Biomedical Signal Processing and Control 55 (2020) 101597

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

A convolutional neural network approach to detect congestive heart failure

Mihaela Porumb[a], Ernesto Iadanza[b], Sebastiano Massaro[c,d], Leandro Pecchia[a,*]

[a] University of Warwick, School of Engineering, Coventry CV4 7AL, UK
[b] University of Florence, v. S. Marta, 3, Florence, Italy
[c] The Organizational Neuroscience Laboratory, London WC1N 3AX, UK
[d] University of Surrey, Guildford GU2 7XH, UK

ARTICLE INFO

Article history:
Received 28 March 2019
Received in revised form 23 May 2019
Accepted 7 June 2019

Keywords:
Convolutional neural networks
Congestive heart failure
Machine learning

ABSTRACT

Congestive Heart Failure (CHF) is a severe pathophysiological condition associated with high prevalence, high mortality rates, and sustained healthcare costs, therefore demanding efficient methods for its detection. Despite recent research has provided methods focused on advanced signal processing and machine learning, the potential of applying Convolutional Neural Network (CNN) approaches to the automatic detection of CHF has been largely overlooked thus far. This study addresses this important gap by presenting a CNN model that accurately identifies CHF on the basis of one raw electrocardiogram (ECG) heartbeat only, also juxtaposing existing methods typically grounded on Heart Rate Variability. We trained and tested the model on publicly available ECG datasets, comprising a total of 490,505 heartbeats, to achieve 100% CHF detection accuracy. Importantly, the model also identifies those heartbeat sequences and ECG's morphological characteristics which are class-discriminative and thus prominent for CHF detection. Overall, our contribution substantially advances the current methodology for detecting CHF and caters to clinical practitioners' needs by providing an accurate and fully transparent tool to support decisions concerning CHF detection.

1. Introduction

Congestive Heart Failure (CHF) is a pathophysiological condition responsible for the failure of the heart in pumping blood in the body [1] which has encountered widespread research and societal attention [2]. According to the European Society of Cardiology, around 26 million people worldwide are affected by a form of heart failure [3]. CHF is a strongly degenerative condition, and its prevalence increases quickly with age [4,5]. The mortality rate is closely associated with the degree of severity, reaching peaks of 40% in the most serious events [e.g., New York Heart Association (NYHA) classes III-IV] [6]. CHF is also one of the foremost reasons for hospitalization in the elderly, and it is characterized by a resilient relapse rate, with half of the outpatients readmitted within a few months from hospital discharge [7]. Moreover, just among the most industrialized countries, the healthcare expenditure for CHF consumes 2–3% of the healthcare budgets, with the cost of hospitalization being the greatest proportion of the spending [8–10]. Thus, with a worldwide aging population and sustained pressures on healthcare systems and resources, there is the compelling demand—among patients, healthcare providers, policymakers, and the society as a whole—to address this scenario by identifying highly accurate methods to improve detection of heart failures [11] and in turn enable early and more efficient diagnoses.

Recently, research has made significant progress in these areas. In particular, given the quantity and complexity of data involved, machine learning techniques and classsifiers (e.g., SVM, MLP, k-NN, CART, Random Forest) [12–18] have been successfully applied to analyze, detect, and classify heart failures, as we discussed and showed in previous works [12,13,19–21]. These approaches able to distinguish between heart failure and healthy subjects are mostly based on Heart Rate Variability (HRV)—the variation over time of the period between consecutive heartbeats extracted from electrocardiographic (ECG) signals [22]—, showing that depressed HRV patterns represent accurate markers for detecting the condition. However, building accurate HRV-based models is time-consuming and prone to error steps, due to the preprocessing and the iterative process of manually selecting appropriate features. Moreover, the

* Corresponding author.
E-mail addresses: Ernesto.Iadanza@unifi.it (E. Iadanza), sebastiano.massaro@theonelab.org (S. Massaro), L.Pecchia@warwick.ac.uk (L. Pecchia).

https://doi.org/10.1016/j.bspc.2019.101597
1746-8094/

best performing HRV-based models generally require either long-term signals (i.e., 24 h) or at least the combination of short-term HRV with non-standard long-term HRV features, as shown in [23].

To tackle these issues, we present a novel framework of CHF detection that does not rely on HRV features, rather it uses raw ECG signals only. This method is based on a 1-D Convolutional Neural Network (CNN). CNNs are hierarchical neural networks that mimic the human visual system and have proven to be effective in recognizing patterns and structures of input data in image classification, localization, and detection tasks, among others [24–26]. Moreover, they have been extensively used for time series analysis in classification tasks. Some successful examples include, among others: general time series classification [27–29], speech recognition tasks [30], arrhythmia detection [31,32], and multivariate diagnostic measurements modeling [33,34].

Inspired by this growing body of research, we aim to detect CHF through a 1-D CNN approach on the ECG signals. Adding to the significant benefit of building upon raw physiological data, this method enables visualization of the input time series subsequences that are class discriminative (i.e., CHF vs. healthy subjects). This feature considerably improves the interpretability of the CNN model and represents a crucial aspect to ensure the 'transparency' of the method [35,36]. Such a transparency is fundamental to help researchers explain how conclusions are reached, and to aid professionals in better understanding correlations of pathophysiological behaviors and properties revealed by the model.

As we shall explain, we used a form of class activation mapping (Grad-CAM) [37] that highlights the class discriminative regions in the input data. In other words, Grad-CAM uses the model's last activation maps to originate a heat map that can be overlapped with the input, thus showing what regions in the input data contribute most to CHF detection. Overall, the proposed framework puts forward several developments for CHF detection, such as refraining from using hefty preprocessing and features selection steps of HRV-based models, as well as enabling visualization of the subsequences in the input time series which are used to reach certain clinically-relevant conclusions.

