Device-Free People Counting in IoT Environments: New Insights, Results, and Open Challenges

Iker Sobron, *Member, IEEE*, Javier Del Ser, *Senior Member, IEEE*, Iñaki Eizmendi, *Member, IEEE*, and Manuel Vélez, *Member, IEEE*

***Abstract*—In the last years multiple Internet of Things (IoT) solutions have been developed to detect, track, count, and identify human activity from people that do not carry any device nor participate actively in the detection process. When WiFi radio receivers are employed as sensors for device-free human activity recognition, channel quality measurements are preprocessed in order to extract predictive features toward performing the desired activity recognition via machine learning (ML) models. Despite the variety of predictors in the literature, there is no universally outperforming set of features for all scenarios and applications. However, certain feature combinations could achieve a better average detection performance compared to the use of a thorough feature portfolio. Such predictors are often obtained by feature engineering and selection techniques applied before the learning process. This manuscript elaborates on the feature engineering and selection methodology for counting device-free people by solely resorting to the fluctuation and variation of WiFi signals exchanged by IoT devices. We comprehensively review the feature engineering and ML models employed in the literature from a critical perspective, identifying trends, research niches, and open challenges. Furthermore, we present and provide the community with a new open database with WiFi measurements in several indoor environments (i.e., rooms, corridors, and stairs) where up to five people can be detected. This dataset is used to exhaustively assess the performance of different ML models with and without feature selection, from which insightful conclusions are drawn regarding the predictive potential of different predictors across scenarios of diverse characteristics.**

***Index Terms*—Device-free people counting, feature selection (FS), Internet of Things (IoT) architectures, machine learning (ML).**

I. Introduction

THE UBIQUITY of wireless technologies has lately increased toward almost any location and physical scenario, mainly ignited by the ever-decreasing hardware costs of radio interfaces, higher chip integrability, and enhanced efficiency of the underlying protocols at different communication layers. This increased multiple connectivity has been specially notable in Internet of Things (IoT) environments, where the heterogeneity of devices and standards coexisting in the same physical space has propelled the massive incorporation of wireless networking capabilities to IoT devices [1], [2]. This has unleashed a great potential for the research and development of new functionalities and services in IoT systems, far beyond the plain data exchange functionality provided by the establishment of wireless links among IoT nodes [3], [4].

Of particular interest for the scope of this manuscript is the noninvasive detection of human movements through the analysis of the propagation effects of radio-frequency signals. This paradigm have lately gained a notable momentum within the research community due to their straightforward use in IoT applications, such as intrusion detection and tracking [5], [6], vital signs monitoring [7], [8], elderly or children in care homes [9], [10], and emotion identification [11], among many others [12]. As such, one person in motion traversing a transmitter–receiver wireless communication link shall lead to perturbations on the propagated signal in terms of refractions, reflections, or fading, which can render noticeable fluctuations in the received signal. Measurements of the channel quality information at the receiver can be exploited and analyzed for localization, tracking, counting, or gesture recognition of people around the wireless communication link.

Since the pioneering work in [13], where the concept of device-free passive localization was introduced and validated with WiFi equipment, many solutions have been thereafter developed in order to infer human activity from people lacking any wireless device and acting as a passive stakeholder in the scenario at hand (see [14]–[16] and references therein). Most of these works have resorted to predictive variables (namely, predictors or features) related to the channel quality, which can be quantified in terms of the received signal strength [e.g., received signal strength indicator (RSSI)] or channel-related information captured at the physical layer [channel state information (CSI)]. RSSI generally considers the overall signal power regardless of the subcarriers, and is expressed as an averaged relative magnitude of signal amplitude or power. In contrast, CSI captures the shape of the received signal or channel response in frequency domain.

This noted difference between metrics belonging to both domains can give some insights about the suitability of each alternative for a specific human activity recognition purpose.

Manuscript received November 29, 2017; revised January 24, 2018; accepted February 8, 2018. Date of publication February 16, 2018; date of current version January 16, 2019. This work was supported in part by the Spanish Ministry of Economy and Competitiveness (TEC2015-66153-P MINECO/FEDER, EU) under Project 5GnewBROS and in part by the Basque Government (IT683-13 and the EMAITEK Program). *(Corresponding author: Iker Sobron.)*

I. Sobron, I. Eizmendi, and M. Vélez are with the Department of Communications Engineering, University of the Basque Country (UPV/EHU), 48013 Bilbao, Spain (e-mail: iker.sobron@ehu.eus; inaki.eizmendi@ehu.eus; manuel.velez@ehu.eus).

J. Del Ser is with the Department of Communications Engineering, University of the Basque Country (UPV/EHU), 48013 Bilbao, Spain, with TECNALIA, 48370 Derio, Spain, and also with the Basque Center for Applied Mathematics, 48009 Bilbao, Spain (e-mail: javier.delser@ehu.eus).

Digital Object Identifier 10.1109/JIOT.2018.2806990

For instance, RSSI measures the attenuation of the channel averaged over all subcarriers and as a consequence, is unable to describe the multipath effect. In the case of CSI, the metric is frequency selective and can provide information about the multipath propagation. In this sense, CSI can give better information of the channel conditions in those IoT scenarios where the human presence significantly modifies the multipath propagation environment. As a result, CSI measurements are usually employed in activity recognition and crowd counting due to the fact that multipath is the major perturbation agent in the channel conditions. Nonetheless, RSSI can also achieve good performance in small distances, i.e., when the channel between transmitter and receiver is dominated by the attenuation caused by the scatterer (human body) [17].

When WiFi radio receivers are employed as IoT sensors for device-free human activity recognition, channel quality measurements (either RSSI or CSI) are preprocessed in terms of time and frequency statistics, correlation estimators, Doppler spectrum computation, and other mathematical manipulations. This preprocessing step yields features that can be fed to discriminative machine learning (ML) models. This preprocessing can be performed at several layers of a generic IoT architecture, namely, on the edge (i.e., at the very sensor at which raw wireless signals are captured) or in the Cloud (correspondingly, in a remote server with higher computation capabilities leveraging the high bandwidths allowed by the latest wireless standards). All in all, preprocessing and feature extraction are crucial steps when counting device-free people by using ML models, having garnered most of the research activity in this field during the last decade [10].

