IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 31 July 2023, accepted 11 September 2023, date of publication 20 September 2023, date of current version 3 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3317793

RESEARCH ARTICLE

Comparison of Supervised Learning Algorithms for Quality Assessment of Wearable Electrocardiograms With Paroxysmal Atrial Fibrillation

ÁLVARO HUERTA[1], ARTURO MARTINEZ-RODRIGO[1], DAVIDE CARNEIRO[2], VICENTE BERTOMEU-GONZÁLEZ[3,4], JOSE J. RIETA[5], (Member, IEEE), AND RAÚL ALCARAZ[1]

[1]Research Group in Electronic, Biomedical, and Telecommunication Engineering, University of Castilla–La Mancha, 16071 Cuenca, Spain
[2]INESC TEC, 4200 Porto, Portugal
[3]Hospital Clínica Benidorm, 03501 Alicante, Spain
[4]Departamento de Medicina Clìnica, Universidad Miguel Hernández, 03202 Alicante, Spain
[5]BioMIT.org, Electronic Engineering Department, Universitat Politecnica de Valencia, 46022 Valencia, Spain

Corresponding author: Álvaro Huerta (alvaro.huerta@uclm.es)

This work was supported in part by the Daiichi Sankyo Sociedad Limitada Unipersonal (SLU) and Public Grants of the Spanish Government 10.13039/501100011033 jointly with the European Regional Development Fund (EU) under Grant PID2021-X128525-IV0, Grant PID2021-123804OB-I00, and Grant TED2021-130935B-I00; in part by Junta de Comunidades de Castilla–La Mancha under Grant SBPLY/21/180501/000186; and in part by Generalitat Valenciana under Grant AICO/2021/286.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethical Review Board of Hospital Universitario San Juan de Alicante under Protocol No. UGP-14-219.

ABSTRACT Emerging wearable technology able to monitor electrocardiogram (ECG) continuously for long periods of time without disrupting the patient's daily life represents a great opportunity to improve suboptimal current diagnosis of paroxysmal atrial fibrillation (AF). However, its integration into clinical practice is still limited because the acquired ECG recording is often strongly contaminated by transient noise, thus leading to numerous false alarms of AF and requiring manual interpretation of extensive amounts of ECG data. To improve this situation, automated selection of ECG segments with sufficient quality for precise diagnosis has been widely proposed, and numerous algorithms for such ECG quality assessment can be found. Although most have reported successful performance on ECG signals acquired from healthy subjects, only a recent algorithm based on a well-known pre-trained convolutional neural network (CNN), such as AlexNet, has maintained a similar efficiency in the context of paroxysmal AF. Hence, having in mind the latest major advances in the development of neural networks, the main goal of this work was to compare the most recent pre-trained CNN models in terms of classification performance between high- and low-quality ECG excerpts and computational time. In global values, all reported a similar classification performance, which was significantly superior than the one provided by previous methods based on combining hand-crafted ECG features with conventional machine learning classifiers. Nonetheless, shallow networks (such as AlexNet) trended to detect better high-quality ECG excerpts and deep CNN models to identify better noisy ECG segments. The networks with a moderate depth of about 20 layers presented the best balanced performance on both groups of ECG excerpts. Indeed, GoogLeNet (with a depth of 22 layers) obtained very close values of sensitivity and specificity about 87%. It also maintained a misclassification rate of AF episodes similar to AlexNet and an acceptable computation time, thus constituting the best alternative for quality assessment of wearable, long-term ECG recordings acquired from patients with paroxysmal AF.

The associate editor coordinating the review of this manuscript and approving it for publication was Long Xu.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

INDEX TERMS Atrial fibrillation, signal quality assessment, long-term ECG monitoring, deep learning, machine learning.

I. INTRODUCTION

Atrial fibrillation (AF) is a common cardiac rhythm disorder, which nowadays presents epidemic proportions by affecting more than 38 million people in the developed world [1]. This arrhythmia often leads the patient to reduced exercise capacity and quality of life, as well as to increased risk for hospitalization and depression [2]. Moreover, although AF is not necessarily lethal, it is also directly associated with increased risk for dementia, chronic disease, heart failure, stroke, and death [2]. The most severe outcome related to AF is often ischemic stroke [3]. This cardiovascular event is highly preventable with anticoagulant treatment, but the arrhythmia has to be timely detected. Unfortunately, paroxysmal nature and absent of symptoms for most patients in the initial stage of AF make its detection difficult, to the point that the number of undiagnosed cases are worryingly increasing in the last years [4]. This, along with the facts that about 20% of subjects suffering from ischemic stroke are firstly diagnosed with AF at the time of the cardiovascular event, and that a stroke is the first manifestation of AF in more than 5% of patients [5], make early detection of the arrhythmia a challenge of utmost priority and urgency.

The massive emergence of wearable devices able to monitor electrocardiogram (ECG) continuously for long weeks and even months represents a great opportunity to improve the current suboptimal diagnosis of paroxysmal AF [6], [7]. To this respect, previous works have proven that continuous cardiac monitoring beyond the common 24–48 hours covered by conventional ambulatory technology increases the rate of AF detection by between 5% and 30% depending on the type and duration of the follow-up [8]. However, most wearable devices capture heart activity during the patient's daily life and the ECG signal is often contaminated strongly with ever-changing artifacts and transient noises [9], [10]. Hence, thoughtful manual interpretation of large amounts of ECG data is still needed to avoid misdiagnosis of AF and other cardiac events. This is highly time-consuming and requires significant medical staff resources that are nowadays unavailable in many healthcare systems, thus delaying the integration of the wearable technology into clinical cardiovascular medical practice [6].

A practical way to palliate this issue is to assist medical workers with automated selection of only ECG segments with sufficient quality for precise interpretation. Indeed, automated ECG quality assessment has received growing attention in the last years, and a broad variety of algorithms have been proposed for that purpose [9], [10]. Most of these methods use common machine learning (ML) classifiers, e.g., decision tree, random forest, support vector machine (SVM), etc., to combine features manually derived from the raw or preprocessed ECG signal and from its delineated intervals and waves [9], [10]. They have reported successful performance on resting, short recordings, where ECG morphology and fiducial points can be clearly identified. However, their efficiency is significantly worse on long-term, wearable ECG signals, because these often present altered waveforms, severe artifacts, and dynamic external noise [9], [10]. To reduce the dependency on ECG morphology as well as the subjectivity associated with the manual feature selection, more recent methods are based on deep learning (DL) approaches. These are able to obtain low-level and abstract representations of the ECG signal, which often result in deeper and more complete feature maps and therefore in better classification between high- and low-quality ECG excerpts than traditional ML algorithms [11], [12].

However, most of these algorithms have been validated on ECG recordings acquired from healthy subjects, and only a few ones have dealt with signals obtained from paroxysmal AF patients. This last context is much more challenging, and a loss of classification performance between 15% and 40% has been seen for almost all methods [12], [13]. For instance, the DL-based techniques proposed by Yoon et al. [14] and Zhang et al. [15] yielded discriminant powers between high- and low-quality ECG excerpts obtained from healthy individuals of about 90%, but they were significantly reduced to less than 75% when the ECG recordings were acquired from paroxysmal AF patients. Precisely, arrhythmic episodes change typical ECG morphology to quick fibrillatory waves, which present very similar aspect and time-frequency characteristics to the most common transient noise and artifacts observed during ECG acquisition [16]. Indeed, these nuisance interferences have been identified as the cause of more than 70% of false alarms of AF in conventional continuous ECG monitoring, both in intensive care units via bedside monitors [17] and in free-living conditions via insertable Holters [18].

In contrast to these previous DL-based algorithms that were trained from scratch [14], [15], a recent work has proven that, after a fine-tuning process on a limited dataset of ECG samples, a pre-trained convolutional neural network (CNN) was able to maintain a successful performance in the context of paroxysmal AF [12]. Thus, distributing more than 100,000 ECG segments in 500 learning-testing cycles of about 5,000 samples each, an accuracy greater than 90% was obtained, only misclassifying around 5% of clean AF segments as noisy excerpts. The core of this algorithm was the well-known CNN architecture of AlexNet, whose development in 2012 meant an unprecedented breakthrough in the field of DL [19]. Compared to common computer vision techniques, this network reduced classification error by 10% on a database with millions of images containing 1,000 classes of different objects [19], [20]. Since then, the transfer learning concept

has been widely exploited and the knowledge acquired in that experiment has been taken as the basis to tune AlexNet for many and diverse classification tasks, where promising outcomes have been mainly obtained [21], [22].

With the development of hardware capabilities and progressive rise of large databases in some fields, especially in computer vision, improved versions of AlexNet and other innovative CNN schemes have been recently proposed [21]. Initial advances were focused on aspects such as modification of processing units, strategies for optimizing parameters and hyper-parameters, connectivity between layers, etc. However, the focus of research was later shifted to improving architectural design of the networks [19]. The main idea was to enhance the performance of the CNN schemes by increasing their size, including the depth (i.e., the number of layers) and the width (i.e., the number of units at each layer). To this respect, Zeiler and Fergus [23] introduced by 2014 the concept of layer-wise visualization of the network to improve understanding of the feature extraction stages, which shifted the trend towards extraction of features at low spatial resolution in deep architectures, such as in VGG [24]. In fact, nowadays many novel CNN architectures are still built by repeating in cascade a simple and homogenous topology. Later, Google Deep Learning Group proposed the innovative idea of a split, transform and merge, with the corresponding block known as inception. This block introduced for the first time the concept of branching within a layer, which allowed abstraction of features at different spatial scales [25]. In 2016, the concept of skip connections was presented with ResNet [26] to obtain highly deep CNN schemes and, since then, it has gained growing popularity [19].