2. Methods

2.1. Data

We performed a retrospective analysis on two publicly available datasets. The data for the normal subjects (i.e., control group) were retrieved from the MIT-BIH Normal Sinus Rhythm Database [38] included in PhysioNet [39]. This dataset includes 18 long-term ECG recordings of normal healthy not-arrhythmic subjects (Females = 13; age range: 20 to 50). The data for the CHF group were retrieved from the BIDMC Congestive Heart Failure Database [40] from PhysioNet [39]. This dataset includes long-term ECG recordings of 15 subjects with severe CHF (i.e., NYHA classes III-IV) (Females = 4; age range: 22 to 63). Altogether, the pool of data available for this study consists of about 20 h of ECG recordings per each subject; it contains two-channel ECG signals sampled at 250 samples per second for the BIDMC dataset, and 128 samples per second for the MIT-BIH dataset.

The two datasets used in this study were published by the same laboratory, the Beth Israel Deaconess Medical Center, and were digitized using the same procedures according to the signal specification line in the header file. These datasets have been widely used in several studies concerning CHF detection, as explained in [11]. Indeed, it is an acknowledged practice to build and validate CHF classification models on these sources to ensure opportunities for reproducibility.

Table 1
Total number of extracted beats from both datasets: MIT-BIH and BIDMC.

Control group beats	CHF group beats
275,974	214,531

2.2. ECG preprocessing

The ECG recordings from the BIDMC dataset were down-sampled in order to match the sampling frequency of the ECG signals from the MIT-BIH dataset (i.e., 128 Hz). Records from both databases were already made available with computed beat annotations, which we used to isolate and extract individual heartbeats. We considered a window of 235 ms before the annotated R peak (equivalent to 30 samples = 128 samples/s*0.235 s), and a window of 390 ms after the R peak (equivalent to 50 samples). Only the beats that were annotated as normal (N) were retained for further analysis.

Each heartbeat was z-normalized to obtain an approximately zero mean value (i.e., amplitude) and a standard deviation close to 1. Because the number of heartbeats extracted for each subject was very large (~70,000 beats), and because temporally close beats hold similar patterns, we randomly selected one single beat every 5 s of the ECG signal. The total number of beats used in this work is shown in Table 1.

2.3. Beats classification

Each heartbeat was labeled with a binary value of 1 or 0 (hereafter "class" not to be confused with the NYHA classes) according to the status of the subject: healthy or suffering from CHF, respectively. As customary in machine learning, the dataset was randomly split into three smaller subsets for training, validation, and testing (corresponding to 50%, 25%, and 25% of the total data, respectively). Because one person's heartbeats are highly intercorrelated, these were included only in one of the subsets (i.e., training, testing, or validation set) at a time. In order to reduce the variance of the classification results we repeated the random splitting process 10 times. In this way, the CNN was trained and evaluated over 10 runs; the reported results represent the average performance on these runs, unless otherwise stated. Performing 10 different random splits of the dataset, and then averaging the results of the models, is a similar procedure to that of cross-validation (i.e., averaging the variance and bias of the estimator). The sole difference between the two methods is that the additional validation dataset is used for early stopping of the training process and tuning the hyperparameters of the CNN.

Two different performance evaluation strategies were employed: a) individual beat classification, and b) classification on 5 min of ECG excerpts through what we called a "majority voting scheme." Specifically, by denoting the number of individual heartbeats classified as normal with N_1, and the number of heartbeats that were classified as CHF with N_0 within 5-minutes ECG excerpts, the final class associated with a given 5-minutes ECG excerpt was calculated as $\max(N_0, N_1)$.

2.4. Performance measures

We computed the following measures for binary classification to assess the performance of the CNN classifier: Accuracy (Acc), Precision (P), Sensitivity (Sens), Specificity (Spec) and Area under the Curve (AUC) [23]. These performance indicators correspond to the golden standard in the literature [17,41], thereby enabling ongoing models' comparison.

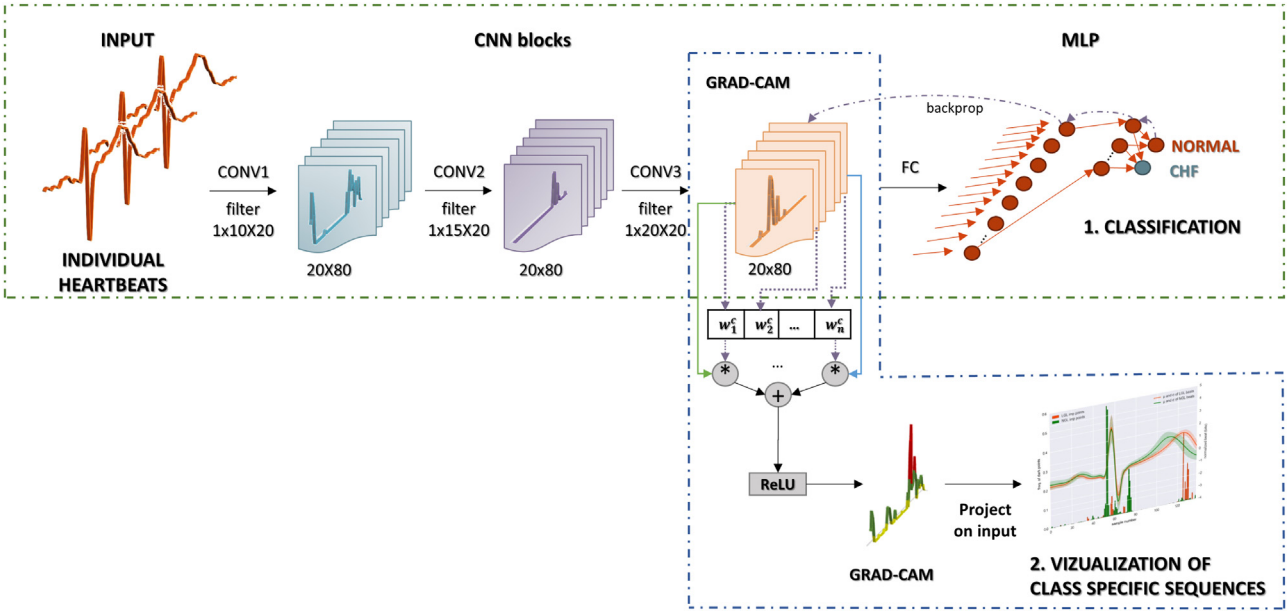

Fig. 1. The 1-D CNN architecture for beat classification and visualization of the class discriminative sequences in the input heartbeats.