Despite the upsurge of contributions dealing with this topic, there is still no consensus on which features prevail over others when used to count people in an IoT environment. This issue is not only exclusive for this application itself: it is known that most predictive models are specially sensitive to collinearity between features, and that a high number of noninformative predictors could lead to overfitted models by capturing irrelevant patterns over the training data. Therefore, a proper choice of features, along with a principled selection and fine-tuning of a learning algorithm capitalizing on such feature subset, are of paramount importance in order to achieve a good accuracy when counting people in a certain scenario. However, to the knowledge of the authors the literature lacks a thorough analysis on how different predictors and models behave across IoT scenarios of diverse propagation characteristics. As a result, some questions remain unanswered: under which criterion should a given model be selected over another? Which features appear to be more predictive than others when exploited in different IoT environments?

This paper aims at thoroughly addressing the above questions from manifold interrelated perspectives.

1) We present a comprehensive, critical review of different features and ML models employed for human activity recognition in IoT environments, emphasizing on those gravitating on crowd counting and tracking problems.
2) We present and describe a new open dataset with wireless signals captured over diverse physical IoT environments.
3) We exhaustively explore different features and ML models over these traces so as to draw empirically informed conclusions about their performance and predictive potential for other IoT scenarios and cases.
4) We outline technical trends and open challenges to which future research efforts should be targeted in this field.

As will be clearly shown by the presented results, features proposed in the literature offer varying predictive potential over different physical scenarios, yet a subset of them is found to remain significantly informative universally across all considered setups.

The rest of this paper is structured as follows. Section II presents an overview of the state of the art in device-free people counting, whereas Section III outlines the architecture of a reference IoT framework along with design guidelines to implement and deploy a people counting service at different levels. Next, details of the newly proposed dataset for counting device-free people are given in Section IV, followed by a discussion on the experimental results in Section V. Finally, Section VI ends the paper by summarizing the conclusions drawn from the performed experiments and motivating open research lines related to this field.

II. Device-Free People Counting and Activity Recognition: Introduction and Review

As aforementioned in the introduction, the literature has been notably rich in the last years in what relates to technical approaches for counting crowds by solely relying on the study of the fluctuations of wireless signals, as opposed to classical systems based on dedicated devices. In this section an overview of the different signal features proposed by the community is done, complemented by a prior introduction on how such predictors can be extracted by processing the CSI captured from the received signal.

A. Channel State Information As Detection Measurement

In typical indoor scenarios, a transmitted signal propagates to the receiver through reflection, scattering, and attenuation. As a result, each multipath signal component is characterized by a different time delay, amplitude attenuation, and phase shift. Therefore, the channel impulse response of the received signal can be expressed as

$$h(t) = \sum_{i=1}^{N} \alpha_i(t) e^{-j\phi_i(t)} \delta(\tau - \tau_i(t)) \tag{1}$$

where N denotes the number of multipath signal components, and $\alpha_i(t)$, $\phi_i(t)$, and $\tau_i(t)$ represent the amplitude, phase, and time delay of the ith multipath component, respectively. The fast Fourier transform (FFT) of $h(t)$ as per (1) corresponds to the so-called *channel frequency response* or PHY layer CSI, which can be in general obtained easily by means of channel estimation processes in orthogonal frequency division multiplexing (OFDM) systems. Most WiFi standards, such as IEEE 802.11a/g/n/ac utilize OFDM waveforms. As a result, one of the first steps in the reception process is the channel estimation for equalization. Thus, the complex CSI of an

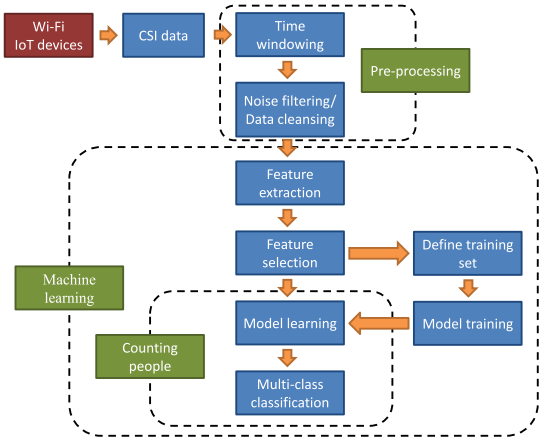

Fig. 1. Processing flow followed a free-device people counting WiFi device.

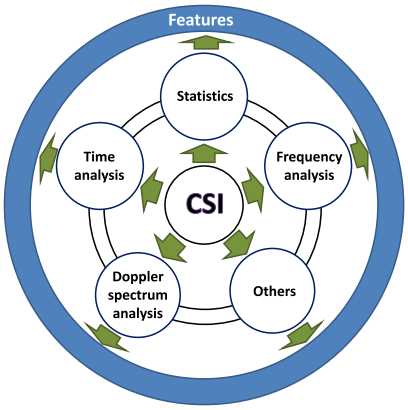

Fig. 2. Domains on which the feature engineering process for device-free people counting can be performed.

OFDM symbol with K subcarriers sampled at time slot n is given by

$$\mathbf{H}(n) = [H_1(n), \ldots, H_K(n)] \tag{2}$$

where $H_i(n) \doteq |H_i(n)|e^{j\angle H_i(n)}$ with $|H_i(n)|$ and $\angle H_i(n)$ being the amplitude and phase of the CSI at the ith subcarrier, respectively.

The effects of human presence shall modify the channel response at symbol time-varying amplitude, phase and time delay of the N paths in (1), thus leading to temporal fluctuations in the amplitude and phase of the CSI samples in (2). Based on such variations along the time domain different features can be furnished and fed to an ML model for estimating the current number of people in the scenario under study. Such predictors can be engineered in a very diverse fashion within a general processing flow, which is next explained in detail.

B. CSI-Based Feature Engineering and Learning Models

The combination of the feature engineering process with the learning model comprises a smart detection system that can be embedded in WiFi IoT end-points, as shown in Fig. 1. In general, the system architecture consists of two stages: 1) data preprocessing and 2) learning algorithm. The former comprises noise filtering, data cleansing, and feature engineering, whereas the latter implies feature selection (FS) (if any), model training/validation and testing. As shown in Fig. 1, the processing flow in a WiFi IoT sensing architecture closely resembles the classical methodology for an ML problem. However, considering the underlying physical meaning of CSI in commodity WiFi devices, several manipulations are necessary to transform the CSI data, traditionally employed for channel equalization, into valuable information for the smart sensing application. In the first stage, CSI data are gathered and segmented in time windows. After that, customized noise filtering methods, such as low-pass filters or wavelets are usually applied to remove unwanted components. Following the data preprocessing stage, data cleansing and feature extraction are performed over time-windowed CSI data.