These novel CNN architectures, as well as others resulting from their combination, have provided better results than AlexNet when dealing with large, heterogeneous, and complex classification problems [21], [22]. Hence, considering this context and the fact that different networks extract diverse representations of the input data [19], the main goal of the proposed work is to analyze whether deeper and wider pre-trained CNN schemes than AlexNet can provide improved quality assessment of single-lead ECG recordings acquired by wearable devices from paroxysmal AF patients. Precisely, five common CNN models, which have been pre-trained on the same database as AlexNet [20], will be compared in terms of classification performance between high- and low-quality ECG excerpts and computational time required for training, validation, and testing. Under the same experimental setup, these algorithms will also be directly compared with other previously proposed and well-known ML-based techniques.

II. METHODS

All the tested networks were two-dimensional (2)-D) CNN schemes, receiving an image as input. Thus, ECG recordings acquired from paroxysmal AF patients were firstly segmented into 5 second-length excerpts with no overlap. Although most previous works have analyzed 10 second-length ECG segments, the window size seems to only have a negligible impact on the performance of many quality assessment algorithms based both on ML and DL concepts [27], [28]. Hence, to increase the number of ECG samples and more precisely delimit brief, transient noise [9], a window size of 5 seconds was selected. The resulting ECG portions were then transformed into a 2-D image and inputted to the CNN-based models. Note that no preprocessing (such as, filtering, transformation, etc.) was applied to the raw ECG signal to preserve its original morphology and avoid common artificial distortion provoked by most denoising methods [29].

A. DATABASES

Two databases were analyzed to consider a broad range of noises, artifacts, and ECG morphologies. Each set of ECG signals was acquired under diverse noisy conditions and ever-changing environments, as well as using different wearable recording systems. For training of the algorithms, a proprietary database (PDB) was firstly collected. To avoid common bias towards the majority class when training is conducted on imbalanced data [30], well-balanced subsets of high- and low-quality ECG excerpts were selected from single-lead, 2 hour-length ECG segments. These signals were extracted from longer continuous cardiac monitoring of 25 patients presenting paroxysmal AF episodes (12 women and 13 men, aged between 52 and 68 years). Briefly, the patients were continuously monitored for some weeks after catheter cryoablation making use of a textile wearable Holter system (Nuubo™), which acquired a continuous ECG signal from the patient's thorax with 12 bits of resolution over a dynamic range of ±5 mV and 250 Hz of sampling rate. The Ethical Review Board of Hospital Universitario San Juan de Alicante (Protocol Number UGP-14-219) approved this kind of cardiac monitoring for the patients, who gave express consent before the follow-up. Next, paroxysmal AF episodes were automatically detected by a previously published algorithm [31] and manually supervised by two cardiologists. They also identified high- and low-quality ECG excerpts. Whereas the first group was composed of the ECG segments exhibiting clearly and unequivocally visible R-peaks, the remaining ECG excerpts were included in the second subset. Other rhythms (OR), including supra-ventricular arrhythmias different from AF and ventricular and atrial premature contractions, were also annotated manually by the two experts. At last, the PDB was formed by 10,000 high-quality and 10,000 low-quality 5 second-length ECG portions. As Table 1 shows, the high-quality subset included 7,650, 1,750, and 600 ECG excerpts from normal sinus rhythm (NSR) segments, AF episodes, and OR intervals, respectively.

On the other hand, the analyzed algorithms were externally validated with the freely available training set of the PhysioNet/CinC Challenge 2017 (PC2017DB) [32], [33]. This database includes 8,528 ECG recordings with a duration ranging from 9 to 60 seconds. Connecting a portable AliveCor™ device to a smartphone, heart electrical activity

TABLE 1. Total amount of 5 second-length ECG excerpts for the two databases included in the study.

Class		Database		Total
		PDB	PC2017DB	
High-quality	NSR	7,650	28,413	57,539
	AF	1,750	4,329	
	OR	600	14,697	
Low-quality		10,000	1,168	11,168
Total		20,000	48,607	**68,607**

was on-demand recorded between the patient's hands with a sampling frequency of 300 Hz and 16 bits of resolution over a dynamic range of ±5 mV. Several experts manually labelled the ECG signals into four groups, such as NSR, AF, OR, and noisy excerpts. After segmentation, a total of 47,439 high-quality and 1,168 low-quality 5 second-length ECG portions were analyzed. As before, the high-quality group was composed of NSR, AF and OR excerpts, including 28,413, 4,329 and 14,697 samples, respectively. Note that considering both databases together, about 68,600 ECG segments were finally analyzed, as summarized in Table 1.

B. CONTINUOUS WAVELET TRANSFORM

A variety of alternatives to transform the ECG signal into a 2-D image can be found in the literature, e.g., Stockwell transform [34], modified frequency slice Wavelet transform [35], and short-time Fourier transform [15], among others. However, the most common option for that purpose is the use of continuous Wavelet transform (CWT), which has also reported successful performance when dealing with other physiological recordings [36], [37], [38], [39]. This tool analyzes a time series with variable resolution in a time-frequency map [40]. In short, translated and dilated instances of a wavelet function, which is called mother wavelet, are correlated with the original signal, then generating novel time series of wavelet coefficients with accurate positioning both in time and frequency domains. Mathematically speaking, CWT of the signal $x(t)$ is obtained as [41]

$$CWT(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t)\psi^*\left(\frac{t-b}{a}\right)dt, \quad (1)$$

where a is the dilatation factor, b is the translation parameter, $\psi(t)$ represents the mother wavelet function, and $*$ defines the complex conjugate operator. Preserving time and frequency information, 2-D color representation of the resulting matrix of wavelet coefficients is known as scalogram. This graph has been widely used to facilitate visual interpretation of the time-frequency decomposition of a signal [42], and it was employed as input for the analyzed CNN-based algorithms in the present work.

The parameters used to compute CWT, such as the mother wavelet function and number of time-frequency scales, as well as the chosen colormap for plotting wavelet coefficients, directly impact on the visual aspect of the resulting scalogram. Given that the Morlet function (i.e., a complex exponential function multiplied by a Gaussian window) is characterized by equal variance in time and frequency and has been used in a broad variety of ECG-based applications [12], [43], [44], it was selected as mother wavelet. Additionally, the number of wavelet scales was established to 48 voices per octave to achieve sufficient time resolution in the low frequency range, where physiological information is mainly concentrated for a clean ECG segment [45]. To minimize the impact of the different amplitude presented by the ECG portions in both databases (PDB and PC2017DB), the resulting wavelet coefficients were rescaled to the interval between 0 and 1. Finally, a Jet colormap of 7 bits was employed to draw the wavelet scalogram. As an example, wavelet scalograms for common 5 second-length ECG excerpts from NSR, AF, OR and noisy episodes are presented in Figure 1. The first three scalograms (related to high-quality ECG segments) display a repetitive pattern in the upper part, but the presence of a large motion artifact blurs such a periodical motif and provokes a more arbitrary layout in the last one.

C. PRE-TRAINED CNN MODELS

As previously mentioned, AlexNet is one of the most popular pre-trained CNN schemes [19] and its architecture has been well described [46]. As shown in Figure 2, it consists of five convolutional layers, three max-pooling layers, and three fully-connected layers, all connected in cascade and then resulting in a depth of 8 learnable layers and 60 million parameters [46]. The size of the kernel for convolution decreases as the number of layers increases, starting from a size of 11×11 in the first convolutional layer to 3×3 in the last convolutional one. Pooling layers are included to reduce spatial features without losing much information. Also, rectified linear unit (ReLU) activation functions are inserted after convolutional and fully-connected layers to limit the feature map to a positive range. Finally, to avoid overfitting, two drop-out regularizations are also incorporated.

A modification of AlexNet, named VGG16, was proposed in 2014 [24]. In this case, the depth was significantly increased by stacking convolutional layers and reducing the kernel size. Thus, this network stacks 13 convolutional layers with 3×3 kernel-sized filters, 5 max-pooling layers for feature extraction, and 3 fully-connected layers, as Figure 3 shows. As for AlexNet, the convolution layers are followed by ReLU functions, and different normalization and drop-out functions are interleaved at different points of the structure. This network presents a depth of 16 learnable layers and 138 million parameters.

In contrast to the stack of cascade layers, the network GoogLeNet introduced a new concept to increase the depth and reduce the computational cost [25]. This CNN scheme is based on the *inception module*, where convolutions are performed at various sizes in parallel. As an example, Figure 4 shows an inception module composed of four convolution branches with different kernel-sized filters ($1\times1, 3\times3, 5\times5$). The branches work in parallel to obtain spatial information at different scales, including both fine and coarse grain levels. More precisely, GoogLeNet contains nine inception modules

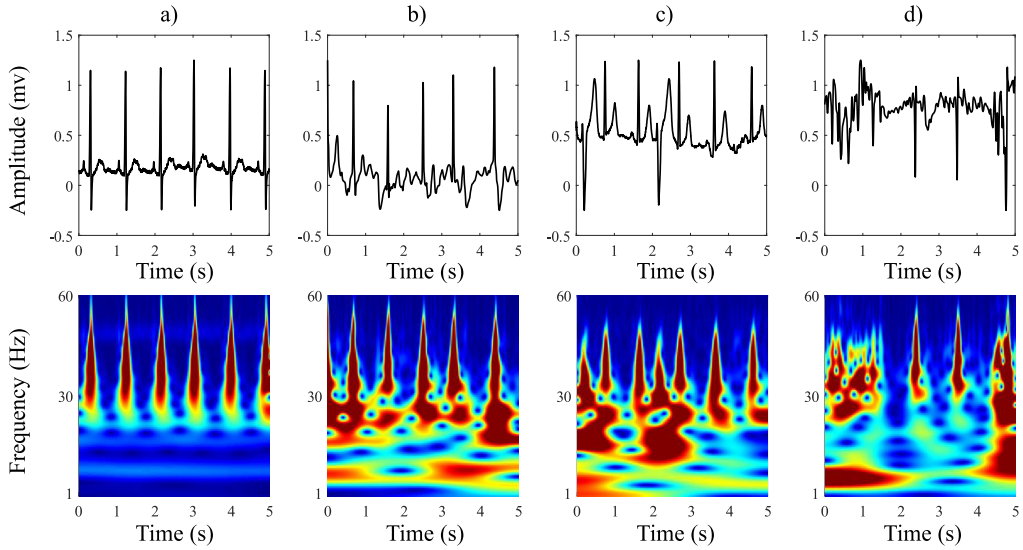


FIGURE 1. Typical 5 second-length ECG segments and wavelet scalograms obtained from (a) NSR, (b) AF, (c) OR, and (d) noisy episodes.