2.5. 1-D CNN classifier

The complete architecture of our model is shown in Fig. 1.

A 1-D CNN classifier was used for feature extraction, and the MLP module for the classification of the raw ECG heartbeats. For the beat classification problem, the CNN acts as a feature extractor block. The final activations obtained from the last layer of convolution are used as inputs in a Multilayer Perceptron (MLP) network. The final output is obtained from a softmax layer [42]. The basic convolutional layer is followed by a batch normalization layer [43] and a ReLU activation function [44]. The convolution operation is performed in each layer by 20 1-D filters (i.e., kernels) with sizes {10, 15, and 20} using a stride of 1, hyperparameters, as in previous studies [28,32]. Thus, the convolutional block can be formalized as a set of three operations—convolution, batch normalization, and non-linear activation—defined by the equations below:

$$y = W \circledast x + b \tag{1}$$

$$bn = BN(y) \tag{2}$$

$$act = ReLU(bn) \tag{3}$$

where $\circledast$ is the convolution operator, W are the convolutional layer parameters (i.e., the learnable filters), x represents the input time series, b the bias, BN the batch normalization function, and $ReLU$ the activation function. The resulting CNN is built by stacking three convolutional blocks, each comprising 20 filters of sizes {10, 15, and 20}.

The final activations are flattened and passed to an MLP with a hidden layer of 30 neurons that are then fully connected to the output neurons. The final label is obtained from the softmax function. Pooling operation is excluded from the convolutional blocks according to recent research showing that it does not affect the classifier performance and might affect overfitting (see, e.g., ResNet [45]).

2.6. Visualization of class specific sequences

We employed a Grad-CAM technique to visualize the subsequences in the input ECG beats discern a certain class [37]. Grad-CAM represents a generalization of CAM [46] applicable to a broad range of CNNs, including those containing fully-connected layers. In Grad-CAM the weights used in the linear combination of the forward activation maps are computed as the global average pooling of the gradients of the score for class c with respect to the last feature maps. Specifically, the method implies the computation of the gradient of the score for class c (y^c) with respect to feature maps A (i.e., in our case the feature maps of the last conv layer) that are global-average-pooled to obtain the weights α_k^c. Therefore, the weights α_k^c are computed as follows:

$$\alpha_k^c = \frac{1}{m}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{3}$$

where α_k^c capture the importance of feature map k for a target class c. The Grad-CAM heatmap for a class c, is obtained as a weighted combination of feature maps, followed by ReLU [37].

$$Grad\_CAM\_map_c = ReLU\left(\sum_k \alpha_k^c A^k\right) \tag{4}$$

As shown in Fig. 1, we used this tool to derive the class activation map for class c, which indicates the importance of the activation at a temporal location x_i; this leads to the classification of the input time series as a class c.

2.7. Experimental settings

We used AdamOptimizer [47] as the optimization function with an initial learning rate of 10^{-3}. The batch size was set to 200. The maximum number of training steps was set to 3000 which represents approximately 2.5 epochs when considering the batch size above and the approximate training dataset size of 250,000 beats. We used the validation set to monitor the training process and account for the occurrence of overfitting. In such a case, an early stopping criterion was implemented: This criterion does stop the training when the AUC on the validation dataset does not improve over a k = 30 optimization steps. We used the categorical cross-entropy as loss function. The weights were initialized using the Xavier initializer [48]; the biases were initialized to zero. The network was implemented in Python using TensorFlow framework [49].

Due to the high number of hyper-parameters in the CNN, we assessed a combination of architecture and hyper-parameters

Table 2
Individual Heartbeats Classification.

	Training	Validation	Test
Accuracy	$0.999 \pm 1.1^{*}10^{-16}$	$0.982 \pm 2.0^{*}10^{-3}$	$0.978 \pm 2.0^{*}10^{-3}$
AUC	$0.999 \pm 1.1^{*}10^{-16}$	$0.990 \pm 1.1^{*}10^{-16}$	$0.980 \pm 1.1^{*}10^{-16}$
Sensitivity	$0.999 \pm 1.1^{*}10^{-16}$	$0.973 \pm 2.0^{*}10^{-3}$	$0.963 \pm 4.0^{*}10^{-3}$
Specificity	$0.999 \pm 1.1^{*}10^{-16}$	$0.978 \pm 2.0^{*}10^{-3}$	$0.986 \pm 7.0^{*}10^{-4}$
Precision	$0.999 \pm 1.1^{*}10^{-16}$	$0.970 \pm 3.0^{*}10^{-3}$	$0.985 \pm 7.0^{*}10^{-4}$

Table 3
Mean Confusion Matrix.

	True CHF	True normal	Total ACC
Classified CHF	67181	694	
Classified normal	2071	61512	
Classified correctly from each class	~97%	~98.8%	~97.9%
Error	~3%	~1.1%	~2.1%

through an iterative process, using grid-search and manual tuning. As concerns the architecture structure, we searched over the number of convolutional layers (3–9), different filter sizes (from 5 to 20), and the number of filters in each convolutional layer (5 to maximum 30). The learning rate was instead manually tuned to achieve faster convergence; the considered values were $\{10^{-1}$ to $10^{-5}\}$. The results we present below were obtained with the CNN architecture that achieved the highest performance on the validation dataset with the minimum number of parameters.

3. Results

3.1. Individual beats classification

Table 2 reports the performance for individual beat classification as the average of the 10 different runs for the 10 random splits of the input data for training, validation, and testing.