Feature extraction can be performed through the analysis of the CSI data in several domains depicted in Fig. 2. When people traverse the monitored area, CSI data captured by the WiFi IoT sensors fluctuate according to the newly appearing scattering elements (i.e., people performing some actions) that modify the multipath conditions of the radio environment. Those variations can be analyzed in time, frequency, and Doppler spectrum domains, using to this end different statistics and other mathematical tools. Features extracted from CSI and RSSI measurements have been most widely employed in the human activity recognition, i.e., human presence detection, human activity detection, and crowd counting. However, recent results have demonstrated that CSI-related features generally provide more stable information on the activity to be recognized than RSSI-based features due to the frequency selectivity of the measurements and less sensitiveness to specific environments [17]. Despite the fine-grained sensitiveness of CSI information, it is worth mentioning the good results scored by RSSI-based methods under certain conditions, such as small-sized rooms, where people is forced to move relatively close to transceiver or receiver and the attenuation of the line-of-sight is the dominant effect when compared to multipath [6], [18].

In the second stage of the crowd counting architecture, selected features are fed into a multiclass ML model that predicts the number of people in the monitoring area based on the patterns learned from a supervised dataset. At this point it should be remarked that unsupervised schemes have been also proposed for human activity recognition. Nonetheless, such approaches require additional mechanisms to previously tune some of their parameters according to the monitored scenario.

C. State of the Art in Device-Free People Counting

In the following, we present a comprehensive review of the feature engineering and ML models portfolio employed in human presence recognition with CSI-based descriptors, which is summarized in Tables I and II.

Since the presence of moving people shall induce temporal fluctuations of the received WiFi signals, time correlation among successive CSI measurements has been considered in several contributions in order to exploit such a temporal disturbance. The design of power-independent features is managed with the computation of the correlation matrix eigenvalues for each time window. In this context, Xiao *et al.* [19] proposed a

CSI feature extraction model in order to detect the presence of people based on the first and second eigenvalues of the Pearson correlation matrix of the CSI amplitudes $|H_i(n)|$. The presence detection is conducted with outliers identification from normal features using density-based clustering.

In contrast, Qian *et al.* [20] proposed to extract features from the respective covariance matrix of normalized CSI amplitude and phase of several sequential measurements over a certain time window. They compute the eigenvalues of both matrices and select the maximum and second maximum eigenvalue of each matrix in order to form a four-tuple of features for device-free passive detection of moving humans with dynamic speed. A support vector machine (SVM) classifier is used for detection outperforming [19]. In [16], they considered both amplitude and phase to compute the maximum eigenvalues of the Pearson correlation matrices since, besides the amplitude of CSI values which is generally the main feature to be analyzed in the detection process, CSI phase information in $\angle H_i(n)$ is also sensitive to environmental changes, thus providing a more accurate detection. An SVM ML model is employed for human moving detection. Also related with time correlation strategies, a principal component analysis (PCA) was applied to CSI information in [21] to extract a target-dependent feature set suitable to enable the robust detection of passive targets (i.e., presence of people) in large indoor areas. Detection is performed through a simple threshold imposed on the values of the principal components.

Time statistics have been widely employed for feature construction in many application areas. Common statistical metrics, such as mean, variance, moments and the like are combined to yield multiple predictors. Following this engineering approach, Gong *et al.* [22] defined a new feature, coined as the *coefficient of variation of phase*, where the ratio between the standard deviation and mean of the ith CSI subcarrier phase $\angle H_i(n)$ is computed within a time window. Human motion is detected when the averaged ratio of the coefficient of variation of phase falls within a predefined confidence interval. In [23], the coefficient of variation of the normalized CSI amplitudes is computed following the same criterion as in [22]. In order to quantify the multipath propagation conditions due to the presence of an intruder, a new metric is derived. This feature consists of the ratio between the kurtosis and the mean of the coefficient of variation of CSI amplitudes. A precalibrated detection threshold is necessary for each scenario. Additionally, in [23] the Rician K-factor was postulated as a possible predictor, however, it was neglected due to the lack of time and phase synchronization on commodity WiFi devices in order to extract accurate information for the Rician estimator. In [24], an speed independent device-free entity detection is proposed for intrusion detection. The presented system captures the variance of variances of CSI amplitudes of each subcarrier as features and uses hidden Markov model to model human motion of different speeds and detect human presence slightly influenced by moving speed. Under this scheme, they transfer the human detection problem into a probability problem. In [25], people counting was performed by formulating a stable monotonic function that characterizes the relationship between the crowd number and variance-based CSI features. To this end, a new metric based on the percentage of nonzero elements in a dilated CSI matrix and the Grey Verhulst model was designed and assessed.

Time and spectral analyses have been employed in [26]–[28] for FS. The work presented in [27] describes a series of features of CSI amplitudes in several domains. In time domain mean, absolute-mean, variance, skewness, kurtosis, range, mean-crossing-rate, max, min, median, and quartiles of CSI amplitude values are computed over every time window, whereas in the frequency domain, normalized entropy, normalized energy, the largest FFT peak and dominant-frequency-ratio are extracted from the normalized OFDM symbol. Two additional features, denoted as DC-mean and DC-area, are computed as the mean and the sum of all amplitudes after a low-pass filter with a cutoff frequency of 1 Hz. Decision tree (DT) and logistic regression (LR) are the ML models adopted for human behavior classification. Based on [27], Zhang *et al.* [26] constructed ten features for counting people; in time domain: mean, max, min, skewness, kurtosis, variance, and mean crossing rate of the CSI amplitudes in a time window. In frequency domain, normalized entropy, normalized energy, and the largest FFT peaks of each OFDM symbol. FS is performed to yield a feature subset fed to a nearest neighbor (NN) and nearest subspace classifier based on a sparse representation [29]. In [28], a device-free identity identification approach is presented based on CSI of WiFi signals. After a data cleansing preprocessing with PCA and low pass filter, features extraction is performed in both time and frequency domain, selecting the five most effective features out of those shown in Table I according to information gain. Based on the selected features, stranger recognition through Gaussian mixture model and identity identification through an SVM with radial basis function kernel are employed.