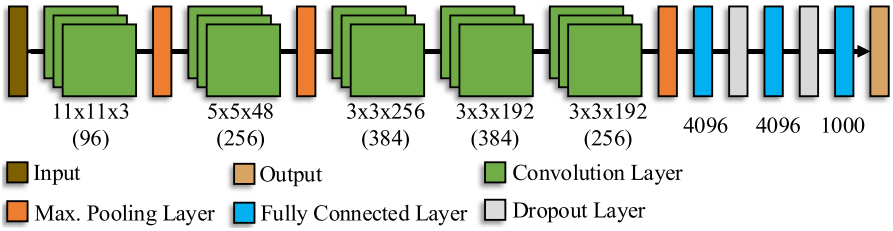


FIGURE 2. Layer-based architecture of AlexNet.

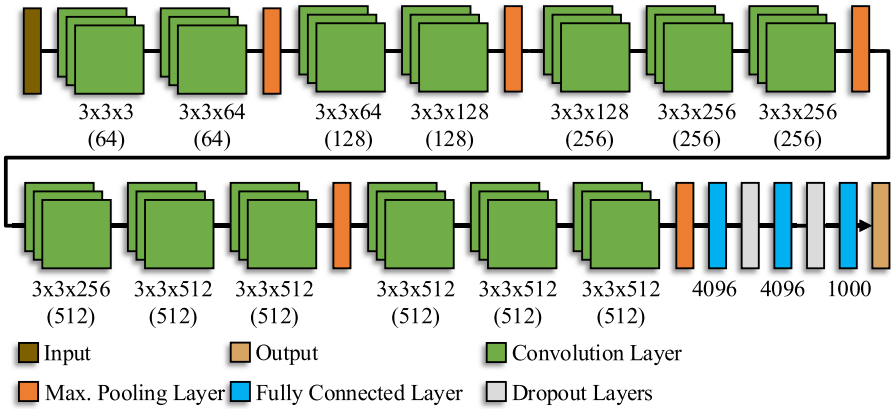


FIGURE 3. Layer-based architecture of VGG16.

as the one presented in Figure 4, two convolutional layers, four max-pooling layers, three average pooling layers, five fully-connected layers, and three soft-max layers, such as Figure 5 displays. Moreover, it uses drop-out regularization after some fully-connected layers and applies ReLU activations after all convolutional layers. Although this network is much deeper and wider than AlexNet, it has a much lower number of parameters [25]. In fact, GoogLeNet presents a depth of 22 learnable layers and 6.7 million parameters.

Recent empirical research has reported that some deep neural networks (stacking many convolutional layers or blocks) perform worse than shallow ones, even when no overfitting has occurred [19]. This is often associated with an extremely large increase (exploiting gradient) or decrease (vanishing gradient) of the gradient during the back-propagation process for the network training. To overcome this problem, the model ResNet proposes the use of residual learning blocks, where the input features of the upper layers are reused in

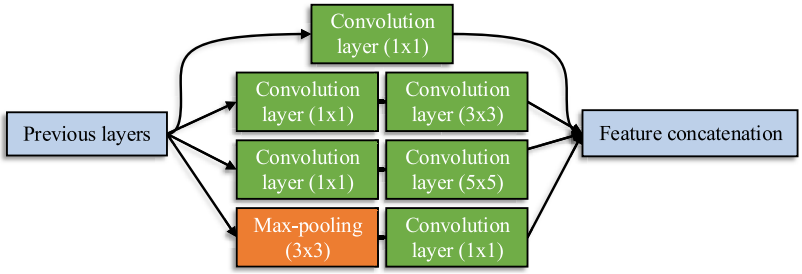


FIGURE 4. Layer-based architecture of a common inception module composed of four branches in parallel.

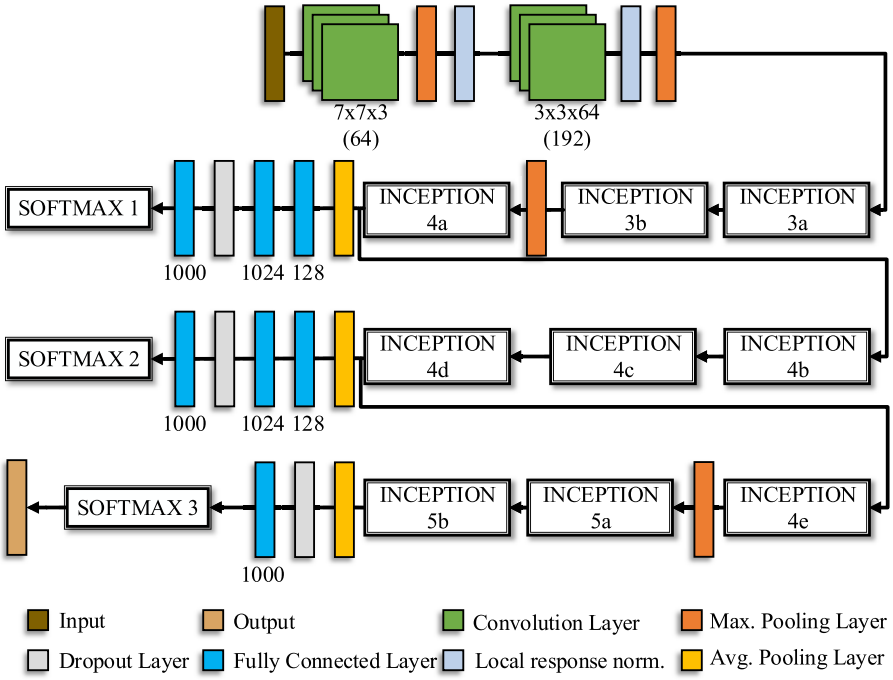


FIGURE 5. Layer-based architecture of GoogLeNet.

deeper ones [26]. This idea is inspired by the connection between neurons in the cerebral cortex and allows to obtain highly deep networks without a loss of generalization. In fact, ResNet presents a sequential architecture similar to AlexNet and VGG, but it is about 20 and 8 times deeper, respectively. Hence, the variant ResNet50 was considered in the present study for ECG quality assessment. This network contains 49 convolutional layers and a fully-connected layer, resulting in a depth of 50 learnable layers and 25.6 million parameters. The architecture of ResNet50 is shown in Figure 6.

In the last years, more compact and lightweight CNN schemes have been pursued to achieve advances in their distributed training, their export of new models from the cloud, and their deployment on resource-constrained FPGA-based systems [19]. These novel networks seek to maintain levels of depth similar to previous ones, but significantly reducing the number of parameters. To this respect, the model SqueezeNet has been proposed as a more compact version of AlexNet, but with 50 times fewer parameters [47]. Indeed, 1×1 kernel-sized filters are used instead of 3×3. Moreover, this network is composed of blocks called *fire modules*, which contain a squeeze convolution layer with 1×1 filters and an expand layer with a mix of 1×1 and 3×3 convolution filters. As Figure 7 shows, SqueezeNet has an initial and a final convolution layer, while the central part is composed of 8 fire module blocks. No fully connected layers are used, but an average pooling is incorporated before the final soft-max classifier. The resulting network has a depth of 18 learnable layers and 1.24 million parameters.

With the same idea of making computation more efficient and simultaneously maintaining the network's depth, the model ShuffleNet has also been recently introduced [48]. This CNN scheme is based on a small structure composed of *shuffle blocks*, where conventional convolution is replaced by 1×1 convolution and a channel shuffle layer. This last one enables cross-group information flow among the three image channels (i.e., R, G and B) for multiple group convolution layers. As Figure 8 displays, ShuffleNet is composed of three

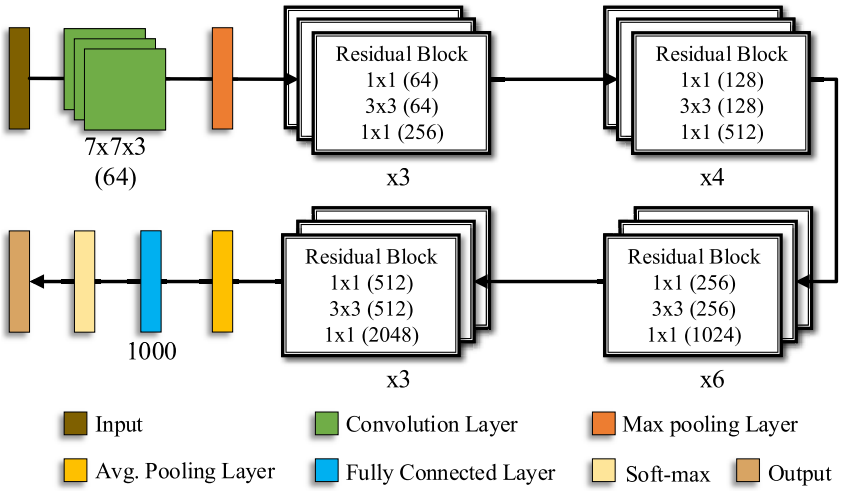


FIGURE 6. Layer-based architecture of ResNet50.