Table 3 shows a confusion matrix for the mean of the 10 runs, without considering the voting strategy, corresponding to the testing dataset. These findings show that on average the number of false positives is around 1% of the mean number of normal heartbeats, while the false negatives are around 3% of the total number of heartbeats in CHF class.

A closer analysis of one of the runs revealed that out of the total of 854 misclassified CHF beats, 614 belonged to the same subject (id: chf01). The same applied to the control group, where 478 beats that were misclassified belonged to the same subject (id:16272). Nonetheless, the number of misclassified beats for these two subjects was negligible, representing around 2%–4% of the total number of beats that were extracted for the two subjects. In other words, subjects 16,272 and chf01 could still be correctly categorized as healthy and CHF subjects respectively, because more than 95% of their heartbeats were appropriately classified. The 5 consecutive heartbeats were accepted to be very similar, thus averaging them should not result in significant changes in the heartbeat morphology. We confirmed this occurrence by performing an additional simulation in which we averaged the heartbeats in 5 s interval. As expected, we observed no change in the performance of the classifier.

3.2. Classification using majority voting in a specific timeframe

We employed a "majority voting scheme" for the heartbeats within a 5-minutes ECG timeframe to enable direct comparison with previous studies using short-term HRV features computed on 5-minutes ECG segments (see Section 2.3 for details). In sum, the class associated with a 5-minutes ECG segment was the majority class of the individual heartbeats in that segment. Using such a voting strategy, we could evaluate the accuracy of our method in discriminating between healthy and CHF patients when using 5-minutes ECG recordings. All the performance measures considered (i.e., ACC, Sens, Spec, and AUC) improved by 99% when employing the 5-minutes voting in comparison to the individual heartbeat predictions. Moreover, the mean number of the misclassified 5-minutes ECG excerpts using the voting scheme on the 10 runs was 23 out of a total of 2,227, which corresponds to about 0.1%.

These results suggest that, differently from existing evidence [23], a highly accurate model for CHF diagnosis can be built by using relatively short (~5-minutes) raw ECG recordings. To further corroborate this argument, we investigated subjects and specific timeframes where any misclassified 5-minutes ECG segment occurred, as shown in Table 4. As a matter of illustration, here we present a randomly selected run. We can observe that out of the total number of 16 subjects included in the test and validation datasets, 5 of them present misclassified 5-minutes ECG segments.

Fig. 2 presents a closer inspection of these results. It illustrates the heartbeat classification on one of the 10 models we used on the training, validation, and testing datasets including all the subjects of this study. Once again, we can readily appreciate that the extracted heartbeats were correctly classified (i.e., the misclassified heartbeats for some subjects are considerably less than the correctly classified ones). Moreover, the misclassified beats show a consecutive trend, which is most likely indicating a noisy ECG segment in the raw recording. It might also be possible that these specific heartbeats were intrinsically not-discriminative, an issue that could be tackled by extending the training dataset. We further support these propositions by illustrating the misclassified ECG excerpts in Fig. 3, showing that our model failed to classify either noisy segments (e.g., subjects 16,272, 19,830) or ECG excerpts affected by movement artifacts, as also acknowledged by independent cardiological assessments. This circumstance can easily be confirmed by inspecting the data with the online viewer LightWAVE [50] from Physionet.

3.3. Class specific sequences

Moving to the visualization feature of the model, Fig. 4 represents a summary of key regions in the input beats, obtained through

Table 4
Test Subjects With Misclassified 5-Minutes ECG Segments.

Subject ID	ECG segment start time → end time	Subject ID	ECG segment start time → end time
16795	13:10 → 13:15	17453	13:10 → 13:15
16795	13:30 → 13:40		
16795	13:55 → 14:20	19093	4:35 → 4:40
16795	14:30 → 14:40	19093	21:55 → 22:25
16795	14:55 → 15:00		
16795	15:30 → 15:40	19830	10:40 → 10:45
16795	15:50 → 15:55		
16795	16:05 → 16:15	16272	16:05 → 16:10
16795	16:50 → 16:55	16272	16:15 → 17:05
16795	14:20 → 17:35	16272	17:30 → 17:35

Fig. 2. Heartbeats classification for all the subjects included in the training, validation and test datasets, evaluated on one of the 10 runs. Panels b) and d) present the predictions for the voting in a 5-minutes window of the individual heartbeats. The red and green bars represent the extracted heartbeats in chronological order for CHF patients and healthy subjects, respectively. The grey gaps in the bars are missing data.

