Feature Contributions to ECG-based Heart-Failure Detection: Deep Learning vs. Statistical Analysis

Agnese Sbrollini[1], Chiara Leoni[1], Marjolein C. de Jongh[2], Micaela Morettini[1], Laura Burattini[1], Cees A. Swenne[2]

[1]Università Politecnica delle Marche, Ancona, Italy
[2]Leiden University Medical Center, Leiden, Netherlands

Abstract

Assessing feature contributions to a specific diagnosis is commonly done by statistical analysis. In the context of heart failure (HF) diagnosis from the electrocardiogram (ECG), this work compares feature contributions assessed by deep learning with those obtained by statistical analysis. Data consists of ECG pairs (baseline and follow-up) from patients with a history of myocardial infarction. When the follow-up ECG was made, controls patients had remained stable, while cases patients had developed HF. The 42 features that characterized each ECG served as inputs of a deep-learning neural network (NN) created by our Repeated Structuring & Learning Procedure. Subject-specific feature ranking was obtained from the local-interpretable model-agnostic explanatory algorithm and processed to obtain feature relevances (FR). Additionally, 42 areas under the curve (AUC) by univariate statistical analysis were obtained. FR and AUC were compared by Pearson's correlation coefficient (ρ). After training, the NN had a 99% classification performance. FR ranged from 0.32 to 4.47; AUC ranged from 23% to 82%. Correlation analysis yielded no significant association between AUC and FR (ρ=0.18, P-value=0.25). Deep-learning and statistical-analysis feature contributions to HF detection were discordant. Further studies will investigate which of the two approaches better reflects clinical interpretation.

1. Introduction

The 12-lead 10-s resting electrocardiogram (ECG) is a standard measurement in the evaluation of patients, especially in cardiology. The ECG, in fact, contains detailed information about the electrical heart action [1]. Thus, clinical ECG interpretation aims to determine if ECG features (wave morphologies, intervals) are normal or pathological [2]. Knowing which ECG features contribute to the diagnosis of a specific disease is essential in this process. Physicians use (combinations of) ECG features when diagnosing an ECG. Automated ECG interpretation programs incorporated in electrocardiographs operate similarly. However, in clinical practice, manual and automatic diagnosis/interpretation are flawed in several ways: manual ECG diagnosis is subjective and depends on the physician's experience; automated interpretation of the ECG cannot perform at the level of a top cardiologist, and it is still unclear, in several clinical scenarios, which ECG features would help diagnose a specific condition.

Thus, further research on the role of specific ECG features in diagnosing a specific cardiac disease/condition is necessary. Such research is usually done by conventional statistical analysis, evaluating the ECG feature performances in separating cases and controls (subjects with/without altered clinical status or disease).

Here, we try to identify ECG features that may be related to heart-failure (HF) development. Currently, the potential role of the ECG in HF diagnosis is not clear. HF affects about 2% of adults and is associated with a risk of death of about 35% at one year from the first diagnosis [3]. HF is characterized by reduced exercise tolerance and/or fluid retention, when it can be demonstrated that these symptoms are related to a form of cardiac pathology: structural and/or functional abnormalities, including changes in the cardiac electrical properties [4]. Thus, considering that timely HF diagnosis helps to slow down its natural development, research on ECG features to detect emerging HF remains imperative.

Recently, new advanced algorithms were presented to help the research on ECG feature interpretation [5]. These innovative methods try to mimic the clinical diagnosis, applying advanced optimization algorithms that rely on deep learning [5]. Results already presented in the literature proved the usefulness of these tools in terms of performance (high classification score) [6,7], but their feature interpretation was never compared with results provided by conventional statistics.

Thus, this work aims to compare the contributions of ECG features in a deep-learning algorithm for the detection of emerging HF with those obtained by conventional statistical analysis.