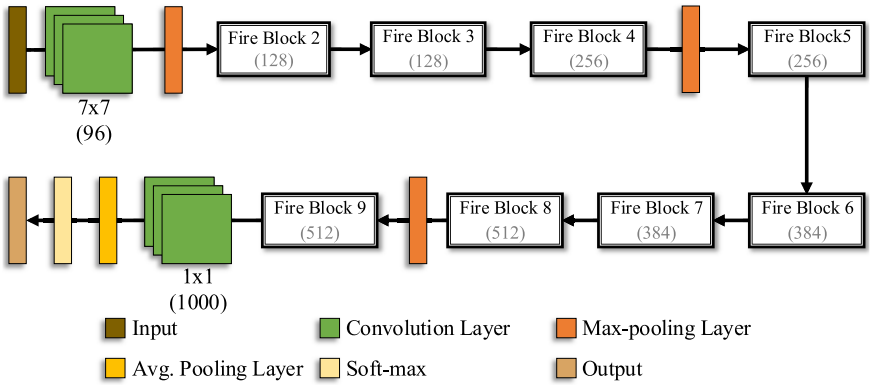


FIGURE 7. Layer-based architecture of SqueezeNet.

shuffle blocks placed in cascade between a group convolution layer and a fully-connected one. As in other networks, diverse pooling layers and other functions to prevent overfitting are also interleaved in several points of the structure. Finally, ShuffleNet presents a depth of 50 learnable layers and only 1.4 million parameters [48].

D. PREVIOUSLY PROPOSED TRADITIONAL ML ALGORITHMS

Some previously proposed and widely used quality indices (QI) for single-lead ECG signals have also been implemented for comparison. These methods are mainly based on combining hand-crafted features derived from the ECG signal with conventional ML classifiers. One of the most referenced methods was developed by Clifford et al. [49]. This, hereinafter referred to as QI_1, was composed of four metrics merged through an SVM classifier [49]. The single metrics were defined as the percentage of R-peaks identified by two published detectors (bSQI), the relative power in the QRS complex (pSQI), the fourth moment (i.e., kurtosis) of the ECG signal (kSQI), and the relative power in the ECG baseline (basSQI). An improved version of this index, which will be here named QI_2, has also been proposed by the same research group in a subsequent work [13]. In this case, three additional single metrics were considered, and a total of seven features were combined by an SVM classifier. The novel metrics were the third moment (skewness) of the ECG signal (sSQI), the ratio of the number of beats detected by the two R-peak detectors (rSQI), and finally the ratio of the sum of the eigenvalues associated with the five principal components and the sum of all eigenvalues obtained by principal component analysis applied to the time-aligned ECG beats (pcaSQI). The combination of the four metrics used by QI_1 through heuristic rules has also been proposed for ECG quality assessment [50], and it was also computed and referred to as QI_3.

More recently, Albaba et al. [51] have studied a broad variety of features derived from the ECG signal to discern between high- and low-quality excerpts in three databases containing recordings acquired by traditional, wearable, and ubiquitous devices. After applying different feature selection techniques, the authors proposed three different methods, one for each database, based on combining several hand-crafted features with an SVM classifier. The first method, here

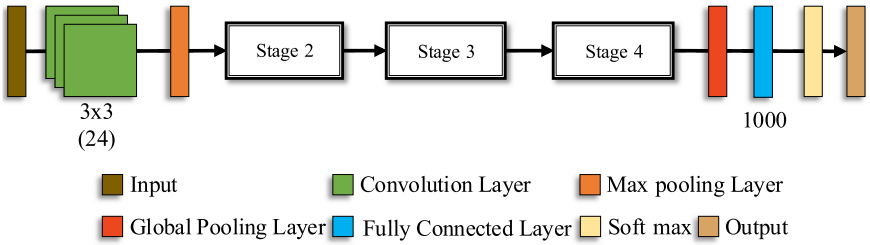


FIGURE 8. Layer-based architecture of ShuffleNet.

referred to as QI_4, merged seven features, i.e., irregularity for wavelet scales 1, 2, and 5 computed via approximate entropy, standard deviation for wavelet scales 3 and 5, and finally location of the first local maximum and the first zero-crossing in the autocorrelation function of the ECG signal. The second index, here named QI_5, was also composed of seven variables, such as mean, maximum, kurtosis and skewness of the spectral distribution of the ECG signal, median absolute deviation of the wavelet scales 3 and 5, and finally location of the first zero-crossing in the autocorrelation function. The last algorithm, here called QI_6, was defined by combining ten features, i.e., mean and irregularity of the ECG signal, standard deviation, skewness, and irregularity of the spectral distribution of the ECG signal, median absolute deviation of the wavelet scale 1, mean of the wavelet scale 2, and finally amplitude of the first local maximum, amplitude of the first local minimum, and location of the first zero-crossing in the autocorrelation function of the ECG signal.

E. PERFORMANCE ANALYSIS

As previously mentioned in Section II-A and following recommendations from the Transparent Reporting of a multivariate prediction model for Individual Prognosis Or Diagnosis (TRIPOD) initiative [52], two totally separate datasets were used for training and testing all the CNN-based algorithms and the indices QI_1–QI_6. Unfortunately, patient-specific training and testing of common supervised learning algorithms, i.e., considering samples from the same patients for training and testing, often prevents them from optimal generalization [53], [54]. Indeed, both ML- and DL-based algorithms are able to memorize patient-specific features, especially when the training dataset is limited, and then classification performance trends to be inflated [53], [54].

On the one hand, the PDB was used for training the models. It was divided into two groups with the idea of monitoring the learning of the CNN-based methods. Precisely, the dataset was stratifiedly split such that 80% of the samples were used for training and the remaining 20% for validation, such as Table 2 summarizes. Although a patient-independent approach was not considered in this stage, unseen ECG samples during training were used for validation. In this way, classification error on the validation subset could still be considered as a reliable estimate of the learning progression for a network during training [55], and it was used to determine the maximum number of epochs to train every

TABLE 2. Distribution of 5 second-length ECG segments for training and validation of the CNN models and the indices QI_1–QI_6 from the PDB.

Class		PDB	
		Training Subset	Validation Subset
High Quality	NSR	6,120	1,530
	AF	1,400	350
	OR	480	120
Low Quality		8,000	2,000
Total		16,000	4,000

CNN model. To this respect, a stable validation error was reached by all CNN-based algorithms after 4 to 7 epochs, and they were then trained for 10 epochs. Note that, after some empirical tests, the batch size was established to 32 and the learning rate to 0.0001, and the cross-entropy, computed for the two output nodes, was used as loss function. The CNN models were initialized with the weights obtained from their pre-training on more than 1.2 million of natural images, i.e., on the ImageNet database [20], and they were then fine-tuned through a stochastic gradient descendent algorithm with a momentum of 0.9.

The SVM classifiers included in the ML-based indices QI_1, QI_2, QI_4, QI_5, and QI_6 were also trained and validated on these subsets of data. Although a gaussian kernel was considered for all cases, different values of scale γ and maximum penalty on margin-violating observations C were used in diverse models. Thus, parameters $\gamma = 1$ and $C = 25$ were used for QI_1 and QI_2, and $\gamma = 2$ and $C = 1$ for QI_4, QI_5 and QI_6. The index QI_3 did not require any kind of training, because it was based on heuristic rules.

On the other hand, the freely available PC2017DB was used for external testing both of the CNN-based models and the ML-based indices. As Table 1 shows, this dataset is strongly unbalanced, presenting much more high-quality ECG samples than low-quality ones. However, this aspect has no great impact in the testing phase, and its use is highly interesting because direct comparison with previous or future methods will be possible. Moreover, to avoid overestimated classification by common performance metrics on unbalanced datasets, specifically designed indices for imbalanced contexts were computed, such as Balanced Accuracy ($BAcc$), F_1 score, and Matthews correlation coefficient (MCC) [56], along with sensitivity (Se), specificity (Sp), and their representation on a receiver operating characteristic (ROC) curve. Hence, considering the true positives (TP) and false

negatives (FN) as the high-quality ECG excerpts correctly and wrongly identified, respectively, and the true negatives (TN) and false positives (FP) as the low-quality ECG segments properly and incorrectly classified, respectively, the following performance metrics were computed:

$$Se = \frac{TP}{TP + FN}, \tag{2}$$

$$Sp = \frac{TN}{TN + FP}, \tag{3}$$

$$BAcc = \frac{Se + Sp}{2}, \tag{4}$$

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}, \text{ and} \tag{5}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{6}$$

whereas Se, Sp, $BAcc$, and F_1 provide values between 0 and 1, MCC ranges from -1 to 1. Nonetheless, to make interpretation of all classification results easier, a normalized version of MCC to the interval [0, 1] was computed as

$$nMCC = \frac{MCC + 1}{2}. \tag{7}$$

Additionally, to quantify how AF episodes were discerned from noisy and OR excerpts, the rates of correctly classified NSR ($\mathcal{R}_{NSR}$), AF ($\mathcal{R}_{AF}$), and OR ($\mathcal{R}_{OR}$) segments belonging to the high-quality group were also estimated. Computation times required by each algorithm for a 5 second-length ECG sample in training, validation, and testing were measured as well. All tests were conducted with MATLAB 2020a on an HP Workstation Z230, equipped with a 64-bit operating system, 32 GB RAM, and an Intel(R) Xeon @ 3.60 GHz processor, running a graphics processing unit GeForce GTX 1060 with 6 GB dedicated VRAM.

III. RESULTS

Table 3 displays classification results obtained both by the CNN-based models and the indices QI_1–QI_6 on the validation subset, which was composed of 4,000 ECG segments extracted from the PDB (i.e., 20% of its total content). As can be observed, all CNN-based algorithms reported values of Se, Sp, $BAcc$, F_1, and $nMCC$ larger than 98%. Similarly, they also correctly classified more than 98% of the ECG excerpts from NSR and AF episodes. Although their ability to identify ECG portions from OR was slightly lower, the ratio $\mathcal{R}_{OR}$ was still higher than 93% for all CNN models. Compared to these results, the indices QI_1–QI_6 obtained a poorer performance. Some methods like QI_1, QI_2, QI_4 and QI_6 achieved values of Se, Sp, $BAcc$, F_1, and $nMCC$ about 95%, but QI_3 and QI_5 only reached values about 70% and 90%, respectively. Moreover, most of the indices failed to properly identify about 20–30% of the ECG samples extracted from OR. To this respect, whereas the indices QI_1–QI_3 reported ratios of $\mathcal{R}_{OR}$ about 80–85%, the methods QI_4-QI_6 only presented values between 65% and 71%. Of note is also that the rule-based index QI_3 was only able to correctly identify about 30% of AF segments.