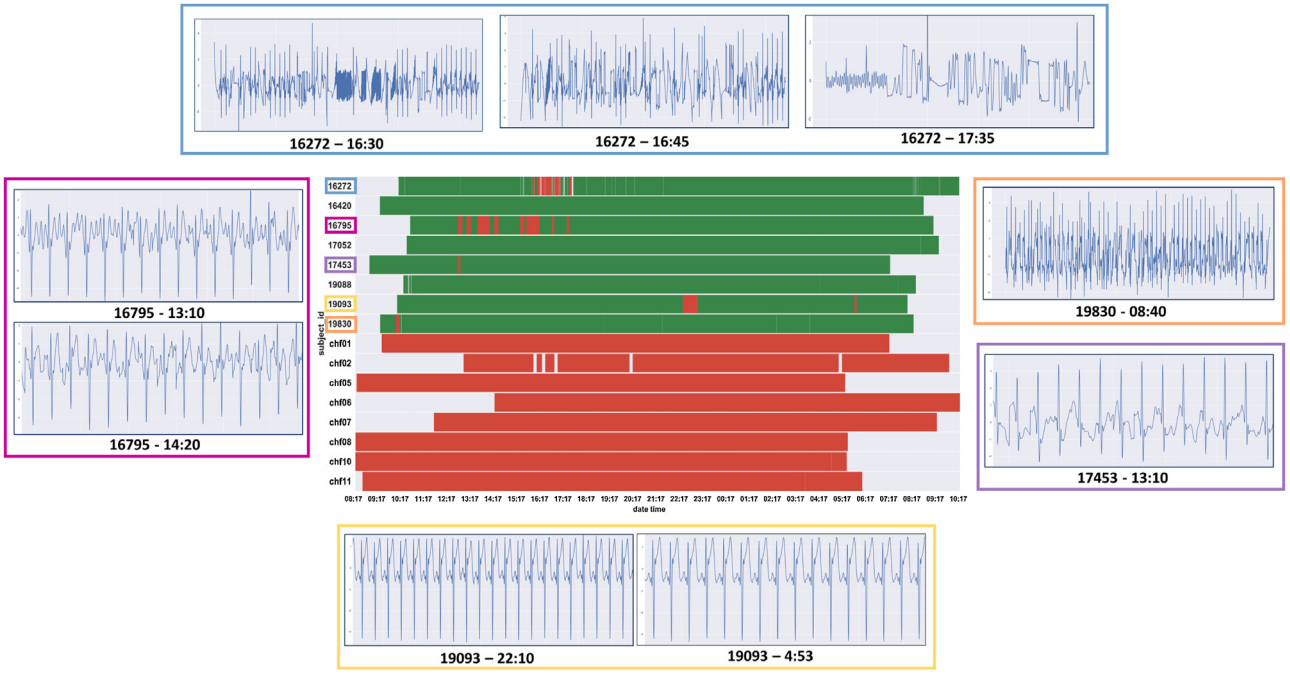

Fig. 3. Example of misclassified ECG segments evaluated on the validation and testing datasets.

Grad-CAM, that are class discriminative. The two ECG beats shown in Fig. 4 represent the average of all the heartbeats from the testing set corresponding to the normal (i.e., green) and CHF (i.e., red) classes with the respective error bands. The histograms feature the data points in the input ECG beats above 0.8 in the normalized heat maps obtained through Grad-CAM. The sample points that were found to be significant (i.e., where the histogram points reached a threshold of 0.25) are presented in Fig. 4. In other words, these are the points that contribute the most in identifying a certain class.

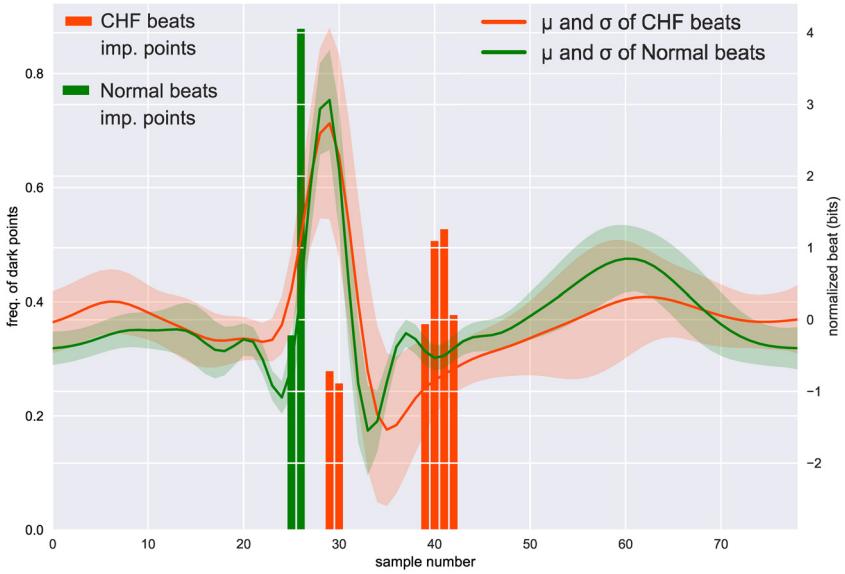

Fig. 4. Histograms obtained from the GRAD-CAM heat map, indicating the most important sample points in the input beats for a certain class.

4. Discussion

We presented a novel and innovative framework for CHF detection based on CNN. While CHF detection using machine learning on ECG has traditionally relied on the extraction of HRV features to build classifiers able to discriminate between healthy and CHF subjects, in this study we substantially advanced this body of knowledge by using as input data raw ECG heartbeats.

The development of such a model allows to bypass HRV features extraction and selection steps, yet, and importantly, without compromising the classification performance of the model. Indeed, as a matter of comparison, the model proposed by Isler et al. [14] based on short-term HRV features extracted from a randomly chosen 5-minutes ECG excerpt out of the total 24 h ECG recordings, reports a classification performance of ∼96% sensitivity and specificity. Thus, not only we improved these classification results, but also we showed that 99% of the 5-minutes ECG excerpts could be used for providing a correct CHF diagnosis. We argue that these benefits are due to the ability of CNNs' to automatically extract and learn patterns in the input data, rather than relying on specialized features that might not fully capture the true characteristics of the underlying physiological signals.

In this this study we purposively chose a shallow architecture (i.e., with 3 convolutional layers), yet using larger filter sizes. Experiments performed with different CNN architectures reveal that similar classification results can also be obtained with a deeper architecture and smaller filter sizes. In our case, the number of neurons in the MLP layer did have great influence on the model's performance and we observed a much faster convergence with batch normalization than without.

Another relevant note concerns the difference between our work and that by Acharya et al. [51] who employed a CNN-based model for CHF detection. The current study puts forward important benefits. First, as regards the evaluation methodology, the authors in [51] used the recordings corresponding to a certain subject for both training and testing the model; here, instead, we considered the extracted heartbeats only in either the training or testing datasets for a run. That is, we divided the subjects into training and testing subsets, rather than using their corresponding ECG heartbeats. Moreover, differently from the architecture proposed in [51], our model is simpler because it contains 3 convolutional layers only and no pooling operations, thus reducing the computational complexity to the benefit of possible future mobile applications. The CNN inputs are different too: In the current study, we extracted individual heartbeats, whereas in [51] ECG excerpts of 2 s were used as inputs. This represents a critical strategy that our model advances because it allows us to investigate the heartbeat morphology in connection to the CHF diagnosis through the Grad-CAM method (see Fig. 4). To the best of our knowledge, this is also the opening evidence in which discriminative segments in the input heartbeat are revealed for the CHF diagnosis.