More recently, both RSSI and CSI descriptors are analyzed in [17] for crowd counting. CSI-based features consist of the mean and standard deviation of the normalized spectral Euclidean distances between the CSI amplitudes of two consecutive WiFi beacon messages. The authors remarked that the CSI-related approach is less sensitive to the specific environment with respect to the performance of an RSSI-related approach. A linear discriminant analysis (LDA, also referred to as Fisher classifier) was selected as the ML model for estimating the number of people in the scenario under analysis. In this same line of research, [30] uses as a basic descriptor the Euclidean distance between two CSI vectors combined with a proper choice of the normalization which can reduce the dependence from the background environment. Normalizations based on the mean and energy of the CSI OFDM symbol have been considered in order to linearly combine the Euclidean distances of several channels. The so-called Davies–Bouldin index is computed for FS and an LDA classifier is finally used for crowd counting. In [10], CSI features are used for distinguishing among walk, fall, sit, and stand actions. In this case, normalized standard deviation, offset of signal strength, period of motion, median absolute deviation, interquartile range, signal entropy, and velocity of signal change are employed over CSI amplitudes, two learning models—a one-class SVM and a random forest (RF) classifier—are compared to each other

TABLE I
LITERATURE ON FEATURE ENGINEERING FOR DEVICE-FREE PEOPLE COUNTING CLASSIFIED BY DOMAIN IN WHICH FEATURES ARE COMPUTED (A [✓] INDICATES THAT THE LABELED FEATURE IS USED IN THE SIMULATIONS PRESENTED IN THIS PAPER)

Domain	**Features**	**CSI used in every reference**	
		Amplitude	**Phase**
Time correlation	Eigenvalues	[16],[19],[20],[21],[✓]	[16],[20],[21],[✓]
	Average	[23]	–
Time statistics	Mean	[17],[23],[26],[27],[28],[30],[✓]	[17],[22],[28],[30]
	Std deviation	[10],[23],[28],[30]	[22],[28],[30]
	Variance	[24],[25],[26],[27],[30],[✓]	[30]
	Median	[10],[27],[28]	[28]
	Skewness	[26],[27],[28],[✓]	[28]
	Kurtosis	[23],[26],[27],[28],[✓]	[28]
	Mean-crossing rate	[26],[27]	–
	Inter-quartile range	[10],[27],[✓]	–
Probability Density Functions	Earth Mover's distance	[31],[32]	–
	Kullback-Leibler divergence	[33]	–
Doppler spectrum	Mean	[34],[✓]	[34],[✓]
	Std deviation	[34],[✓]	[34],[✓]
	Central moments	[34],[✓]	[34],[✓]
	Skewness	[34],[✓]	[34],[✓]
	Kurtosis	[34],[✓]	[34],[✓]
	Decay factor	[34]	[34]
	Spectral spread	[34]	[34]
	Largest peak	[26],[27]	–
Frequency analysis	Energy	[26],[27],[28],[30],[34],[✓]	–
	Max	[10],[26],[27],[28]	–
	Min	[26],[27],[28]	–
Others	Rician K-factor	[23]	–
	Entropy	[10],[26],[27],[28],[34]	[28],[34]
	1st derivative	[10],[34]	[34]
	Flatness	[34]	[34]
	Roll-off factor	[34]	[34]
	Euclidean distance	[17],[30]	[17],[30]
	Average step time	[28]	[28]
	Variance step time	[28]	[28]

in four locations with three different people (not simultaneously) performing the aforementioned actions. RF is shown to achieve slightly higher accuracy scores than SVM.

From another statistical perspective, probability distribution functions are also exploited in the detection process. In [31], histograms of time-windowed CSI amplitudes at each subcarrier were calculated. Earth mover's distance (EMD) among different histograms were subsequently computed at each subcarrier and eventually summed to yield a single, aggregate EMD. The final summation is compared to a predefined threshold to detect the presence of an intruder in the monitoring area. Likewise, Liu *et al.* [32] performed a time analysis through histograms of the variance of the CSI amplitudes along fixed time windows. In this paper, an off-line database stores the histogram fingerprint with no people in the monitoring area and by comparing with the histogram computed online, one can determine whether there is any person moving over the area. Again, EMD was used for comparing both histograms. From a theoretical perspective, the work presented in [33] performs an occupancy estimation model based on the analytical derivation of the probability distribution function of the time-domain received signal amplitude as a function of the number of occupants describing a motion model within a predefined squared workspace. Kullback–Leibler divergence is employed for the estimation of the number of people.

In [34], a Doppler spectral analysis is carried out over the CSI magnitudes for people counting. From the absolute values of 512-FFT-sized Doppler spectrum, a series of features is computed using statistics, information theory and mathematical manipulations as shown in Table I. After FS, the linear combination of the Doppler spectrum kurtosis in the MIMO channel of IEEE 802.11n is selected as the input to a Gaussian Naive Bayes (GNB) classifier.

TABLE II
ML MODELS UTILIZED IN THE LITERATURE

Supervised	Classification	Support Vector Machines (SVM)	[10],[16],[20],[24],[✓]
		Random Forest (RF)	[10],[27],[✓]
		Logistic Regression	[27],[✓]
		k Nearest Neighbor (kNN)	[26],[✓]
		Linear Discriminant Analysis (LDA)	[30],[✓]
		Naive Bayes (NB)	[34],[✓]
		Gaussian Mixture Model (GMM)	[28]
Unsupervised	Clustering	Density based (DBSCAN)	[19]
	Detection	Confidence Interval	[22]
		Threshold	[20],[21],[23]
		Hidden Markov Model	[24]
	Fitting	Earth Mover's Distance	[31],[32]
		Kullback-Leibler Divergence	[33]
		Grey-Verhulst Model	[25]

D. Research Niches

Upon the literature reviewed above, the reader may note that most contributions related to people counting are restricted to the definition and validation of new predictors, yet very

little attention is paid to the ML model to which such features are fed. As evinced by Table II, every work within this field resorts to a very narrow spectrum of ML models, disregarding whether, e.g., an eventually bad choice of classifier might degrade the predictive potential of the derived feature portfolio. A flurry of research activity on ML models has been caused by the momentum gained by hot concepts and paradigms, such as Big Data [35]. Contributions dealing with device-free people counting via ML models should also account for the variety and diversity of existing ML models so as to attain a comprehensive, well-rounded insight on the performance of device-free people counting models based on this family of learning algorithms. This is indeed one of the purposes of the experimental benchmark later discussed in this paper. Specifically we will focus on supervised learning models in light of their more extensive adoption for *counting* people than their unsupervised counterparts, which have instead been utilized for inferring the *absence/presence* of moving people.

Another issue worth to be highlighted in regards to Table I is the lack of a clear understanding about the potential of every newly proposed feature set when they are jointly used with existing predictors. Statistical interactions between features are a key aspect for the performance of supervised learning classifiers, which motivate the adoption of FS techniques [36]. However, most of the contributions reviewed above do not assess the predictive potential of their proposed features by accounting them together with other predictors from the literature, which might make their informativeness vary by virtue of their mutual statistical relationships with respect to the target variable (i.e., the number of people). This crucial aspect lies also at the core rationale of the simulation setup presented in Section V.