Regarding external testing, classification outcomes obtained by all the analyzed algorithms on the PC2017DB are shown in Table 4, and ROC curves are presented in Figure 9. In this case, more notable differences among the performance of the CNN-based algorithms were noticed. Thus, although all provided values of F_1 and $nMCC$ larger than 88% and 63%, respectively, differences about 4–7% were observed between the algorithms providing the best and worst performances. Thus, whereas AlexNet reported values of F_1 and $nMCC$ about 95% and 67%, ResNet50 about 88% and 63%, respectively. The remaining models presented values halfway between these two extremes. Moreover, only VGG16, GoogLeNet and SqueezeNet reported balanced values of Se and Sp, because the remaining networks exhibited differences about 8–10% between these two performance metrics. For instance, whereas AlexNet showed values of Se and Sp about 90% and 80%, ResNet50 reported opposite values about 80% and 90%, respectively. On the contrary, a far less variability about 1% was noticed among the ratios $\mathcal{R}_{NSR}$, $\mathcal{R}_{AF}$ and $\mathcal{R}_{OR}$ for all the CNN-based algorithms.

In comparison with these outcomes, again those obtained by the indices QI_1–QI_6 on the testing subset were strongly poorer. Precisely, they yielded a performance between 5% and 16% lower in terms of F_1 and between 3% and 8% lower in terms of $nMCC$, both performance metrics remaining below 85% and 60.5%, respectively (see Table 4). Also, most indices exhibited remarkable differences about 10% or longer between the values of Se and Sp. To this respect, an extreme case was seen for the index QI_4, which presented values of Se and Sp about 52% and 90%, respectively. Similarly, within the high-quality group, differences between 5% and 10% were also observed among the ratios $\mathcal{R}_{NSR}$, $\mathcal{R}_{AF}$ and $\mathcal{R}_{OR}$ for most of the indices.

Finally, Table 5 and Figure 10 presents average time spent by all algorithms on a 5 second-length ECG excerpt in training, validation and testing. As can be seen, the CNN-based models required much more time for training than for validation and testing. Whereas training time on each ECG-based image ranged from 0.1 to 2.1 seconds, depending on the CNN architecture, validation and testing required always less than 0.12 seconds. Contrarily, the indices QI_1–QI_6 needed a very similar time for training, validation and testing, since the most time-consuming task was computation of the handcrafted features. Nonetheless, notable differences in speed were noticed between the groups of indices QI_1–QI_3 and QI_4–QI_6, those included in the first set being significantly faster than those in the second one. In fact, it is noteworthy that the algorithms of the second group required even much more time for validation and testing than most of the CNN-based algorithms.

IV. DISCUSSION

To the best of our knowledge, the present work has introduced for the first time an exhaustive comparison among

TABLE 3. Classification results obtained by the CNN-based algorithms and the indices QI_1–QI_6 on the validation subset, extracted from the PDB. Bold value indicates the highest value for each performance metric.

Algorithm	Se (%)	Sp (%)	BAcc (%)	F_1 (%)	nMCC (%)	$\mathcal{R}_{NSR}$ (%)	$\mathcal{R}_{AF}$ (%)	$\mathcal{R}_{OR}$ (%)
AlexNet	98.29	**99.40**	98.85	98.84	98.85	**99.48**	**100**	**96.67**
VGG16	**99.55**	98.31	98.93	98.94	98.94	98.19	99.42	**96.67**
GoogLeNet	98.99	99.11	**99.05**	**99.05**	**99.05**	99.42	99.41	93.33
ResNet50	99.35	98.06	98.70	**99.05**	**99.05**	98.00	**100**	93.33
SqueezeNet	98.94	98.46	98.70	98.71	98.70	98.52	**100**	93.33
ShuffleNet	98.89	98.66	98.78	98.78	98.78	98.71	99.42	95.83
QI_1	93.74	95.47	94.61	94.57	94.61	93.35	99.42	82.50
QI_2	97.02	96.17	96.59	96.62	96.60	97.61	99.71	81.67
QI_3	81.53	60.62	71.07	73.90	71.56	92.72	30.14	85.00
QI_4	95.93	94.91	95.42	95.46	95.42	97.35	**100**	65.83
QI_5	87.88	92.30	90.09	89.89	90.12	88.96	88.99	70.83
QI_6	94.84	93.30	94.07	94.14	94.08	95.61	99.71	70.83

TABLE 4. Classification results obtained by the CNN-based algorithms and the indices QI_1–QI_6 on the testing database, i.e, on the PC2017DB. Bold value indicates the highest value for each performance metric.

Algorithm	Se (%)	Sp (%)	BAcc (%)	F_1 (%)	nMCC (%)	$\mathcal{R}_{NSR}$ (%)	$\mathcal{R}_{AF}$ (%)	$\mathcal{R}_{OR}$ (%)
AlexNet	**90.42**	80.99	85.70	**94.74**	**67.27**	**90.31**	**89.77**	**90.81**
VGG16	85.77	86.73	86.25	92.18	65.15	85.60	85.29	86.23
GoogLeNet	88.79	85.66	87.22	93.89	66.95	88.75	88.13	89.08
ResNet50	79.02	**92.89**	85.96	88.19	63.14	78.71	79.35	79.51
SqueezeNet	84.66	90.07	**87.37**	91.57	65.19	84.60	84.48	84.83
ShuffleNet	82.24	91.61	86.93	90.15	64.26	82.19	82.01	82.40
QI_1	66.38	83.90	75.14	79.60	58.08	67.60	66.87	63.88
QI_2	72.13	88.44	80.29	83.67	60.19	72.66	72.70	70.92
QI_3	72.42	72.43	72.43	83.67	57.60	76.32	69.39	65.75
QI_4	51.93	89.81	70.87	68.25	56.39	55.06	52.99	45.57
QI_5	74.20	80.48	77.34	84.96	59.42	75.55	68.42	73.29
QI_6	66.89	91.01	78.95	80.05	59.33	67.12	68.86	65.86

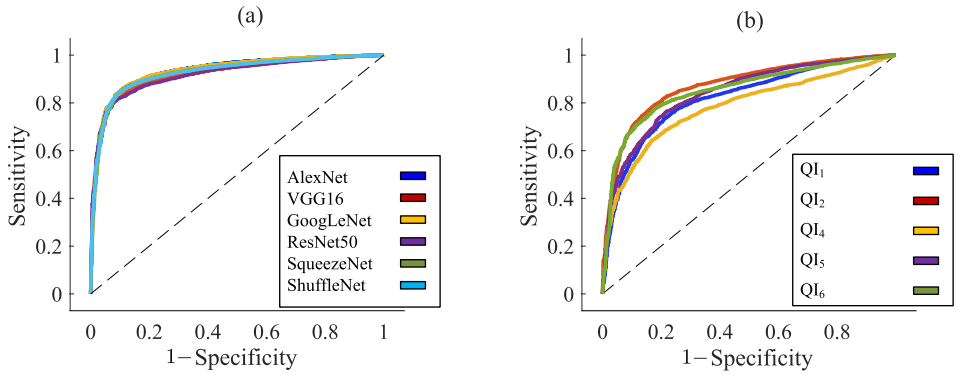


FIGURE 9. ROC curves obtained by (a) CNN-based and (b) ML-based algorithms on the testing database, i.e., on the PC2017DB. Note that a ROC curve for QI_3 was not computed because it was based on heuristic rules.

pre-trained CNN-based models and previously proposed traditional ML algorithms for quality assessment of single-lead ECG recordings acquired with wearable devices from paroxysmal AF patients. All the methods have been trained, validated and tested under the same experimental conditions by making use of the same personal computer, databases, and learning parameters. Thus, a fair and comparable analysis was ensured, and indirect comparison of results previously reported on different databases and varied validation approaches was also prevented. Moreover, patient-independent training and testing processes have been conducted on two separate databases to obtain a realistic view of the methods' generalization capability. In this way, potential bias associated with memorization of patient-specific features during training was avoided [53], [54], a fact that could explain the remarkable loss of performance exhibited by all the analyzed algorithms from validation to testing. To this respect, Tables 3 and 4 shown decreases higher than

TABLE 5. Average time required by all algorithms on a 5 second-length ECG interval in training, validation and testing. The values are expressed in seconds.

