performance uplift is limited. The explainability achieved by using features also outlines the potential for use in the field, as the decisions are easier to trace and validate by experts. It could help clinicians to assess a diagnosis faster while being based upon reliable and meaningful attributes.

5. Conclusions

Our work outlined that extracted features of ECG data is competitive to an extensively used deep learning model, while maintaining several advantages. The reproducibility, as a one of the key factors for open science, its fast convergence and detailed explainability are crucial benefits for researchers. The feature-based attribution is retractable and allows full transparency. These key elements empower clinicians to assess the findings while maintaining trustworthiness over black box decision-support-systems.

References

[1] Deep learning and the electrocardiogram: review of the current state-of-the-art - PMC [Internet]. [cited 2024 Mar 14]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8350862/.
[2] Borisov V, Leemann T, Seßler K, et al. Deep Neural Networks and Tabular Data: A Survey. IEEE Trans Neural Netw Learn Syst. 2022;1–21.
[3] Shwartz-Ziv R, Armon A. Tabular data: Deep learning is not all you need. Inf Fusion. 2022;81:84–90.
[4] Middlehurst M, Schäfer P, Bagnall A. Bake off redux: a review and experimental evaluation of recent time series classification algorithms. Data Min Knowl Discov [Internet]. 2024 [cited 2024 May 14]; Available from: https://doi.org/10.1007/s10618-024-01022-1.
[5] Macfarlane PW, Devine B, Clark E. The university of glasgow (Uni-G) ECG analysis program. Comput Cardiol 2005 [Internet]. 2005 [cited 2024 Mar 5]. p. 451–454. Available from: https://ieeexplore.ieee.org/document/1588134.
[6] Subasi A. Chapter 4 - Feature Extraction and Dimension Reduction. In: Subasi A, editor. Pract Guide Biomed Signals Anal Using Mach Learn Tech [Internet]. Academic Press; 2019 [cited 2024 Mar 6]. p. 193–275. Available from: https://www.sciencedirect.com/science/article/pii/B9780128174449000040.
[7] Wagner P, Strodthoff N, Bousseljot R-D, et al. PTB-XL, a large publicly available electrocardiography dataset [Internet]. [cited 2024 Mar 5]. Available from: https://physionet.org/content/ptb-xl/1.0.2/.
[8] Wagner P, Strodthoff N, Bousseljot R-D, et al. PTB-XL, a large publicly available electrocardiography dataset. Sci Data. 2020;7:154.
[9] Goldberger AL, Amaral LAN, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet. Circulation. 2000;101:e215–e220.
[10] Strodthoff N, Mehari T, Nagel C, et al. PTB-XL+, a comprehensive electrocardiographic feature dataset [Internet]. [cited 2024 Mar 5]. Available from: https://physionet.org/content/ptb-xl-plus/1.0.1/.
[11] Strodthoff N, Mehari T, Nagel C, et al. PTB-XL+, a comprehensive electrocardiographic feature dataset. Sci Data. 2023;10:279.
[12] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min [Internet]. New York, NY, USA: Association for Computing Machinery; 2016 [cited 2024 Mar 5]. p. 785–794. Available from: https://dl.acm.org/doi/10.1145/2939672.2939785.
[13] Ismail Fawaz H, Lucas B, Forestier G, et al. InceptionTime: Finding AlexNet for time series classification. Data Min Knowl Discov. 2020;34:1936–1962.
[14] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Proc 31st Int Conf Neural Inf Process Syst. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 4768–4777.
[15] Bickmann L, Plagwitz L, Varghese J. Post Hoc Sample Size Estimation for Deep Learning Architectures for ECG-Classification. Stud Health Technol Inform. 2023;302:182–186.
[16] Plagwitz L, Vogelsang T, Doldi F, et al. The Necessity of Multiple Data Sources for ECG-Based Machine Learning Models. Stud Health Technol Inform. 2023;302:33–37.