2. Materials and Method

2.1. Database

Data consist of 58 10-second 12-lead ECG pairs (baseline and follow-up ECGs) constituting a retrospective observational database of the Leiden University Medical Center. All subjects had a history of myocardial infarction (MI) and were clinically stable during the recording of baseline ECG (routinely performed at least six months after the acute MI). Of these subjects, 33 (controls) remained clinically stable during follow-up ECG recording (one year after the acute MI). The remaining 25 subjects (cases) developed HF; their follow-up ECG was made on the occasion of their first presentation with HF.

The ECGs were processed by LEADS software [8]. Each ECG pair was characterized by 42 features (see Table 1). Among the 42 features, 27 were computed as serial features (*i.e.*, by subtracting baseline ECG feature values from follow-up ECG feature values), while 15 features were computed in the follow-up ECG only.

2.2. Deep Learning Analysis

Data was divided into training set (70%) and validation set (30%), maintaining the case/control prevalence in both sets. A deep-learning neural network (NN) with 42 inputs and case/control outputs was obtained by Repeated Structuring & Learning Procedure (RS&LP) [9]. The NN was created with neurons having sigmoid activation functions and coefficients (weights and biases) that ranged between -1 and +1. The scaled-conjugate-gradients algorithm [10] was used as optimization algorithm. Classes were balanced according to the inverse of their prevalence to compensate for case-control disproportion [11]. The NN was automatically constructed during training by using the training set. The RS&LP algorithm alternates phases of structuring, adding, and initializing neurons, and phases of training, evaluating the increment of the classification task. A validation-based early stopping criterion [12] was applied to avoid overfitting, using the validation set.

To interpret the feature contributions to classification, the local-interpretable model-agnostic explanatory (LIME) algorithm [13-15] was applied to the learned NN. LIME is an explainer algorithm that interprets NN predictions by combining features and coefficients of the trained NN. It locally approximates the NN with an interpretable model, ranking features according to their impact on classification. Thus, for each patient, a feature ranking was constructed by analysing the coefficients of the trained NN. Finally, feature relevance (FR) was obtained as the weighted average (by ranking) of the percentage of patients presenting a specific feature in each of the ranking positions. Thus, 42 FRs were obtained, reflecting the relevance of each of the ECG features listed in Table 1.

Table 1. Feature list and description.

		Feature Description
SERIAL FEATURES	F1	QRS-duration difference
	F2	Modulus of QRS-duration difference
	F3	Difference in maximal QRS-vector magnitude
	F4	Modulus of difference in maximal QRS-vector magnitude
	F5	QRS-integral vector magnitude difference
	F6	Modulus of QRS-integral vector magnitude difference
	F7	QRS-complexity difference
	F8	Modulus of QRS-complexity difference
	F9	Magnitude of J-vector difference
	F10	Difference in maximal T-vector magnitude
	F11	Modulus of difference in maximal T-vector magnitude
	F12	T-integral vector magnitude difference
	F13	Modulus of T-integral vector magnitude difference
	F14	T-wave complexity difference
	F15	Modulus of T-wave complexity difference
	F16	T-wave symmetry difference
	F17	Modulus of T-wave symmetry difference
	F18	Difference in the number of leads with positive T waves
	F19	Number of leads with a T-wave polarity change
	F20	QT-duration difference
	F21	Modulus of QT-duration difference
	F22	Magnitude of the ventricular-gradient difference vector
	F23	QRS-T spatial-angle difference
	F24	Modulus of QRS-T spatial-angle difference
	F25	Heart-rate difference
	F26	Modulus of heart-rate difference
	F27	Difference in ECG-derived ventricular gradient optimized for right ventricular pressure overload
FOLLOW-UP FEATURES	F28	QRS duration
	F29	Maximal QRS-vector magnitude
	F30	QRS-integral vector magnitude
	F31	QRS complexity
	F32	Magnitude of J-vector
	F33	Maximal T-vector magnitude
	F34	T-integral vector magnitude
	F35	T-wave complexity
	F36	T-wave symmetry
	F37	Number of leads with positive T waves
	F38	QT duration
	F39	Magnitude of the ventricular gradient
	F40	QRS-T spatial-angle
	F41	Heart rate
	F42	ECG-derived ventricular gradient optimized for right ventricular pressure overload

2.2. Statistical Analysis

Conventional univariate statistical analysis was performed for each feature by computing the area under the curve (AUC) of the receiver operating characteristic (ROC). Thus, 42 AUCs were obtained.