III. IoT Framework for Device-Free People Counting

Before proceeding with the experimental part of this paper we delve into the design of a device-free people counting IoT framework, which can be regarded as a cyber-physical system in the IoT context. Apart from the regular application of the WiFi systems, IoT nodes can act as sensors by collecting data about the surrounding radio environment. The sensed CSI, necessary for channel decoding in the wireless communication chain, can be converted into meaningful information about the human activity in a monitored area. WiFi IoT devices are usually connected to a local network where one or several routers grant connectivity to Internet. As a result, sensing data can be shared, gathered, and analyzed in a cloud platform in order to provide ubiquitous device-free people counting services over IoT deployments.

Additionally, a people-counting IoT scheme can achieve different degrees of intelligence and self-resilience depending on the computing model adopted in the implemented design. That is, a centralized computation model where ML models and decision making strategies are performed in the Cloud allows for a higher combination and coordination between physical and computational elements (i.e., a cyber-physical system) when compared with a distributed IoT scheme, where the Cloud can be seen as a visualization platform with the info-graphic of the monitored process. In this sense, Edge and Cloud computing are the most common topologies for data processing in IoT systems [37]. Unlike Cloud computing where almost all data processing tasks are centralized to alleviate the computational load at IoT nodes, Edge computing enables data processing to be performed closer to the data source, hence reducing the communication overhead between IoT nodes and the data center. Built on this identified trade-off and in connection to Fig. 3, we next discuss on how these computation models can be used to deploy a device-free crowd counting service within an IoT environment.

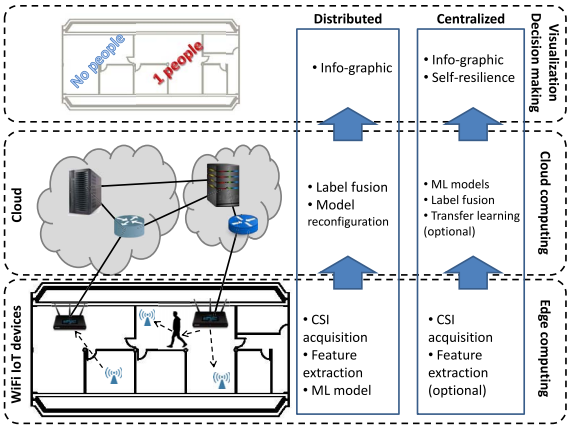

Fig. 3. Schematic of the IoT architecture on which a device-free counting functionality can be deployed.

1) *Edge Computing:* As has been mentioned previously, this architectural scheme is usually adopted for decreasing the volume of data transmitted from the IoT nodes to the Cloud. However, due to the fact that nodes can be connected to the Cloud via wired links, the adoption of Edge Computing is not motivated here by data traffic limitations, but rather by the required computational effort and by possible latencies that data processing can yield, which might be of relevance for certain time-critical applications demanding the counting service (e.g., intrusion detection). As a result, raw CSI data can be processed at each end device such that some elaborated features or testing decisions are eventually transmitted to the Cloud for further processing and/or storage. In this regard two computation levels are proposed: a) feature extraction and selection can be performed at each IoT node, so that the selected feature set is transmitted and fed to the ML model in the Cloud, whose (re-)training algorithm leverages a higher amount of computational resources; b) both FS and learning/testing are performed at the IoT devices, after which predictions are delivered to the Cloud and served therefrom to applications demanding the people counting service.
2) *Cloud Computing:* When raw CSI data are acquired at each WiFi IoT end-point and sent to the Cloud,

feature extraction and selection is performed in a centralized manner and specific trained models are built and used for classification. This scheme would also allow for new inductive strategies for knowledge transfer among different environments, such as transfer learning [38]. Knowledge extracted from different radio scenarios would ease the creation of a global classification model that would permit a tether-free human activity recognition scheme. When information about final decisions are sent from the WiFi IoT devices, shared information can be contrasted taking into account positional information about the monitored areas and can give a global viewpoint of the collected data.

According to [39], a device-free people counting IoT system can follow a cyber-physical architecture with attributes at the following levels.

1) *Smart Connection Level:* CSI data is acquired by WiFi IoT devices, which act as radio sensors.
2) *Data-to-Information Conversion Level:* CSI is transformed into features and ultimately, in an estimated number of people by virtue of predictive ML models.
3) *Cyber Level:* Information (CSI, features or number of people) is gathered and shared through the Cloud. Additionally, specific learning and validation models can be used to extract additional information that provide a global insight over the status of individual WiFi end-points.
4) *Cognition Level:* At this level, the presentation of the acquired knowledge is given to expert users in order to take proper decisions. For instance, the detection of human presence in several rooms of a flat when only one person is monitored can be self-corrected by the system.
5) *Configuration Level:* This level depends on the selected computation scheme. When Edge computing is employed, feedback from the cyber space (Cloud) to the physical space (WiFi IoT end-devices) is necessary in order to apply corrective decisions into the ML models.

IV. EHUCOUNT: New Dataset for Device-Free People Counting

In order to address the research questions posed previously, a measurement campaign capturing actual WiFi signal measurements was carried out in different indoor scenarios (Fig. 4) over facilities of the Faculty of Engineering of the University of the Basque Country (Bilbao, Spain). A portable testbench comprising a Anritsu MG3700A vector signal generator was used to transmit a prerecorded IEEE 802.11n trace signal during 15 s over the 2.4 GHz band. Reception was performed by a Anritsu MS2690A signal analyzer, which recorded data samples in IQ format at a sampling frequency of 25 Msamples per second for later off-line decoding. The decoding process produced the CSI values $\mathbf{H}(n)$ of the $K = 53$ subcarriers of the WiFi signal by means of the Agilent 89600B VSA software suite.

The produced dataset, coined as EHUCOUNT,[1] provides full CSI from all $K = 53$ subcarriers of the OFDM signal at the expense of an off-line processing. Nonetheless, this must be deemed irrelevant for the scope of the study: the assessment and comparison of the portfolio of features and supervised models utilized in the literature. As a matter of fact, on-line processing could be performed with a modified driver of the Intel NIC card 5300, which might provide the channel estimates of $K = 30$ subcarriers of an IEEE 802.11n MIMO signal in several commercial implementations [41]. The number of CSI traces per class and scenario ranges between 12 000 and 15 000, depending on synchronization issues at the signal decoding process.

[1]The EHUCOUNT dataset has been made public to the community in [40].