Finally, mean differences in the heartbeat morphology are clearly shown in Fig. 4 and the histograms highlight the regions in the input heartbeats that were mostly used by the CNN in making the predictions. Thus, our model allows to visualize those features that are automatically learned during the classification process. According to a recent study by Taylor and Hobbs [52], heart failure is associated with abnormal ECG in 89% of the cases. Moreover, James et al. [53] showed that heart failure patients have significant prolonged QRS duration, prolonged QT duration, prolonged QTc and more rightward T wave axis. Also, de Luna et al. [54] reported long QT interval and visible variations of the shape, amplitude or polarity of the T wave on an alternate-beat basis as ECG changes associated with heart failure. In Fig. 4, we presented two heartbeats, obtained as average on all normal and CHF heartbeats included in the available datasets. We can appreciate that some of the morphological changes reported in the literature can be easily identified in Fig. 4. These include, among others, flattened, prolonged T wave, visible changes in the amplitude of the T wave. The highlighted parts in the normal and CHF heartbeats represent the ECG heartbeat subsequence used by the CNN network to generate a prediction. Whilst not entering into clinical interpretations, we can anyway observe that the identified important parts in the ECG are mainly located near the fiducial points (Q, R and T waves).

As shown in Table 5, the classification results suggest that this method is highly competitive with, and in some respect superior to, previous models using HRV features. It is worth noticing that, while the classification process presented here is performed at the heartbeat level, the obtained results can be easily compared with previous studies by performing a per-subject classification, using the presented voting scheme, either for 5-minutes ECG excerpts or on the entire ECG recording (i.e., by quantifying the majority class predicted for all the heartbeats corresponding to each subject).

Table 5
Classification performance of the current study as compared to the existing literature.

Author(s) (Year)	Methods	Data	Results
Asyali et al. (2013) [17]	Linear discriminant analysis Bayesian classifier Based on HRV features	1. Normal Sinus Rhythm RR Interval Database 2. Congestive Heart Failure RR Interval Database 54 normal subjects 29 CHF subjects	Sensitivity: 0.82 Specificity: 0.98 Accuracy: 0.93
Isler et al. (2003) [14]	k-NN Based on HRV features	1. Normal Sinus Rhythm RR Interval Database 2.Congestive Heart Failure RR Interval Database 54 normal subjects 29 CHF subjects	Sensitivity: 0.96 Specificity: 0.96 Accuracy: 0.96
Melillo et al. (2014) [12]	CART Based on HRV features	1.Normal Sinus Rhythm RR Interval Database 2. MIT-BIH Normal Sinus Rhythm Database 3. Congestive Heart Failure RR Interval Database 4. BIDMB Congestive Heart Failure 72 normal subjects 44 CHF subjects	Sensitivity: 0.9 Specificity: 1.0 Accuracy: 0.96
Chen et al. (2015) [18]	Non-equilibrium decision-tree Nased on support vector machine classifier Based on dynamics of 5-minute short-term HRV measures	1.Normal Sinus Rhythm RR Interval Database 2. MIT-BIH Normal Sinus Rhythm Database 3. Congestive Heart Failure RR Interval Database 4. BIDMB Congestive Heart Failure 72 normal subjects 44 CHF subjects	Accuracy: 0.96
Liu et al. (2016) [15]	SVN and k-NN Non-standard HRV features	1. Normal Sinus Rhythm RR Interval Database 2.Congestive Heart Failure RR Interval Database 30 normal subjects 17 CHF subjects	Sensitivity: 1.0 Specificity: 1.0 Accuracy: 1.0
Acharya et al [51]	Raw, 2 seconds ECG CNN	1. MIT-BIH Normal Sinus Rhythm Database 2. BIDMB Congestive Heart Failure 3. Fantasia	Sensitivity: 0.98 Specificity:0.99 Accuracy: 0.98
This study	**Raw ECG heartbeats** **CNN based method**	**1. MIT-BIH Normal Sinus Rhythm Database** **2. BIDMB Congestive Heart Failure** **18 normal subjects** **15 CHF subjects**	**Sensitivity: 1.0** **Specificity:1.0** **Accuracy: 1.0**

In our model, the error in the subjects' classification is nil: Each subject is correctly classified as being either healthy or affected by CHF. Previously, Liu et al. [15] presented what to the best of our knowledge is the only other system achieving a similar accuracy in CHF detection. Yet, our work focuses on a purposively selected sample of subjects rather than on the entire pool available in the original datasets. The misclassified heartbeats are trivial in comparison to the correctly classified ones, indicating 100% accuracy for the classification at the subject level, as evident in Fig. 2.

Added to their superior performance, our results bring forward the intuition that short ECG recordings, of just about 5 min, can be sufficient to correctly diagnose severe CHF. This is an important result because, with the increasing availability of wearable devices capturing interim ECG recordings (e.g., smart-watches), accurate CHF detection and prediction might be soon performed through devices people carry in everyday situations. In this respect, current approaches have looked at HRV, which is conventionally analyzed using 5-min excerpts, if not longer (e.g., nominally 24 h). However, physiological phenomena running over much shorter timespans might not be fully captured when considering such long excerpts. Thus, by using individually extracted heartbeats to provide an accurate diagnostic prediction, the approach proposed here facilitates future implementation into real-time systems towards the realization of a Clinical Decision Support System (CDDS). In CDDS envisioned end-users are not only experienced cardiologists, but also patients, caregivers, general practitioners, nurses, trainees, and so forth. Moreover, once trained, the network responds in a very short time and can be embedded into mobile phones or deployed to cloud architectures, as we already discussed elsewhere [13,55].