	CNN-based Algorithms					
Database	AlexNet	VGG16	GoogLeNet	ResNet50	SqueezeNet	ShuffleNet
Training	0.115	2.138	0.404	0.681	0.134	0.429
Validation	0.041	0.121	0.054	0.065	0.043	0.050
Testing	0.040	0.119	0.059	0.066	0.043	0.049
	ML-based Algorithms					
Database	QI_1	QI_2	QI_3	QI_4	QI_5	QI_6
Training	0.007	0.010	0.006	0.247	0.075	0.128
Validation	0.004	0.009	0.005	0.247	0.074	0.128
Testing	0.004	0.008	0.006	0.247	0.074	0.128

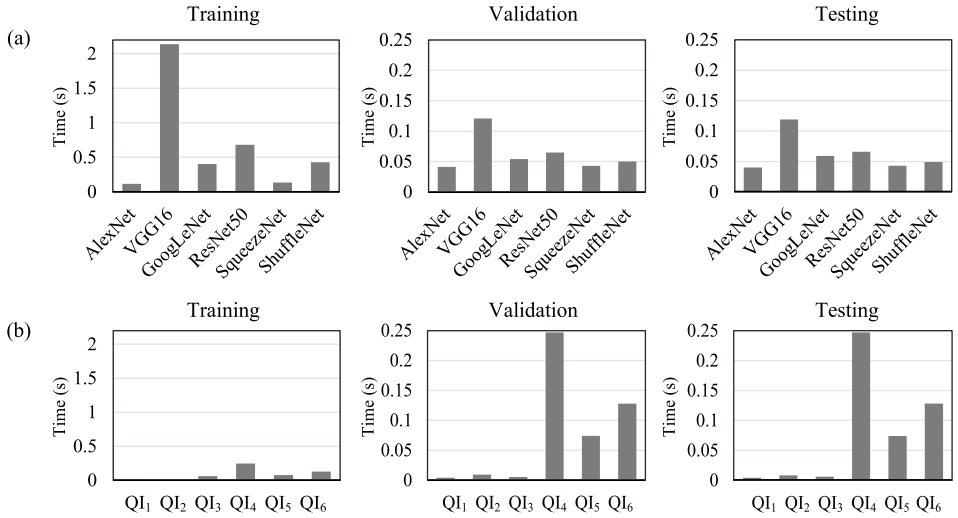


FIGURE 10. Average time required by (a) CNN-based and (b) ML-based algorithms on a 5 second-length ECG interval in training, validation and testing.

4% and 31% in terms of F_1 and *nMCC* from the scenario where ECG excerpts of the same patients were shared in training and validation to the one where an external database was used for testing.

On average, the loss of performance was significantly higher for most of the traditional ML algorithms than for CNN-based algorithms, because F_1 and *nMCC* reported mean decreases of 14.83±7.99% and 35±3.08% versus 7.11±2.42% and 33.57±1.58%, respectively, when the index QI_3 (based on heuristic rules) was excluded. Moreover, in absolute terms, all the CNN-based models also exhibited better performance than the indices QI_1–QI_6 on the testing database. Precisely, whereas the CNN-based models always reported values of F_1 and *nMCC* greater than 88% and 63%, traditional ML algorithms presented values lower than 85% and 60.5%, respectively. Both outcomes suggest that the neural networks were able to obtain more complete and abstract feature maps than those manually derived for ECG quality assessment, and consequently higher levels of generalization. Similar findings have been also reported by previous works where both kinds of supervised learning approaches were compared in ECG-based contexts different from quality assessment, such as automated detection of apnea [57], identification of shockable rhythms for automated defibrillation [58], classification of diverse cardiovascular diseases [59], or screening of paroxysmal AF [60].

Moreover, these results also agree with other well-known observations in the scientific literature. In fact, conventional ML-based algorithms has often been reported to need a huge number of observations to achieve high levels of generalization [61]. To this respect, in the specific context of ECG quality assessment, Albaba et al. [51] corroborated that the set of indices QI_4–QI_6 trained on about 10,000 ECG excerpts were not sufficiently general to work with ECG signals obtained from other different datasets. For instance, whereas the index QI_5 reported a value of *BAcc* about 95% on the dataset used for its training, values lower than 50% were obtained on two external databases [51]. Contrarily, stacking of multiple linear and non-linear processing units in a layer-wise fashion has recently provided a strong ability to extract morphology-independent ECG features and then reach deep levels of generalization, even when the number of training samples is reduced [19]. This idea is also supported in the present work by the fact that the CNN-based models

presented a notably higher ability to detect ECG excerpts from OR than the indices QI_1–QI_6. Although the number of these samples was notably reduced in the training subset (only 480), the CNN-based models were able to identify correctly more than 93% and 80% on the validation and testing databases, whereas the indices QI_1–QI_6 lower than 85% and 75%, respectively.

Nonetheless, it is worth noting that the CNN-based models achieved such high levels of abstraction on the ECG signal at the cost of computationally intensive training. Indeed, this is an iterative process of layer-by-layer evolution, which often takes from several minutes to some hours, depending on the model's depth. However, once the networks were trained, they required a processing time significantly higher than the set of indices QI_1–QI_3, but lower than the group of indices QI_4–QI_6. This result could be explained by the fact that computation of some hand-crafted features, such as approximate entropy, are also computationally expensive [62]. Anyway, today there exist programmable devices which have been specifically developed to run CNN models in real-time [63], if needed. Moreover, many wearable systems for continuous ECG monitoring are based on a device that captures the signal and transmits it to a cloud platform via an internet connection [7]. Then, the cloud server computes off-line tasks for ECG assessment, processing, and data extraction for interpretation and diagnosis. In this context, the computational cost is not really a limitation for the use of CNN-model models in ECG quality assessment [64].

On the other hand, establishing a direct comparison among the performance of all the CNN-based models, in absolute terms no great differences were noticed even on the testing database. Nonetheless, the best performers (i.e., AlexNet and GoogLeNet) reported about 6% and 4% greater values in F_1 and $nMCC$ than the worst one (i.e., ShuffleNet and ResNet50). As well, these last CNN models reported the largest relative loss of performance from validation to testing. Although both networks contained a moderate number of learnable parameters (ShuffleNet about 1.4 million and RestNet50 about 25.6 million) compared to the remaining ones (from 1.26 million for SquezeeNet to 138 millions for VGG16), they were the most depth with 50 layers. Precisely, regardless of the number of learnable parameters, a direct relationship between the network's depth and the relative loss of performance from validation to testing was noticed. To this respect, results in Tables 3 and 4 show that the shallower network (i.e., AlexNet with 8 layers) reduced its performance about 4% and 31.5% in terms of F_1 and $nMCC$, the networks with a moderate depth (i.e., VGG16 with 16 layers, SqueezeNet with 18 layers, and GoogLeNet with 22 layers) about 5–7% and 32–33.5%, and finally the deeper networks (ResNet50 and ShuffleNet with 50 layers) about 8.5–11% and 34.5%–36%, respectively.

Despite the fact that all CNN-based models were trained on a totally balanced dataset of clean and noisy ECG segments, a clear relationship between the network's depth and a higher disparity between the values of Se and Sp obtained on the testing database was also observed. In this case, the shallowest network (i.e., AlexNet) reported a notably better ability to identify high-quality ECG excerpts ($Se = 90.42\%$) than low-quality ones ($Sp = 80.99\%$), the networks with a moderate depth (i.e., VGG16, SqueezeNet, and GoogLeNet) exhibited notably balanced values of Se and Sp with differences lower than 5%, and finally the deeper networks (i.e., ResNet50 and ShuffleNet) provided an opposite trend to AlexNet by revealing higher values of Sp than Se, more precisely, $Se = 79.02\%$ and $Sp = 92.89\%$ for ResNet50, and $Se = 82.24\%$ and $Sp = 91.61\%$ for ShuffleNet.

These results suggest that the learning ability presented by shallow CNN schemes, such as AlexNet, could only be sufficient to abstract the most relevant features from high-quality ECG excerpts, which present high similarity and low morphological diversity. Contrarily, the subset of low-quality ECG signals includes a higher pool of different and chaotic morphologies, and much deeper neural networks could then be required to reach high levels of abstraction. However, the scarcity of morphological variability among the high-quality ECG excerpts could lead to overtrain highly deep CNN-based models for this group of signals, thus explaining the lowest values of Se reported by ResNet50 and ShuffleNet. As a consequence, CNN schemes with a moderate depth of about 20 layers seem to reach optimal trade-off of generalization for both high- and low-quality groups of ECG excerpts. It should be noted that a balanced classification between both groups is desired in the context of ECG quality assessment to equally reduce the risk of misdiagnosis by interpreting noisy signals, as well as of loss of clinical information by discarding clean ECG excerpts.

The best trade-off between Se and Sp was achieved by VGG16, with less than 1% difference between both metrics. However, this network is an extended version of AlexNet by stacking many additional layers and therefore making its computational load heavy (about 13 times more time than AlexNet to classify each 5 second-length ECG excerpt). Moreover, although F_1 and $nMCC$ were only about 2% lower than for AlexNet, VGG16 reported decreases about 5% in classification rates of NSR, AF and OR intervals on the testing dataset (see Table 4). Hence, a better option with still well-balanced values of Se and Sp (difference about 3%) and a computational cost only 3 times more than AlexNet was GoogLeNet. In absolute values, this CNN model also reported a very similar performance to AlexNet in terms of F_1, $nMCC$, $\mathcal{R}_{NSR}$, $\mathcal{R}_{AF}$, and $\mathcal{R}_{OR}$ on the testing database. This good and balanced performance might be explained by the parallel configuration of GoogLeNet, which leads to a multi-scale analysis. Indeed, previous works have already proven the strong capability of multi-scale processing, combined with common entropy-based metrics, to identify typical artifacts and noises in the ECG signal [65].

SqueezeNet may also be an interesting alternative to AlexNet for ECG quality assessment, because it yielded a

moderate difference of about 5% between values of *Se* and *Sp*. Moreover, the network also achieved a larger classification rate of low-quality ECG segments ($Sp = 90.07\%$) than high-quality ones ($Se = 84.66\%$), which is more interesting than the opposite trend in the context of quality assessment of long-term, wearable ECG recordings. Indeed, in this case, maximizing the detection of low-quality ECG segments at the cost of slightly increasing the rate of false positives would not be a significant drawback, since many hours of ECG signals are recorded and many high-quality portions could be easily analyzed for making precise interpretations and diagnoses. The network also showed a similar computation load than AlexNet, but its performance in terms of $\mathcal{R}_{NSR}$, $\mathcal{R}_{AF}$, and $\mathcal{R}_{OR}$ was about a 5% lower.

Finally, some limitations deserve mention. Thus, only pre-trained CNN schemes on a popular database, such as ImageNet [20], were compared in the present work. Although the knowledge learned by many CNN schemes on such dataset of natural images has been helpful to improve their performance on many classification tasks [21], [22], it is today controversial whether pre-training on images more similar to the target ones could make transfer learning more effective [66], [67]. However, this aspect has still not been tackled in the specific context of ECG quality assessment. Similarly, a thorough analysis about the performance of diverse CNN architectures trained from scratch to discern between high- and low-quality ECG portions is not found in the literature. Hence, in a future work the impact of pre-training and training from scratch several CNN schemes using in-domain ECG-based images and out-of-domain natural images on quality assessment of wearable ECG recordings will be addressed.