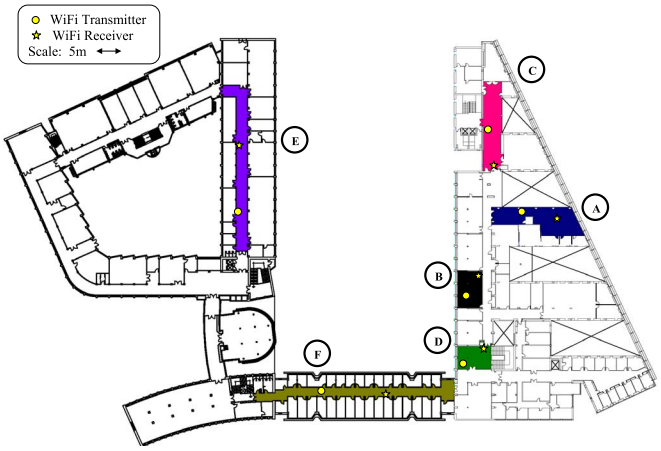

Fig. 4. Experimental scenarios (A to F) and testbed deployments embedded in the EHUCOUNT dataset presented in this paper.

TABLE III
DESCRIPTION AND PARAMETERS OF THE SCENARIOS IN EHUCOUNT

Scenario	People (#)	P_t (dBm)	f_C (GHz)
A (Office)	[0,1,2,3]	-10	2.462
B (Lab)	[0,1,2,3,4,5]	-10	2.462
C (Corridor)	[0,1,2,3,4]	0	2.437
D (Hall+Stairs)	[0,1,2,3,4,5]	0	2.437
E (Corridor)	[0,1,2,3,4]	2	2.422
F (Corridor)	[0,1,2,3,4,5]	1	2.452

Measurements have been carried out in six indoor scenarios where up to five people walked casually in the colored areas depicted in Fig. 4. In general, people tended to maintain the direction for a while before changing it according to the scenario topology. That is, one direction is maintained for a longer distance in corridors, such as scenarios C, E, and F. In contrast, people roamed around the rooms in scenarios A, B, and D. In all cases, people were instructed to walk by the area of interest at a speed in the typical range for indoor walking dynamics (i.e., $\leq$3 km/h). Details about the number of people that took part in the measurements at each scenario and some setup parameters are given in Table III, where P_t and f_C denote transmission power and carrier frequency, respectively.

V. Experiments and Discussion

As aforementioned in the introduction and concluded from the survey in Section II, it is difficult to ensure the existence of specific features performing better than others for all applications and scenarios. Nonetheless, there might be some of them that occur to be more frequently selected—in average terms—than others after FS across different scenarios. For that

TABLE IV
FEATURES FOR THE ML MODEL CONSTRUCTION

Features	Indexes
Time-windowed amplitude-based eigenvalues	1-50
Time-windowed phase-based eigenvalues	51-100
Time-windowed mean of CSI amplitudes	101-153
Time-windowed variance of CSI amplitudes	154-206
Time-windowed skewness of CSI amplitudes	207-259
Time-windowed kurtosis of CSI amplitudes	260-312
Time-windowed inter-quartile range of CSI amplitudes	313-365
Doppler spectrum mean	366
Doppler spectrum variance	367
Doppler spectrum centroid	368
2nd order moment of Doppler spectrum	369
Doppler spectrum skewness	370
Doppler spectrum kurtosis	371
Time-frequency-windowed mean-std deviation ratio of CSI energy	372
Time-frequency-windowed skewness of CSI energy	373
Time-frequency-windowed kurtosis of CSI energy	374

reason, we have chosen a variety of features computed in time, frequency, and Doppler spectrum domains over the provided CSI dataset. Time windows of 50 OFDM symbols have been used for the feature extraction in the time domain. Table I denotes with a checkmark [✓] those features considered in this paper. As a result, a total of 374 features have been indexed as shown in Table IV.

Considering the supervised architecture for device-free people counting in Fig. 1, an exhaustive assessment of features and supervised learning models have been applied to the provided dataset. The following ML models have been included in this paper, part of which have been used in the literature as detailed in Table II (further details on the models listed below can be found in [42]–[44]).

1) *GNB:* Which falls within the broad set of Naive Bayes classification models, all based on the assumption of statistical independence between every pair of features subject to the target variable to be predicted. GNB assumes that the conditional probability of every feature given the target is normally distributed.
2) *Adaptive Boosting:* An ensemble that hinges on the concatenation of individual weak learners so that they progressively concentrate on training samples misclassified by previous models (i.e., the boosting ensemble subsequently focuses on difficult training classes). The model configuration essentially comprises the weak learner, the number of learners in the ensemble and the learning rate.
3) *kNN:* A distance-based lazy classifier that assigns every test instance the majority class among their k closest examples (under a certain measure of distance) in the training dataset. Parameters to be tuned include the number of neighbors k, the distance measure and the voting scheme (e.g., unweighted or weighted depending on the relative distance value among the k neighbors).
4) *DT:* Which is based on discriminating classes among the training examples via a tree of nested splits. Such splits are driven by greedily adjusted threshold values imposed on specific features. The most influential parameter of this ML model is the depth to which branches of the tree are permitted to grow, which is a key to control model overfitting.
5) *Extreme Learning Machine (ELM):* A single-layer neural network whose hidden nodes are not tuned, yet initialized randomly. The model only needs to determine the output weights of the network, which can be done analytically and extremely fast with respect to the time taken by the training algorithm of other feed-forward neural classifiers. Typical ELM parameters to be tuned are the number of hidden nodes, and their activation function.
6) *LDA:* A linear classifier whose decision boundary is estimated by assuming that features corresponding to every class are normally distributed. This model is not parametric.
7) *LR:* Which predicts the likelihood that a new test sample belongs to a certain class by modeling the relationship between features and the target as a logistic function, whose parameters are fitted via maximum likelihood. Since we deal with a multiclass classification problem, a one-versus-rest scheme of the LR model is selected. Again, this model is nonparametric.
8) *RF:* A bagging ensemble composed by a number of weak DT learners trained over sample subsets drawn with replacement from the training set, whose outputs are voted for producing a predictive estimate of the overall model. Parameters to be tuned include the number of weak learners and their maximum depth, among others.
9) *Gradient Boosting Classifier (GBC):* A gradient-based version of a boosting ensemble based on generalizing the error committed by the weak learners along the ensemble to other arbitrary loss functions. The learning rate and the number of weak learners are arguably among the main parameters to be tuned in these models.
10) *SVM, νSVM:* Which hinge on discriminating classes with a gap as wide as possible, for which points can be mapped to high-dimensional feature spaces so as to find a maximally wide separating hyperplane between the classes. A parameter is included in the computation of the maximal hyperplane to vary the cost of misclassified training examples. Instead, νSVM regularizes the model based on the number of support vectors.