Our study must also be seen in light of its limitations, which anyway represent valuable calls for future research avenues. First, the CHF subjects used in this study suffer from severe CHF only (i.e., NYHA classes III-IV); the discrimination could yield less accurate results for milder CHF (i.e., NYHA classes I-II). Second, the use of acknowledged public datasets including raw ECG signals was aimed at improving the opportunity for comparison and transparency of science. Yet, this choice necessarily resulted in a dataset which, at the subject level, is on average smaller than those used in other models. Future studies will need to extend the results presented here in larger datasets.

Finally, we would like to encourage future research as well as clinical practice to apply this framework to the pre-diagnostic assessment of CHF. This strategy will not only allow further model's validation, but it will also enable moving from retrospective detection to prospective diagnosis, potentially improving human well-being, reducing healthcare costs, and hopefully even saving lives.

5. Conclusions

In this article we proposed a novel and highly-effective method for CHF detection based on CNNs. Compared to existent models the added value of our model is that it uses raw ECG signals, rather than HRV features, and yields prominent accuracy in detecting CHF. Indeed, we showed that the performance of the CNN model holds considerably high scores when the classification is performed at the heartbeat level, and it is error-free as pertains the subject classification component. Moreover, this is the first study using advanced machine learning approaches to reveal which morphological characteristics of the ECG beats are the most important to efficiently detect CHF. We believe that this is a crucial contribution for clinical practice because clinicians, who are ultimately responsible for CHF patients, are generally refractory to the adoption of methods that are not fully transparent in showing how certain decisions were reached. Our method successfully responds to this issue.

Finally, we are hopeful that the model provided here will be used as a highly-effective and generalizable classification method for other time series classification tasks in medicine and beyond (e.g.,

in neuro-behavioral research [56]). We thus call for future works to expand the results presented in this study towards improving human wellbeing and reducing current healthcare financial and societal burdens.