On the other hand, some 5 second-length ECG excerpts could have been mislabeled in the PC2017DB, because the complete recordings (lasting between 9 and 60 seconds) received a single manual annotation [32]. Thus, ECG recordings annotated as low-quality by the presence of highly localized artifacts might still present some high-quality 5 second-length excerpts, which would have inherited the inappropriate original label. However, to facilitate comparison with other works, these ECG excerpts were not relabelled with the same criteria as in the PCB.

Also of note is that a binary classification between high- and low-quality ECG segments was only considered. Although discerning among more levels of quality could improve further manual ECG interpretation by discarding those portions with clear R-peaks but masked or distorted P- and/or T-waves, no standard and strict limits exist for such categorization [68]. Hence, multiclass classification could involve a subjective bias, regarding the even-handed used criterion of identifying clear R-peaks, that might reduce generalization of the CNN-based models when used prospectively in databases annotated by different experts. Moreover, it should be remarked that heart rhythm analysis based on R-peak information is the most commonly conducted on very long-term ECG recordings acquired with wearable systems from paroxysmal AF patients [69].

V. CONCLUSION

The established comparison of supervised learning algorithms under a unified framework for ECG quality assessment has revealed that those based on modern deep learning concepts reached better generalization, and therefore higher discriminant power on unseen data, than those based on handcrafted features and conventional machine learning classifiers. Among the deep learning methods, deeper and wider CNN models than AlexNet only reported a similar or slightly poorer performance in global terms. However, it was clearly noticed that, regardless of the number of learnable parameters, shallow networks trended to detect better high-quality ECG excerpts and deep CNN models to identify better noisy ECG portions. As a desired trade-off, the networks with a moderate depth (of about 20 layers) presented the best balanced performance on both groups of ECG excerpts. To this respect, GoogLeNet yielded close classification rates for both high- and low-quality ECG segments, maintaining a similar misclassification rate of AF episodes to AlexNet and an acceptable computation time. As a consequence, this pre-trained CNN scheme seems to be the best alternative for quality assessment of wearable, long-term ECG recordings acquired from patients with paroxysmal AF. In the future, its ability to reduce AF false alarms and improve AF detection without manual intervention will be systematically analyzed on patients under continuous ECG monitoring.

REFERENCES

[1] G. Lippi, F. Sanchis-Gomar, and G. Cervellin, "Global epidemiology of atrial fibrillation: An increasing epidemic and public health challenge," *Int. J. Stroke*, vol. 16, no. 2, pp. 217–221, Feb. 2021.

[2] G. Hindricks et al., "2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS): The Task Force for the diagnosis and management of atrial fibrillation of the European Society of Cardiology (ESC) developed with the special contribution of the European Heart Rhythm Association (EHRA) of the ESC," *Eur. Heart J.*, vol. 42, no. 5, pp. 373–498, Feb. 2021.

[3] I. Escudero-Martínez, L. Morales-Caba, and T. Segura, "Atrial fibrillation and stroke: A review and new insights," *Trends Cardiovascular Med.*, vol. 33, no. 1, pp. 23–29, Jan. 2023.

[4] M. P. Turakhia, J. D. Guo, A. Keshishian, R. Delinger, X. Sun, M. Ferri, C. Russ, M. Cato, H. Yuce, and P. Hlavacek, "Contemporary prevalence estimates of undiagnosed and diagnosed atrial fibrillation in the United States," *Clin. Cardiol.*, vol. 46, no. 5, pp. 484–493, May 2023.

[5] S. A. Lubitz, X. Yin, D. D. McManus, L.-C. Weng, H. J. Aparicio, A. J. Walkey, J. R. Romero, C. S. Kase, P. T. Ellinor, P. A. Wolf, S. Seshadri, and E. J. Benjamin, "Stroke as the initial manifestation of atrial fibrillation: The Framingham Heart Study," *Stroke*, vol. 48, no. 2, pp. 490–492, Feb. 2017.

[6] E. Y. Ding, G. M. Marcus, and D. D. Mcmanus, "Emerging technologies for identifying atrial fibrillation," *Circulat. Res.*, vol. 127, no. 1, pp. 128–142, Jun. 2020.

[7] S. Soon, H. Svavarsdottir, C. Downey, and D. G. Jayne, "Wearable devices for remote vital signs monitoring in the outpatient setting: An overview of the field," *BMJ Innov.*, vol. 6, no. 2, pp. 55–71, Jan. 2020.

[8] Z. Kalarus et al., "Searching for atrial fibrillation: Looking harder, looking longer, and in increasingly sophisticated ways. An EHRA position paper," *Europace*, vol. 25, no. 1, pp. 185–198, Feb. 2023.

[9] U. Satija, B. Ramkumar, and M. S. Manikandan, "A review of signal processing techniques for electrocardiogram signal quality assessment," *IEEE Rev. Biomed. Eng.*, vol. 11, pp. 36–52, 2018.

[10] K. van der Bijl, M. Elgendi, and C. Menon, "Automatic ECG quality assessment techniques: A systematic review," *Diagnostics*, vol. 12, no. 11, p. 2578, Oct. 2022.

[11] F. Liu, S. Xia, S. Wei, L. Chen, Y. Ren, X. Ren, Z. Xu, S. Ai, and C. Liu, "Wearable electrocardiogram quality assessment using wavelet scattering and LSTM," *Frontiers Physiol.*, vol. 13, Jun. 2022, Art. no. 905447.

[12] Á. H. Herraiz, A. Martínez-Rodrigo, V. Bertomeu-González, A. Quesada, J. J. Rieta, and R. Alcaraz, "A deep learning approach for featureless robust quality assessment of intermittent atrial fibrillation recordings from portable and wearable devices," *Entropy*, vol. 22, no. 7, p. 733, Jul. 2020.

[13] J. Behar, J. Oster, Q. Li, and G. D. Clifford, "ECG signal quality during arrhythmia and its application to false alarm reduction," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 6, pp. 1660–1666, Jun. 2013.

[14] D. Yoon, H. S. Lim, K. Jung, T. Y. Kim, and S. Lee, "Deep learning-based electrocardiogram signal noise detection and screening model," *Healthcare Inform. Res.*, vol. 25, no. 3, pp. 201–211, Jul. 2019.

[15] Q. Zhang, L. Fu, and L. Gu, "A cascaded convolutional neural network for assessing signal quality of dynamic ECG," *Comput. Math. Methods Med.*, vol. 2019, Oct. 2019, Art. no. 7095137.

[16] S. K. Bashar, E. Ding, A. J. Walkey, D. D. McManus, and K. H. Chon, "Noise detection in electrocardiogram signals for intensive care unit patients," *IEEE Access*, vol. 7, pp. 88357–88368, 2019.

[17] B. J. Drew, P. Harris, J. K. Zègre-Hemsey, T. Mammone, D. Schindler, R. Salas-Boni, Y. Bai, A. Tinoco, Q. Ding, and X. Hu, "Insights into the problem of alarm fatigue with physiologic monitor devices: A comprehensive observational study of consecutive intensive care unit patients," *PLoS One*, vol. 9, no. 10, Oct. 2014, Art. no. e110274.

[18] M. R. Afzal, J. Mease, T. Koppert, T. Okabe, J. Tyler, M. Houmsse, R. S. Augostini, R. Weiss, J. D. Hummel, S. J. Kalbfleisch, and E. G. Daoud, "Incidence of false-positive transmissions during remote rhythm monitoring with implantable loop recorders," *Heart Rhythm*, vol. 17, no. 1, pp. 75–80, Jan. 2020.

[19] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5455–5516, Apr. 2020.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[21] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, Apr. 2016.

[22] X. Han et al., "Pre-trained models: Past, present and future," *AI Open*, vol. 2, pp. 225–250, Jan. 2021.

[23] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 818–833.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[27] S. Rahman, C. Karmakar, I. Natgunanathan, J. Yearwood, and M. Palaniswami, "Robustness of electrocardiogram signal quality indices," *J. Roy. Soc. Interface*, vol. 19, no. 189, Apr. 2022, Art. no. 20220012.

[28] J. N. John, C. Galloway, and A. Valys, "Deep convolutional neural networks for noise detection in ECGs," 2018, *arXiv:1810.04122*.

[29] S. Luo and P. Johnston, "A review of electrocardiogram filtering," *J. Electrocardiol.*, vol. 43, no. 6, pp. 486–496, Nov./Dec. 2010.

[30] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, pp. 1–54, Dec. 2019.

[31] J. Ródenas, M. García, R. Alcaraz, and J. J. Rieta, "Combined nonlinear analysis of atrial and ventricular series for automated screening of atrial fibrillation," *Complexity*, vol. 2017, Oct. 2017, Art. no. 2163610.

[32] G. D. Clifford et al., "AF classification from a short single lead ECG recording: The PhysioNet/computing in cardiology challenge 2017," in *Proc. Comput. in Cardiol. (CinC)*, Rennes, France, 2017, pp. 1–4.

[33] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. 215–220, Jun. 2000.

[34] G. Liu, X. Han, L. Tian, W. Zhou, and H. Liu, "ECG quality assessment based on hand-crafted statistics and deep-learned S-transform spectrogram features," *Comput. Methods Programs Biomed.*, vol. 208, Sep. 2021, Art. no. 106269.

[35] Z. Zhao, C. Liu, Y. Li, Y. Li, J. Wang, B.-S. Lin, and J. Li, "Noise rejection for wearable ECGs using modified frequency slice wavelet transform and convolutional neural networks," *IEEE Access*, vol. 7, pp. 34060–34067, 2019.

