An AutoML Pipeline for the Classification of Noisy and Clean Electrocardiogram Signals Using Bayesian Optimization

Lucia Billeci[1], Lorenzo Bachi[1], Chiara Podrecca[2], Maurizio Varanini[1], Pia Cincione[3]

[1]Institute of Clinical Physiology, National Research Council of Italy (CNR-IFC), Pisa, Italy
[2]Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy
[3]Medical Department, University of Foggia, Foggia, Italy

Abstract

Electrocardiogram (ECG) recordings are affected by notorious source of noise which can limit the reliability of the clinical decisions. Several methods have been implemented to automatically identifying noisy signals. Recently Machine Learning (ML) has helped in the automatic classification of ECG signals. In particular, Automated Machine Learning (AutoML), allows to overcome the limits of more common ML approaches.

In this study, we aimed to implement an AutoML pipeline able to classify ECG signals as noisy or clean. Several databases were considered which were annotated on a 2 s window basis. After pre-processing and balancing the dataset, AutoGluon was applied for signals classification. For the model with the best trade-off between accuracy and inference time, Bayeasian Hyperparameter Optimization (HPO) was then applied. The Neural Network model was the best performing model on both the training and test set. HPO increased the performance without increasing significantly the inference time.

This study showed the efficiency of an AutoML pipeline for ECG signal classification encouraging its applicability in a clinical setting.

1. Introduction

Electrocardiogram (ECG) recordings are susceptible to a myriad of elements that can plague and obfuscate the true cardiac signal. The magnitude of the changes must be substantial in order to significantly alter a named ECG feature to result in a clinical diagnosis [1].

Techniques rooted in the frequency domain are typically employed to mitigate issues like powerline interference and baseline drift. Challenges persist with muscular noise and movement artifacts due to their frequency overlap with genuine ECG signals. Moreover, ECG signal variability and its morphology change over time and are dependent on each individual. Due to its dynamic nature, the ECG exhibits stochastic and nonstationary behavior [2].

Considering all these aspects, it emerges that the study of ECG signals is a time-consuming task, with a high probability of evaluation mistakes. The most frequent issues related to ECG analysis are the prediction or identification of cardiac abnormalities or localization of cardiac pathological events, but if noise and artifacts are not accurately removed, they inevitably degrade the accuracy of the analysis report, potentially leading to false positives or missed diagnoses in the clinical assessment. Therefore, there is a great effort to develop new methods for processing and analyzing ECG records.

In this context, numerous studies for the classification of ECG signals and, in detail, the identification of noise segments through Machine Learning and Deep Learning methodologies are included [3].

Notably, Automated Machine Learning (AutoML) represents a paradigm shift in overcoming the traditional challenges associated with Machine Learning. By automating the entire model development process, AutoML not only facilitates the exploration of a more extensive set of methodological approaches but also significantly optimizes the time and quality of mode development. The fundamental advantage of AutoML lies in its ability to systematically evaluate numerous models and configurations, simplifying the identification of the most effective solutions and avoiding the need for costly manual parameter tuning, which characterizes traditional ML methodologies [4]. This automation, therefore, increases the likelihood of discovering the most effective models with less manual effort and in less time [5].

Thus, the aim of this study was to develop an AutoML pipeline for noise detection in ECG signals.

2. Methods

2.1. Dataset

The dataset examined consists of a compilation of ECG signals sourced from various datasets, with a predominant amount of data being authentic ECG signals, while a portion consisting of simulated signals:

- The Physionet/Computing in Cardiology Challenge 2011 Database, featuring a collection of 1500 standard 12-lead ECG records, each 10 sec long, sampled at 500 Hz.

- The MIT-BIH Malignant Ventricular Database, containing 22 2-lead Holter ECG records, each lasting 30 minutes, sampled at 250 Hz.
- Cardioline S.p.A (Italy)'s Private R&D Database, a private database owned by Cardioline SpA. This database includes 20 records of either 3-lead or 12-lead Holter ECG records, 20 minutes long each, sampled at 250 Hz.
- A simulated database consisting of 60 12-lead ECG records, each 1 minute long, sampled at 1000 Hz. Data was synthesized with the simulator described in [6].

2.2. Data labeling

Each of these databases was labeled by an expert human observer regarding the presence or absence of noise. For each database, the same procedure was performed, as described below. Each independent lead was labeled over windows of 2 seconds, independently of the others. Only independent leads were labeled, meaning that leads III, aVR, aVF, and aVL of 12- lead records were discarded.
The observer assigned a binary outcome Y to each 2 s ECG record k, with "class = 0" denoting clean ECG and "class=1" denoting noise.