2.2. Deep Leaning vs. Statistical Analysis

The trained NN was characterized in terms of architecture, and its performance was quantified by ROC analysis, computing the AUC and its 95% confidence intervals (95% CI). The agreement between deep-learning analysis and statistical analysis was evaluated by Pearson's correlation coefficient (ρ) and linear regression analysis of FR on AUC.

3. Results

The trained NN had a [13,7,6] architecture and an AUC of 99% (95% CI: 98%-100%). Values of FR and AUC are reported in Figure 1. FR ranged from 4.47% (F31; QRS complexity of the follow-up ECG) to 0.32% (F3; difference in maximal QRS-vector magnitude between the follow-up and baseline ECGs), while AUC ranged from 82% (F24; modulus of QRS-T spatial-angle difference between follow-up and baseline ECGs) to 23% (F37; number of leads with positive T waves in the follow-up ECG). Agreement between FR and AUC was poor (ρ=0.18; *P-value*=0.25; FR=0.02·AUC+1.46; Figure 2).

4. Discussion

The aim of the paper was to evaluate the contribution of ECG features to HF diagnosis, and to compare the analyses performed by deep-learning interpretation and by conventional statistical analysis.

To be reliable, the LIME algorithm should be applied to NN providing very high classification performance. Thus, in this work, LIME was applied to the trained NN, considering the subjects used to construct the NN (AUC=99%). These subjects have been used by the RS&LP to optimize the NN architecture and performance and, thus, can constitute the perfect dataset to interpret the reasoning performed by the trained NN. However, in deep learning, a high training performance may be a symptom of a poor generalization property of the NN. For this reason, RS&LP was used to create the NN since this constructive procedure proved reliable in preserving the generalization property of the trained NN thanks to its construction rules [9,16,17].

Despite their common statistical background, conventional statistics and deep learning present differences. Firstly, statistical approach is based on linear methods to discriminate cases and controls.

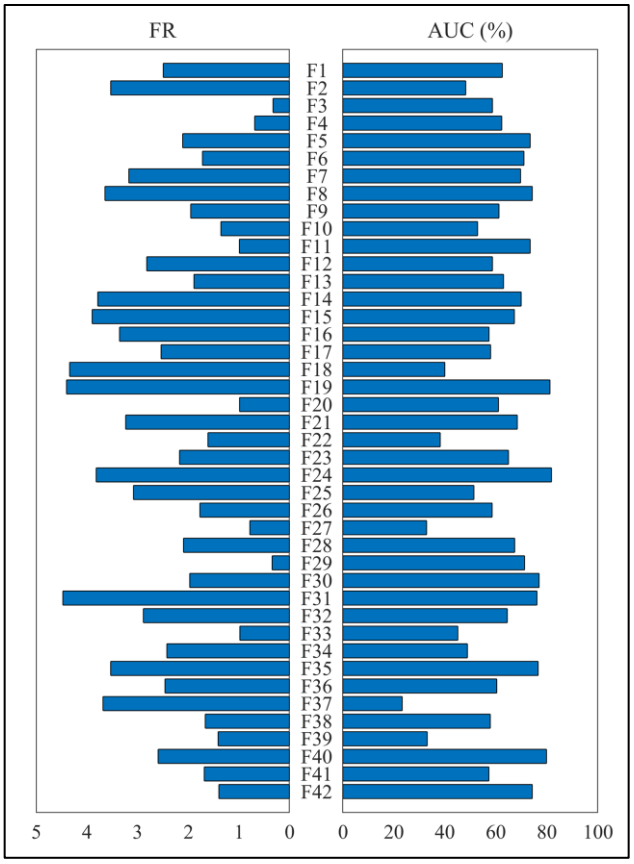

Figure 1. Values of FR and AUC for all features (from F1 to F42).