[36] T. Li and M. Zhou, "ECG classification using wavelet packet entropy and random forests," *Entropy*, vol. 18, no. 8, p. 285, Aug. 2016.

[37] F. Demir, N. Sobahi, S. Siuly, and A. Sengur, "Exploring deep learning features for automatic classification of human emotion using EEG rhythms," *IEEE Sensors J.*, vol. 21, no. 13, pp. 14923–14930, Jul. 2021.

[38] J. Allen, H. Liu, S. Iqbal, D. Zheng, and G. Stansby, "Deep learning-based photoplethysmography classification for peripheral arterial disease detection: A proof-of-concept study," *Physiolog. Meas.*, vol. 42, no. 5, Jun. 2021, Art. no. 054002.

[39] W. Chen, Q. Sun, X. Chen, G. Xie, H. Wu, and C. Xu, "Deep learning methods for heart sounds classification: A systematic review," *Entropy*, vol. 23, no. 6, p. 667, May 2021.

[40] H. Khorrami and M. Moavenian, "A comparative study of DWT, CWT and DCT transformations in ECG arrhythmias classification," *Expert Syst. Appl.*, vol. 37, no. 8, pp. 5751–5757, Aug. 2010.

[41] P. S. Addison, *The Illustrated Wavelet Transform Handbook: Introductory Theory and Applications in Science, Engineering, Medicine and Finance*. Boca Raton, FL, USA: CRC Press, 2017.

[42] V. J. Bolós and R. Benítez, "The wavelet scalogram in the study of time series," in *Advances in Differential Equations and Applications*. Berlin, Germany: Springer, 2014.

[43] S. A. Singh and S. Majumder, "A novel approach OSA detection using single-lead ECG scalogram based on deep neural network," *J. Mech. Med. Biol.*, vol. 19, no. 4, Jun. 2019, Art. no. 1950026.

[44] Y.-H. Byeon, S.-B. Pan, and K.-C. Kwak, "Intelligent deep models based on scalograms of electrocardiogram signals for biometrics," *Sensors*, vol. 19, no. 4, p. 935, Feb. 2019.

[45] B. K. Pradhan, B. C. Neelappu, J. Sivaraman, D. Kim, and K. Pal, "A review on the applications of time-frequency methods in ECG analysis," *J. Healthcare Eng.*, vol. 2023, Feb. 2023, Art. no. 3145483.

[46] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[47] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50× fewer parameters and < 0.5 MB model size," 2016, *arXiv:1602.07360*.

[48] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.

[49] G. D. Clifford, J. Behar, Q. Li, and I. Rezek, "Signal quality indices and data fusion for determining clinical acceptability of electrocardiograms," *Physiolog. Meas.*, vol. 33, no. 9, pp. 1419–1433, Aug. 2012.

[50] Z. Zhao and Y. Zhang, "SQI quality evaluation mechanism of single-lead ECG signal based on simple heuristic fusion and fuzzy comprehensive evaluation," *Frontiers Physiol.*, vol. 9, p. 727, Jun. 2018.

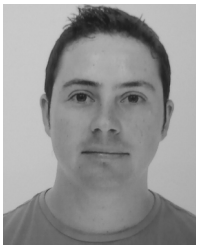
[51] A. Albaba, N. Simões-Capela, Y. Wang, R. C. Hendriks, W. De Raedt, and C. Van Hoof, "Assessing the signal quality of electrocardiograms from varied acquisition sources: A generic machine learning pipeline for model generation," *Comput. Biol. Med.*, vol. 130, Mar. 2021, Art. no. 104164.

[52] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement," *Circulations*, vol. 131, no. 2, pp. 211–219, 2015.

[53] M. A. Little, G. Varoquaux, S. Saeb, L. Lonini, A. Jayaraman, D. C. Mohr, and K. P. Kording, "Using and understanding cross-validation strategies. Perspectives on Saeb et al." *GigaScience*, vol. 6, no. 5, pp. 1–6, Mar. 2017.

[54] S. Saeb, L. Lonini, A. Jayaraman, D. C. Mohr, and K. P. Kording, "The need to approximate the use-case in clinical machine learning," *GigaScience*, vol. 6, no. 5, pp. 1–9, Mar. 2017.

[55] Q. Dong and G. Luo, "Progress indication for deep learning model training: A feasibility demonstration," *IEEE Access*, vol. 8, pp. 79811–79843, 2020.
[56] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F_1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, Jan. 2020.
[57] M. Bahrami and M. Forouzanfar, "Sleep apnea detection from single-lead ECG: A comprehensive analysis of machine learning and deep learning algorithms," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022.
[58] K. Dahal and M. H. Ali, "Overview of machine learning and deep learning approaches for detecting shockable rhythms in AED in the absence or presence of CPR," *Electronics*, vol. 11, no. 21, p. 3593, Nov. 2022.
[59] M. B. Abubaker and B. Babayigit, "Detection of cardiovascular diseases in ECG images using machine learning and deep learning methods," *IEEE Trans. Artif. Intell.*, vol. 4, no. 2, pp. 373–382, Apr. 2023.
[60] B. Pourbabaee, M. J. Roshtkhari, and K. Khorasani, "Deep convolutional neural networks and learning ECG features for screening paroxysmal atrial fibrillation patients," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 48, no. 12, pp. 2095–2104, Dec. 2018.
[61] Z. Ebrahimi, M. Loni, M. Daneshtalab, and A. Gharehbaghi, "A review on deep learning methods for ECG arrhythmia classification," *Expert Syst. Appl., X*, vol. 7, Sep. 2020, Art. no. 100033.
[62] G. Manis, M. Aktaruzzaman, and R. Sassi, "Low computational cost for sample entropy," *Entropy*, vol. 20, no. 1, p. 61, Jan. 2018.
[63] M. J. H. Pantho, P. Bhowmik, and C. Bobda, "Towards an efficient CNN inference architecture enabling in-sensor processing," *Sensors*, vol. 21, no. 6, p. 1955, Mar. 2021.
[64] J. A. Rincon, S. Guerra-Ojeda, C. Carrascosa, and V. Julian, "An IoT and fog computing-based monitoring system for cardiovascular patients with automatic ECG classification using deep neural networks," *Sensors*, vol. 20, no. 24, p. 7353, Dec. 2020.
[65] Y. Zhang, S. Wei, Y. Long, and C. Liu, "Performance analysis of multiscale entropy for the assessment of ECG signal quality," *J. Electr. Comput. Eng.*, vol. 2015, Apr. 2015, Art. no. 563915.
[66] A. Maracani, V. P. Pastore, L. Natale, L. Rosasco, and F. Odone, "In-domain versus out-of-domain transfer learning in plankton image classification," *Sci. Rep.*, vol. 13, no. 1, p. 10443, Jun. 2023.
[67] D. S. Terzi and N. Azginoglu, "In-domain transfer learning strategy for tumor detection on brain MRI," *Diagnostics*, vol. 13, no. 12, p. 2110, Jun. 2023.
[68] J. Xie, L. Peng, L. Wei, Y. Gong, F. Zuo, J. Wang, C. Yin, and Y. Li, "A signal quality assessment-based ECG waveform delineation method used for wearable monitoring systems," *Med. Biol. Eng. Comput.*, vol. 59, no. 10, pp. 2073–2084, Oct. 2021.
[69] G. Hirsch, S. H. Jensen, E. S. Poulsen, and S. Puthusserypady, "Atrial fibrillation detection using heart rate variability and atrial activity: A hybrid approach," *Expert Syst. Appl.*, vol. 169, May 2021, Art. no. 114452.



ÁLVARO HUERTA received the master's degree in telecommunications engineering from the University of Castilla–La Mancha, Cuenca, Spain, in 2019, where he is currently pursuing the Ph.D. degree in information technology advanced. His main research interest relies on biomedical signal processing, specially electrocardiograms and employing deep learning and machine learning techniques.

ARTURO MARTINEZ-RODRIGO received the M.Sc. degree in telecommunications engineering and the Ph.D. degree in medical care research from the University of Castilla–La Mancha (UCLM), Cuenca, Spain, in 2010 and 2013, respectively. He was an Assistant Professor with the Department of Mathematics, UCLM, in 2010, where he is currently with the Department of Computer Science. His research interests include signal processing, sensor networks, and applied AI.

DAVIDE CARNEIRO received the Ph.D. degree from the Joint Doctoral Program in Computer Science of three top Portuguese universities (MAP-i Program–Minho, Aveiro, and Porto). He is an Adjunct Professor with the School of Management and Technology, Polytechnic Institute of Porto, and a Researcher with INESC TEC. His research interests are on the domain of artificial intelligence and addressing topics, such as distributed and streaming machine learning, meta-learning, and AI ethics.

VICENTE BERTOMEU-GONZÁLEZ is a Cardiologist at Hospital Clínica Benidorm, Alicante, Spain. He is also an Associate Professor with the Department of Clinical Medicine, Universidad Miguel Hernández. His main lines of research are epidemiology of atrial fibrillation in Spain, diagnostic and prognostic markers in acute heart failure, hemorrhagic and thrombotic complications in cardiac electrophysiology interventions in patients, with atrial fibrillation and ablation with minimal scope.

JOSE J. RIETA (Member, IEEE) received the Ph.D. degree in biomedical signal processing from Universitat Politecnica de Valencia (UPV), Valencia, Spain, in 2003. He is a Full Professor with the Electronic Engineering Department, UPV. He currently coordinates the Biomedical Synergy Research Group, UPV. His research interests include biomedical signal processing, blind signal separation, engineering education, arrhythmias, and educational data mining.

RAÚL ALCARAZ received the Ph.D. degree in electronic engineering from Universitat Politecnica de Valencia, Valencia, Spain, in 2008. He is an Associate Professor with the Department of Electrical, Electronic, Automatic Control and Communications, University of Castilla–La Mancha, Cuenca, Spain. His research interests include biomedical signal processing, engineering education, and educational data mining.