2.3. Noise indices

A set of twelve noise indices were considered in this study. Ten of these indices were established noise indices in the literature: kurtosis (kur) [7], skewness (skew) [7], relative power in the QRS complex (rpow) [8], relative power in the baseline (bas) [7], in-band to out-of-band spectral power ratio (ior) [9], modulation spectrum quality index (msqi) [10], signal-to-noise ratio (snr) [11], signal-to-noise ratio computed through wavelet decomposition (wave) [12], high order statistic (hos) [13], sample entropy (se) [14].

Additionally, two novel noise indices designed by the authors were evaluated: i. ECG derivative pattern index (edp), which quantifies the presence of ECG by exploiting its main characteristic: quasi-periodic impulsive signal. ECG is considered as sequence of high-energy events (QRS). QRSs are characterized by high derivative, occur with a period approximately in the range of 0.5-1.5 s and have a duration approximately in the range of 0.05-0.2 s, ii. signal slope inversions over a minimum threshold (inv) – based on the observation that clean ECG features few derivative sign changes.

2.3. Data preparation

The dataset is heavily imbalanced with respect to the target class: the distribution is significantly skewed towards clean signals, marked as clean (class=0), which constitute 76% of the entire dataset, while the noisy signals account for 24%. The dataset presents a further layer of complexity due to the interrelated nature of its rows: it encompasses 1,602 unique patients, with each individual contributing multiple samples obtained by segmenting the ECG signal into 2-second intervals.

For database splitting, we adopted a data division approach where 70% of the dataset was originally allocated to the training set and 30% was set aside for testing. Furthermore, from the training set, we extracted an additional 15% of the data to create the validation set, resulting in a final distribution of 55% for training, 15% for validation, and 30% for testing.

To prevent any cross-distribution of data, we identified and separated the individual patient records before splitting: this ensures that each patient's data were exclusively allocated to one set at a time, thus maintaining the purity of the training set and the robustness of the validation and testing process.

Since the data were imbalanced, we opted for Partial Balancing strategy by using the RandomUnderSampler function: we reduced the number of instances in the majority class to 1.4 times the number of instances in the minority class.

2.3. AutoML pipeline

Among the various AutoML frameworks available, our study has intentionally chosen to utilize AutoGluon, an AutoML toolkit developed by the Amazon Web Services (AWS) team and released under an open-source license [15].

AutoGluon stands out for two primary features: its Quality Preset options and a cutting-edge Multi-layer Stacking Ensemble Technique.

AutoGluon uses user-friendly 'presets', predefined configurations of models and hyperparameters that facilitate the balancing act between training duration, model quality and the consumption of computational resources. Specifically, AutoGluon offers four distinct presets: medium_quality, good_quality, high_quality, and best_quality.

AutoGluon introduces a novel form of Multi-Layer Stack Ensembling [15] exemplified in Figure 1.

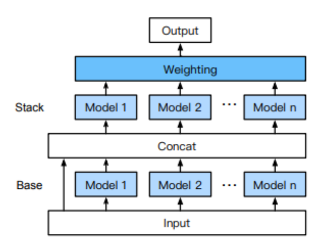

Figure 1. AutoGluon's multi-layer stacking strategy.

In our binary classification task, we set up and trained four distinct 'TabularPredictor' instances, each corresponding to one of the four AutoGluon presets, and Balanced Accuracy (BA) served as chosen metric for evaluating performance.

To inspect the training results, we utilized the leaderboard() command to gather information about the models produced during the fit() process, along with their performance metrics. Considering the top 5 models across the 4 distinct presets, we proceeded to compare their performances, both within each preset and among the presets.

For the model with the best trade-off between accuracy and inference time, we performed Bayesian Hyperparameter Optimization (HPO).

3. Results

3.2. Comparison of different presets

Within each preset, the top 5 models demonstrate closely comparable accuracies, albeit with varying inference and fitting times (Figure 2). Notably, across all presets, the most outstanding model consistently emerges as an Ensemble model.