Acknowledgement

We would like to thank Dr. Dindeal Alin and Dr. Moldovan Emil for assessing the ECG excerpts that resulted as misclassified by our model.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J.J. McMurray, S. Stewart, Epidemiology, aetiology, and prognosis of heart failure, Heart 83 (May (5)) (2000) 596–602.
[2] D.M. Lloyd-Jones, et al., Lifetime risk for developing congestive heart failure: the Framingham Heart Study, Circulation 106 (December (24)) (2002) 3068–3072.
[3] P. Ponikowski, et al., Heart failure: preventing disease and death worldwide, ESC Heart Fail. 1 (September (1)) (2014) 4–25.
[4] S. Stewart, K. MacIntyre, S. Capewell, J.J.V. McMurray, Heart failure and the aging population: an increasing burden in the 21st century? Heart 89 (January (1)) (2003) 49–53.
[5] M.W. Rich, Epidemiology, pathophysiology, and etiology of congestive heart failure in older adults, J. Am. Geriatr. Soc. 45 (August (8)) (1997) 968–974.
[6] B.M. Massie, N.B. Shah, Evolving trends in the epidemiologic factors of heart failure: rationale for preventive strategies and comprehensive disease management, Am. Heart J. 133 (June (6)) (1997) 703–712.
[7] J.M. Vinson, M.W. Rich, J.C. Sperry, A.S. Shah, T. McNamara, Early readmission of elderly patients with congestive heart failure, J. Am. Geriatr. Soc. 38 (December (12)) (1990) 1290–1295.
[8] J.B. O'Connell, The economic burden of heart failure, Clin. Cardiol. 23 (March (3 Suppl.)) (2000) III6–10.
[9] A.P. Ambrosy, et al., The global health and economic burden of hospitalizations for heart failure: lessons learned from hospitalized heart failure registries, J. Am. Coll. Cardiol. 63 (April (12)) (2014) 1123–1133.
[10] C. Berry, D.R. Murdoch, J.J.V. McMurray, Economics of chronic heart failure, Eur. J. Heart Fail. 3 (June (3)) (2001) 283–291.
[11] E.E. Tripoliti, T.G. Papadopoulos, G.S. Karanasiou, K.K. Naka, D.I. Fotiadis, Heart failure: diagnosis, severity estimation and prediction of adverse events through machine learning techniques, Comput. Struct. Biotechnol. J. 15 (January) (2017) 26–47.
[12] P. Melillo, R. Fusco, M. Sansone, M. Bracale, L. Pecchia, Discrimination power of long-term heart rate variability measures for chronic heart failure detection, Med. Biol. Eng. Comput. 49 (January (1)) (2011) 67–74.
[13] L. Pecchia, P. Melillo, M. Bracale, Remote health monitoring of heart failure with data mining via CART method on HRV features, IEEE Trans. Biomed. Eng. 58 (March (3)) (2011) 800–804.
[14] Y. İşler, M. Kuntalp, Combining classical HRV indices with wavelet entropy measures improves to performance in diagnosing congestive heart failure, Comput. Biol. Med. 37 (October (10)) (2007) 1502–1510.
[15] G. Liu, L. Wang, Q. Wang, G. Zhou, Y. Wang, Q. Jiang, A new approach to detect congestive heart failure using short-term heart rate variability measures, PLoS One 9 (April (4)) (2014), e93399.
[16] Z. Masetic, A. Subasi, Congestive heart failure detection using random forest classifier, Comput. Methods Programs Biomed. 130 (July) (2016) 54–64.
[17] M.H. Asyali, Discrimination power of long-term heart rate variability measures, Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society vol. 1 (2003) 200–203.
[18] W. Chen, L. Zheng, K. Li, Q. Wang, G. Liu, Q. Jiang, A novel and effective method for congestive heart failure detection and quantification using dynamic heart rate variability measurement, PLoS One 11 (November (11)) (2016).
[19] G. Guidi, M.C. Pettenati, R. Miniati, E. Iadanza, Random Forest for automatic assessment of heart failure severity in a telemonitoring scenario, Annual International Conference of the IEEE Engineering in Medicine and Biology Society vol. 2013 (2013) 3230–3233.
[20] G. Guidi, L. Pollonini, C.C. Dacso, E. Iadanza, A multi-layer monitoring system for clinical management of Congestive Heart Failure, BMC Med. Inform. Decis. Mak. 15 (Suppl. 3) (2015) S5.
[21] G. Guidi, M.C. Pettenati, R. Miniati, E. Iadanza, Heart failure analysis dashboard for patient's remote monitoring combining multiple artificial intelligence technologies, Annual International Conference of the IEEE Engineering in Medicine and Biology Society vol. 2012 (2012) 2210–2213.
[22] U. Rajendra Acharya, K. Paul Joseph, N. Kannathal, C.M. Lim, J.S. Suri, Heart rate variability: a review, Med. Biol. Eng. Comput. 44 (December (12)) (2006) 1031–1051.
[23] L. Pecchia, P. Melillo, M. Sansone, M. Bracale, Discrimination power of short-term heart rate variability measures for CHF assessment, IEEE Trans. Inf. Technol. Biomed. 15 (January (1)) (2011) 40–46.
[24] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, USA, in: Proceedings of the 25th International Conference on Neural Information Processing Systems, vol. 1, 2012, pp. 1097–1105.
[25] Stanford University CS231n: Convolutional Neural Networks for Visual Recognition. [Online]. Available: http://cs231n.stanford.edu/.
[26] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (May (7553)) (2015) 436–444.
[27] Y. Zheng, Q. Liu, E. Chen, Y. Ge, J.L. Zhao, Exploiting multi-channels deep convolutional neural networks for multivariate time series classification, Front. Comput. Sci. China 10 (February (1)) (2016) 96–112.
[28] Z. Wang, W. Yan, T. Oates, Time series classification from scratch with deep neural networks: a strong baseline, 2017 International Joint Conference on Neural Networks (IJCNN) (2017) 1578–1585.
[29] Z. Cui, W. Chen, Y. Chen, Multi-Scale Convolutional Neural Networks for Time Series Classification, ArXiv160306995 Cs, Mar., 2016.
[30] A. Graves, A. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 6645–6649.
[31] A.Y. Hannun, et al., Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network, Nat. Med. 25 (January (1)) (2019) 65.
[32] S. Kiranyaz, T. Ince, M. Gabbouj, Real-time patient-specific ECG classification by 1-D convolutional neural networks, IEEE Trans. Biomed. Eng. 63 (March (3)) (2016) 664–675.
[33] Z.C. Lipton, D.C. Kale, C. Elkan, R. Wetzel, Learning to Diagnose With LSTM Recurrent Neural Networks, ArXiv151103677 Cs, Nov., 2015.
[34] M. Fiterau, J. Fries, E. Halilaj, N. Siranart, C. Ré, Similarity-based LSTMs for time series representation learning in the presence of structured covariates, Conference on Neural Information Processing Systems (2016) 7.
[35] F. Doshi-Velez, B. Kim, Towards A Rigorous Science of Interpretable Machine Learning, ArXiv170208608 Cs Stat, Feb., 2017.
[36] S. Chakraborty, et al., Interpretability of deep learning models: a survey of results, in: 2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation, 2017, pp. 1–6.
[37] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626.
[38] The Arrhythmia Laboratory At Boston's Beth Israel Hospital (Now The Beth Israel Deaconess Medical Center), The MIT-BIH Normal Sinus Rhythm Database, 1990 physionet.org.
[39] A.L. Goldberger, et al., PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, Circulation 101 (June (23)) (2000) e215–e220.
[40] D.S. Baim, et al., The BIDMC Congestive Heart Failure Database, 2000 physionet.org.
[41] P. Melillo, E. Pacifici, A. Orrico, E. Iadanza, L. Pecchia, heart rate variability for automatic assessment of congestive heart failure severity, XIII Mediterranean Conference on Medical and Biological Engineering and Computing (2013) 1342–1345.
[42] J.S. Bridle, Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition, Neurocomputing (1990) 227–236.
[43] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, International Conference on Machine Learning (2015) 448–456.
[44] V. Nair, G.E. Hinton, Rectified linear units improve restricted Boltzmann machines, Proceedings of the 27th International Conference on International Conference on Machine Learning (2010) 807–814.
[45] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 770–778.
[46] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 2921–2929.
[47] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, Proceedings of the 3rd International Conference on Learning Representations (2015).
[48] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (2010) 249–256.
[49] M. Abadi, et al., TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, 2015.
[50] LightWAVE 0.64. [Online]. Available: https://www.physionet.org/lightwave/. (Accessed 28 January 2019).
[51] U.R. Acharya, et al., Deep convolutional neural network for the automated diagnosis of congestive heart failure using ECG signals, Appl. Intell 49 (January (1)) (2019) 16–27.

[52] C. Taylor, R. Hobbs, Diagnosing heart failure – experience and 'Best pathways', Eur. Cardiol. Rev. 6 (3) (2010) 10–12.
[53] S. James, et al., 22 role of 12-lead electrocardiography in predicting heart failure in the community, Heart 101 (September (Suppl. 5)) (2015) 46–49.
[54] A.B. de Luna, Basic Electrocardiography: Normal and Abnormal ECG Patterns, 1 edition, Wiley-Blackwell, Malden, MA, 2007.
[55] P. Melillo, A. Orrico, P. Scala, F. Crispino, L. Pecchia, Cloud-based smart health monitoring system for automatic cardiovascular and fall risk assessment in hypertensive patients, J. Med. Syst. 39 (October (10)) (2015) 109.
[56] S. Massaro, L. Pecchia, Heart rate variability (HRV) analysis: a methodology for organizational neuroscience, Organ. Res. Methods 22 (January (1)) (2019) 354–393.