Methodologically this paper aims to shed light on how different ML models and features perform in the context of device-free people counting without any bias on how such models are parametrized. To this end, in all cases a nested cross-validation (CV) methodology will be adopted to compute different performance scores over each simulated case. Specifically, for every (scenario, model) combination the dataset containing examples for the selected scenario is partitioned in $M_{\text{outer}} = 4$ equal parts or folds. Then, one of such M_{outer} folds is left apart for testing, and the parameters of the model are tuned by performing a stratified K_{inner}-fold CV over the remaining $M_{\text{outer}} - 1$ folds, where M_{inner} is set to 4 in all cases. The parameter tuning is done based on a fine-grained value grid of the hyper-parameters of the model at hand, whose optimality is given by the average values of two different scores computed over the M_{inner} folds.

1) Accuracy, given by the ratio of correctly classified examples to the total number of validation/test samples.

TABLE V
ACCURACY AND COHEN'S KAPPA COEFFICIENT (FIRST AND SECOND ROW OF EVERY CELL, RESPECTIVELY), WITHOUT AND WITH FS. RESULTS ARE EXPRESSED AS MEAN±STD COMPUTED OVER NESTED CV WITH $M_{\text{inner}} = M_{\text{outer}} = 4$ FOLDS AND 20 REPETITIONS

Model	Scenarios											
	A		B		C		D		E		F	
	Without FS	With FS	Without FS	With FS	Without FS	With FS	Without FS	With FS	Without FS	With FS	Without FS	With FS
GNB	0.57 ± 0.00 0.43 ± 0.00	0.58 ± 0.01 0.43 ± 0.01	0.46 ± 0.00 0.35 ± 0.00	0.48 ± 0.02 0.38 ± 0.01	0.48 ± 0.00 0.35 ± 0.01	0.49 ± 0.00 0.37 ± 0.01	0.28 ± 0.00 0.14 ± 0.01	0.31 ± 0.01 0.16 ± 0.01	0.61 ± 0.00 0.51 ± 0.01	0.60 ± 0.01 0.50 ± 0.01	0.40 ± 0.00 0.28 ± 0.00	0.41 ± 0.00 0.30 ± 0.01
ADA	0.60 ± 0.01 0.46 ± 0.02	0.58 ± 0.01 0.45 ± 0.01	0.40 ± 0.00 0.27 ± 0.01	0.40 ± 0.01 0.28 ± 0.02	0.50 ± 0.01 0.37 ± 0.02	0.50 ± 0.01 0.36 ± 0.02	0.38 ± 0.01 0.26 ± 0.01	0.38 ± 0.01 0.27 ± 0.01	0.56 ± 0.02 0.45 ± 0.03	0.57 ± 0.02 0.47 ± 0.01	0.37 ± 0.01 0.25 ± 0.01	0.36 ± 0.01 0.26 ± 0.01
kNN	0.62 ± 0.01 0.50 ± 0.01	0.65 ± 0.01 0.53 ± 0.01	0.60 ± 0.01 0.52 ± 0.01	0.66 ± 0.00 0.58 ± 0.01	0.62 ± 0.01 0.52 ± 0.01	0.68 ± 0.00 0.60 ± 0.01	0.43 ± 0.01 0.31 ± 0.01	0.44 ± 0.01 0.33 ± 0.01	0.68 ± 0.01 0.60 ± 0.01	0.74 ± 0.02 0.68 ± 0.01	0.53 ± 0.00 0.44 ± 0.00	0.59 ± 0.01 0.51 ± 0.01
DT	0.61 ± 0.01 0.48 ± 0.02	0.60 ± 0.01 0.50 ± 0.02	0.65 ± 0.00 0.58 ± 0.01	0.68 ± 0.01 0.59 ± 0.01	0.62 ± 0.01 0.52 ± 0.01	0.64 ± 0.02 0.53 ± 0.02	0.42 ± 0.01 0.31 ± 0.01	0.42 ± 0.01 0.32 ± 0.01	0.67 ± 0.01 0.58 ± 0.01	0.67 ± 0.00 0.60 ± 0.01	0.48 ± 0.01 0.37 ± 0.01	0.50 ± 0.01 0.40 ± 0.01
ELM	0.66 ± 0.02 0.56 ± 0.01	0.71 ± 0.01 0.60 ± 0.01	0.66 ± 0.01 0.60 ± 0.01	0.68 ± 0.01 0.61 ± 0.01	0.58 ± 0.01 0.47 ± 0.01	0.64 ± 0.00 0.55 ± 0.01	0.39 ± 0.02 0.27 ± 0.01	0.44 ± 0.01 0.34 ± 0.01	0.72 ± 0.01 0.65 ± 0.01	0.75 ± 0.01 0.70 ± 0.02	0.48 ± 0.00 0.38 ± 0.00	0.53 ± 0.01 0.44 ± 0.01
LDA	0.67 ± 0.01 0.56 ± 0.01	0.72 ± 0.01 0.61 ± 0.01	0.68 ± 0.01 0.62 ± 0.01	0.70 ± 0.01 0.63 ± 0.01	0.65 ± 0.01 0.56 ± 0.01	0.68 ± 0.01 0.60 ± 0.01	0.41 ± 0.00 0.29 ± 0.01	0.44 ± 0.01 0.33 ± 0.01	0.73 ± 0.01 0.66 ± 0.01	0.77 ± 0.02 0.72 ± 0.01	0.55 ± 0.01 0.46 ± 0.01	0.58 ± 0.01 0.50 ± 0.01
LR	0.68 ± 0.01 0.57 ± 0.01	0.73 ± 0.01 $\mathbf{0.64 \pm 0.01}$	0.69 ± 0.01 0.62 ± 0.01	0.74 ± 0.01 0.69 ± 0.01	0.63 ± 0.02 0.54 ± 0.01	$\mathbf{0.74 \pm 0.01}$ 0.67 ± 0.01	0.40 ± 0.01 0.28 ± 0.01	0.46 ± 0.01 0.35 ± 0.01	0.72 ± 0.01 0.65 ± 0.01	0.77 ± 0.01 0.72 ± 0.01	0.53 ± 0.01 0.44 ± 0.01	0.58 ± 0.01 0.50 ± 0.01
RF	$\mathbf{0.71 \pm 0.01}$ $\mathbf{0.61 \pm 0.01}$	0.70 ± 0.00 0.62 ± 0.01	0.72 ± 0.01 0.66 ± 0.01	0.73 ± 0.00 0.68 ± 0.01	$\mathbf{0.71 \pm 0.00}$ $\mathbf{0.65 \pm 0.01}$	$\mathbf{0.73 \pm 0.01}$ 0.67 ± 0.01	$\mathbf{0.55 \pm 0.01}$ $\mathbf{0.46 \pm 0.01}$	$\mathbf{0.57 \pm 0.01}$ $\mathbf{0.48 \pm 0.01}$	0.79 ± 0.00 0.74 ± 0.01	0.80 ± 0.01 0.75 ± 0.01	$\mathbf{0.61 \pm 0.01}$ $\mathbf{0.54 \pm 0.01}$	0.64 ± 0.01 0.57 ± 0.01
GBC	$\mathbf{0.71 \pm 0.01}$ $\mathbf{0.61 \pm 0.01}$	0.70 ± 0.01 0.60 ± 0.02	$\mathbf{0.76 \pm 0.00}$ $\mathbf{0.71 \pm 0.01}$	$\mathbf{0.76 \pm 0.01}$ $\mathbf{0.72 \pm 0.01}$	$\mathbf{0.71 \pm 0.01}$ $\mathbf{0.64 \pm 0.01}$	0.72 ± 0.01 0.65 ± 0.01	0.53 ± 0.01 0.44 ± 0.01	0.54 ± 0.00 0.45 ± 0.01	$\mathbf{0.81 \pm 0.01}$ $\mathbf{0.76 \pm 0.01}$	0.81 ± 0.01 0.76 ± 0.01	0.58 ± 0.01 0.50 ± 0.00	0.60 ± 0.01 0.53 ± 0.01
νSVM	0.69 ± 0.01 0.58 ± 0.01	$\mathbf{0.75 \pm 0.01}$ $\mathbf{0.65 \pm 0.01}$	0.71 ± 0.01 0.66 ± 0.01	$\mathbf{0.75 \pm 0.01}$ 0.68 ± 0.01	0.64 ± 0.01 0.55 ± 0.01	0.72 ± 0.00 0.66 ± 0.01	0.48 ± 0.00 0.38 ± 0.01	0.54 ± 0.01 0.43 ± 0.01	0.78 ± 0.01 0.73 ± 0.01	0.80 ± 0.01 0.75 ± 0.01	0.58 ± 0.01 0.50 ± 0.01	$\mathbf{0.65 \pm 0.02}$ 0.56 ± 0.01
SVM	0.69 ± 0.01 0.59 ± 0.01	$\mathbf{0.76 \pm 0.01}$ $\mathbf{0.65 \pm 0.02}$	0.71 ± 0.01 0.66 ± 0.01	$\mathbf{0.75 \pm 0.01}$ 0.70 ± 0.01	0.64 ± 0.01 0.55 ± 0.01	0.72 ± 0.02 $\mathbf{0.69 \pm 0.01}$	0.48 ± 0.00 0.38 ± 0.01	0.54 ± 0.00 0.43 ± 0.01	$\mathbf{0.80 \pm 0.01}$ $\mathbf{0.75 \pm 0.01}$	$\mathbf{0.85 \pm 0.01}$ $\mathbf{0.79 \pm 0.01}$	0.59 ± 0.01 0.51 ± 0.01	$\mathbf{0.66 \pm 0.01}$ $\mathbf{0.59 \pm 0.01}$