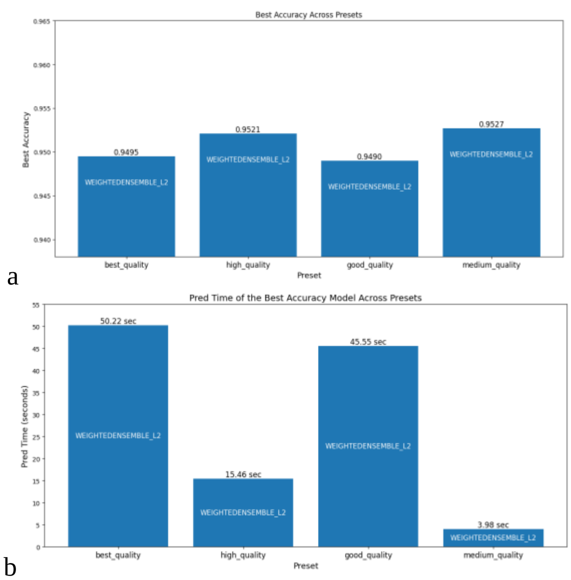

Figure 2. Ensemble model accuracy (a) and prediction time (b) by quality preset.

The NeuralNetTorch model from the medium quality preset is particularly interesting: it surpasses the accuracies of the top models from other presets, with one of the lowest inference times both on the training and test set (Table1).

Table 1. Performance of the best models of the medium quality preset before HPO.

Model	BA	Prediction time (s)
Training set		
Weighted Ensamble	0.952	3.979
NeuralNetTorch	0.949	0.138
Test set		
Weighted Ensamble	0.947	2.638
NeuralNetTorch	0.945	0.131

3.2. Hyperparameters optimization

Based on the findings of the presets' comparison, we decided to perform a Bayeasian HPO on NeuralNetTorch model from medium preset in an attempt to further improve its performance. Notably, AutoGluon perform HPO on this model with different configurations and on an ensemble of these models.

The comparison between "Pre HPO" and "Post HPO" suggests that the optimization process has fine-tuned the NeuralNetTorch model. In particular, on the training set, compared to the pre-HPO NeuralNetTorch model, all 5 top models show an improvement in accuracy, with the best score increasing to 0.960.

We then evaluated the performance of the top three models after HPO on the test set (Table 2).

Table 2. Performance of the best models of the medium quality preset after HPO.

Model	BA	Prediction time (s)
Training set		
Weighted Ensamble	0.960	1.313
NeuralNetTorch	0.955	0.605
Test set		
Weighted Ensamble	0.952	0.563
NeuralNetTorch	0.950	0.221

4. Discussion and conclusions

In this study we implemented an AutoML pipeline to discriminate noisy from clean ECG recordings.

The results of our study showed that the performance of the tested models was good both in term of accuracy and inference time.

Notably, across all presets, the most outstanding model consistently emerges as an Ensemble model, highlighting the effectiveness of AutoGluon's Multi-Layer Stacking Ensemble technique. While ensemble models exhibit strong performance, they also show longer times and higher resource consumption, thus confirming the trade-off between model complexity, time, and accuracy.

The second top performing model is a neural networks-based model called NeuralNetTorch. Indeed, neural networks are able to handle complex, high-dimensional data and to model intricate patterns and for these reasons are increasingly used in the medical field [16]. The stability of these models is also confirmed by the fact that performance remains stable when running the model on the training and test set.

Surprisingly enough, the top performing models were not from the best quality preset, but from the medium quality preset. The efficiency of the medium quality preset suggests that for this particular dataset and set up, more complex models do not significantly outperform in terms of validation accuracy and inference time.

When we applied HPO the performance of the models increased, demonstrating that the combination of AutoML and Post-Tuning can be a winning strategy, even in computational development contexts that are not excessively high-performing. In particular, Bayesian optimization is a very effective algorithm specifically in complex and noisy optimization tasks outperforming other global optimization algorithms [17].

Further research could build on this work to explore more targeted feature selection and to broaden the analysis to more diverse noisy data and signals with arrhythmias, thus widening the study's applicability.

Acknowledgments

We would like to thank Cardioline S.p.A, Italy for supporting this work and for providing part of the dataset.

References

[1] L. Sörnmo, P. Laguna, “Chapter 7 - ECG Signal Processing”, in: L. S.rnmo, P. Laguna (Eds.), Bioelectrical Signal Processing in Cardiac and Neurological Applications, Biomedical Engineering, Academic Press, Burlington, pp. 453–566, 2005.

[2] M. D’Aloia, A. Longo, M. Rizzi, "Noisy ECG Signal Analysis for Automatic Peak Detection" Information, vol. 10, no. 2, Art. no. 35, 2019.