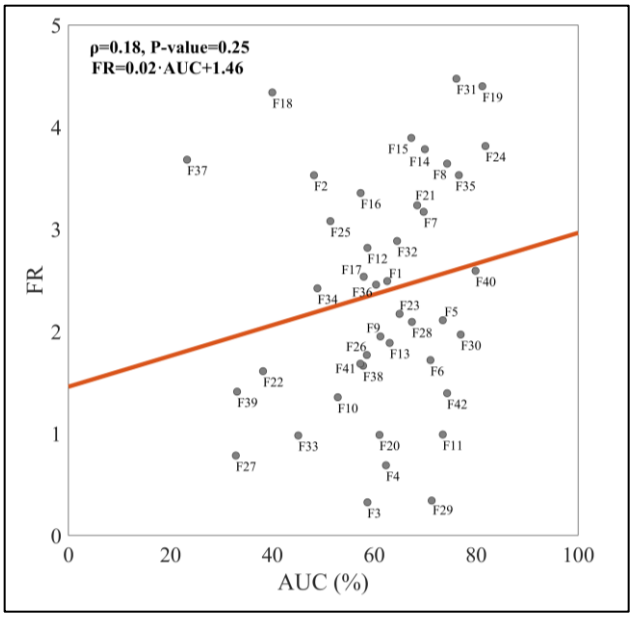

Figure 2. Scatter plot of AUC and FR obtained by conventional statistical analysis and deep learning, respectively. The regression line is depicted in orange.

Differently, deep learning applies innovative nonlinear methodologies, optimizing the shape of hyperplanes with the aim of improving classification performance. Moreover, conventional statistical analysis is based on a univariate statistical approach, evaluating only the discriminant power of each variable without considering possible feature interactions. On the other hand, all features participate in the training of a NN, irrespective of the associations between features.

Results obtained by the deep-learning algorithm are not in agreement with those obtained by conventional statistical analysis (ρ=0.18; *P-value*=0.25). The most prominent features (F19 and F31 with the NN, F9 and F24 with univariate AUC) all make sense, however. Features F19 (number of leads with a T-wave polarity change) and F24 (modulus of the QRS-T spatial-angle difference) are both serial features and can be interpreted as a decrease of concordance or an increase in discordance of the ECG. This indicates that the relation between the depolarization and repolarization processes in the heart deteriorates, a clear trend towards electrical dysfunctioning. In addition, the role of feature F31 (QRS complexity in the follow-up ECG) suggests the presence of a deteriorated depolarization process in HF patients (increased QRS complexity signals QRS fragmentation).

5. Conclusions

Our study is an initial step in identifying important ECG features to HF diagnosis; it suggests that serial ECG comparison may be helpful in HF diagnosis because both statistical and NN approaches identified a change in discordance as a potentially ominous sign for HF development. Additionally, the NN approach suggests that high QRS complexity might be indicative of HF.

Identifying diagnostic features by means of a deep-learning model helps to counterweigh the black-box character of artificial intelligence. Further studies will investigate which of the two approaches superiorly reflects the clinical diagnosis, which remains the gold standard.

References

[1] E. Merdjanovska, A. Rashkovska, "Comprehensive survey of computational ECG analysis: databases, methods and applications," *Expert Syst. Appl*., vol. 203, no. 117206, Oct. 2022.

[2] Y. Sattar, L. Chhabra, "Electrocardiogram," In: StatPearls Treasure Island (FL): StatPearls Publishing, Jan. 2022.

[3] M. Metra, J. R. Teerlink, "Heart failure," *Lancet*, vol. 390, no. 10106, pp. 1981–1995, Oct. 2017.