[3] K. van der Bijl, M. Elgendi, C. Menon, “Automatic ECG Quality Assessment Techniques: A Systematic Review”, Diagnostics, vol. 12, no. 11 Art. no. 2578, 2022.

[4] F. Hutter, L. Kotthoff, J. Vanschoren J. Automated Machine Learning: Methods, Systems, Challenges. New York: Springer; 2019.

[5] Q. Yao, M. Wang, Y. Chen, W. Dai, Y.F. Li, W.W. Tu, Q. Yang, Y. Yu, “Taking Human out of Learning Applications: A Survey on Automated Machine Learning”, arXiv preprint arXiv:1810.13306, 2018.

[6] L. Bachi, H. Halvaei, C. Perez, A. Martin-Yebra, A. Petrenas, A. Sološenko, L. Johnson, V. Marozas, J. P. Mart.nez, E. Pueyo, M. Stridh, P. Laguna, L. Sörnmo, “ECG Modeling for Simulation of Arrhythmias in Time-Varying Conditions”, IEEE Transactions on Biomedical Engineering, vol. 70, no. 12, pp. 3449–3460, 2023.

[7] G. D. Clifford, J. Behar, Q. Li, I. Rezek, “Signal Quality Indices and Data Fusion for Determining Clinical Acceptability of Electrocardiograms”, Physiological Measurement, vol. 33, no. 9, Art. no. 1419, 2012.

[8] Q. Li, R. G. Mark, G. D. Clifford, “Robust Heart Rate Estimation from Multiple Asynchronous Noisy Sources Using Signal Quality Indices and A Kalman Filter”, Physiological Measurement, vol. 29, no. 1, Art. no. 15, 2007.

[9] G. D. Clifford, F. Azuaje, P. McSharry, “ECG Statistics, Noise, Artifacts, and Missing Data, Advanced Methods and Tools for ECG Data Analysis”, vol. 6, no. 1, pp. 55–99, 2006.

[10] D. P. Tobon V., T. H. Falk, M. Maier, “MS-QI: A Modulation Spectrum-Based ECG Quality Index for Telehealth Applications”, IEEE Transactions on Biomedical Engineering, vol. 63, no. 8, pp. 1613–1622, 2016.

[11] S. Rahman, C. Karmakar, I. Natgunanathan, J. Yearwood, M. Palaniswami, “Robustness of Electrocardiogram Signal Quality Indices”, Journal of The Royal Society Interface, vol. 19, no. 189, Art. No. 20220012, 2022.

[12] L. Smital, C. R. Haider, M. Vitek, P. Leinveber, P. Jurak, A. Nemcova, R. Smisek, L. Marsanova, I. Provaznik, C. L. Felton, B. K. Gilbert, D. R. Holmes III, “Real-Time Quality Assessment of Long- Term ECG Signals Recorded by Wearables in Free-Living Conditions”, IEEE Transactions on Biomedical Engineering, vol. 67, no. 10, pp. 2721–2734, 2020.

[13] H. Tan, J. Lai, Y. Liu, Y. Song, J. Wang, M. Chen, Y. Yan, L. Zhong, Q. Feng, W. Yang, “Neural Architecture Search for Realtime Quality Assessment of Wearable Multi-Lead ECG on Mobile Devices”, Biomedical Signal Processing and Control, vol. 74, Art. No. 103495, 2022.

[14] M. Li, Y. Qin, F. Gao, W. Zhu, X. He, “Discriminative Analysis of Multivariate Features from Structural MRI and Diffusion Tensor Images”, Magnetic Resonance Imaging, vol. 32 no. 8, pp. 1043–1051, 2014.

[15] N. Erickson, J. Mueller, A. Shirkov, H. Zhang, P. Larroy, M. Li, A. Smola, “AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data”, arXiv preprint, arXiv:2003.06505, 2020.

[16] N. Shahid, T. Rappon, W. Berta “Applications of Aneural Networks in Health Care Organizational Decision-Making: A Scoping Review”, PLoS One. vol. 14, no. 2, Art. No. 2e0212356.

[17] D.R. Jones, “A Taxonomy of Global Optimization Methods Based on Response Surfaces”, Journal of Global Optimization, vol. 21, pp. 345–383, 2001.

Address for correspondence:

Lucia Billeci
via Moruzzi 1, 56124, Pisa, Italy
lucia.billeci@cnr.it