[4] P. Ponikowski, et al., "2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) Developed with the special contribution of the Heart Failure Association (HFA) of the ESC," *Eur. J. Heart Fail.*, vol. 18, no. 8, pp. 891-975, Aug. 2016.

[5] J. W. Hughes, J. E. Olgin, R. Avram, S. A. Abreau, T. Sittler, K. Radia, H. Hsia, T. Walters, B. Lee, J. E. Gonzalez, G. H. Tison, "Performance of a convolutional neural network and explainability technique for 12-lead electrocardiogram interpretation," *JAMA Cardiol.*, vol. 6, no. 11, pp. 1285-1295, Nov. 2021.

[6] M. Bodini, M. W. Rivolta, R. Sassi, "Interpretability analysis of machine learning algorithms in the detection of ST-elevation myocardial infarction," *2020 Comput Cardiol. Conf.*, vol. 47, no. 9344165, pp. 1-4, Sep. 2020.

[7] W. Liu, F. Wang, Q. Huang, S. Chang, H. Wang, J. He, "MFB-CBRNN: A hybrid network for MI detection using 12-lead ECGs," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 2, pp. 503-514, Feb. 2020.

[8] H. H. M. Draisma, C. A. Swenne, H. Van De Vooren, A. C. Maan, B. H. Van Huysduynen, E. E. Van Der Wall, M. J. Schalij, "LEADS: An interactive research oriented ECG/VCG analysis system," *2005 Comput. Cardiol. Conf.*, vol. 32, no. 1588151, pp. 515-518, Sep. 2005.

[9] A. Sbrollini, M. C. De Jongh, C. C. Ter Haar, R. W. Treskes, S. Man, L. Burattini, C. A. Swenne, "Serial electrocardiography to detect newly emerging or aggravating cardiac pathology: A deep-learning approach," *Biomed. Eng. Online*, vol.18, no. 15, pp. 1-17, Feb. 2019.

[10] M. F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Networks*, vol. 6, no. 4, pp. 525-533, 1993.

[11] G. King, L. Zeng, "Logistic regression in rare events data," *J. Stat. Softw.*, vol. 8, no. 2, pp. 137-163, Jan 2003.

[12] L. Prechelt, "Early stopping — but when?," *In: Montavon G, Orr GB, Müller K-R, editors. Neural Networks Tricks Trade Lect. Notes Comput. Sci.,* second edition, Berlin: Springer, pp. 53–67, 2012.

[13] M. T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," *22nd ACM SIGKDD Int. Conf.*, pp. 1135–1144, Aug. 2016.

[14] S. Grzegorz, A. Naoki, C. L. Aurélie, "Grouped orthogonal matching pursuit for variable selection and prediction," *Adv. Neural Inf. Process. Syst.*, pp. 1150-1158, 2009.

[15] A. C. Lozano, S. Grzegorz, A. Naoki, "Group orthogonal matching pursuit for logistic regression," *JMLR*, vol. 15, pp. 452–460, Apr. 2011.

[16] D. Marinucci, A. Sbrollini, I. Marcantoni, M. Morettini, C. A. Swenne, L. Burattini, "Artificial neural network for atrial fibrillation identification in portable devices," *Sensors*, vol. 20, no. 12, pp. 1-16, Jun. 2020.

[17] A. Sbrollini, M. C. De Jongh, C. C, Ter Haar, R. W. Treskes, S. Man, L. Burattini, C. A. Swenne, "Serial ECG analysis: Absolute rather than signed changes in the spatial QRS-T angle should be used to detect emerging cardiac pathology," *2018 Comput. Cardiol. Conf.*, vol. 45, no. 8744016, pp. 1-4, Sep. 2018.

Address for correspondence:

Laura Burattini
Department of Information Engineering,
Università Politecnica delle Marche,
via Brecce Bianche 12,
60131, Ancona, Italy.
E-mail address. l.burattini@univpm.it