2) Cohen's Kappa coefficient, which measures the agreement between two datasets by accounting for the chance that data could coincide just by chance [45]. A fair level of agreement between the predicted and the true number of people is given by values of this coefficient above 0.2, with 0.8 acknowledged as the lower bound for a *very good* agreement [46]. It should be remarked that this coefficient removes any interpretative bias due to an imbalanced class distribution in the datasets.

Once the model is selected based on any of the above scores, a prediction is obtained for the fold left apart for testing. The same model tuning and testing procedure is repeated by iterating on the remaining $M_{\text{outer}} - 1$ folds, after which an average score can be computed from the results obtained for every fold. Then the overall process is replicated for 20 repetitions by randomly shuffling the instances within the dataset so as to remove any influence of the partitioning on the performance of the models.

Features of training examples are transformed to their standard score prior to model learning, whereas validation and test instances are also standardized using the statistics computed over the training subset. This is crucial for certain ML models in the benchmark with sensitiveness to varying feature support (e.g., distance-based schemes such as *k*NN).

A. Results and Discussion

We begin the discussion with Table V, where the mean and standard deviation statistics of the accuracy and Kappa scores computed over the 20 repetitions of the nested CV method are displayed for all models and scenarios within the EHUCOUNT dataset. The table includes results obtained: 1) with the entire set of features in Table IV fed to every model and 2) with a selected feature subset resulting from the application of an FS method prior to model training. Among the myriad of FS approaches reported in the ML literature a simple model-based approach is adopted: an RF model is first built over the training dataset with all features included, from which those accounting for an *importance* greater than the median across all predictors are retained as the selected feature subset. The so-called importance is computed as the average impurity decrease of the feature when it is utilized to grow a compounding tree of the RF model.

Several conclusions can be drawn by inspecting these results. To begin with, the best Kappa scores obtained for every scenario are above 0.4, which indicate a fairly good agreement between the true and predicted number of people. However, scores vary significantly among scenarios: while some setups yield very good prediction results for the best model (e.g., the E scenario with SVM and FS, which achieves an average accuracy of 85% and an average Kappa of approximately 0.8), best scores are notably lower for other cases (e.g., the D scenario, for which the performance scores of the same model degrade down to an average accuracy of 54% and an average Kappa of 0.43). This observed fact confirms that the physical characteristics of the scenario under study have a strong influence on the ultimate performance of the model. Furthermore, the benchmark also evinces that the incorporation of FS techniques improves in general the performance of the model, yet the obtained gain is also dependent on both the model and the scenario. For instance, in scenario D the selection of features does not improve significantly except for LR, SVM, and νSVM. Interesting is to note that the gain due to FS is particularly high in SVM-based models in all considered scenarios, a result that goes in line with the reported impact of a good choice of features for this family of classifiers [47].

We follow the discussion by commenting on the statistical behavior of the scores in Table V. To this end, the outperforming models without and with FS are highlighted in dark blue for every scenario, with the statistical significance of the performance gaps being determined by a Wilcoxon [48] signed-rank test with level of significance equal to $\alpha = 5\%$. This test permits to determine whether the scores obtained for two different models belong to populations drawn from distributions with significant mean differences. In some cases, more than a single model has been found to be superior for every scenario since the performance gaps of each other are found to be statistically irrelevant for the established confidence interval. So occurs in scenario A, where SVM-based models, GBC and RF are found to perform statistically similar

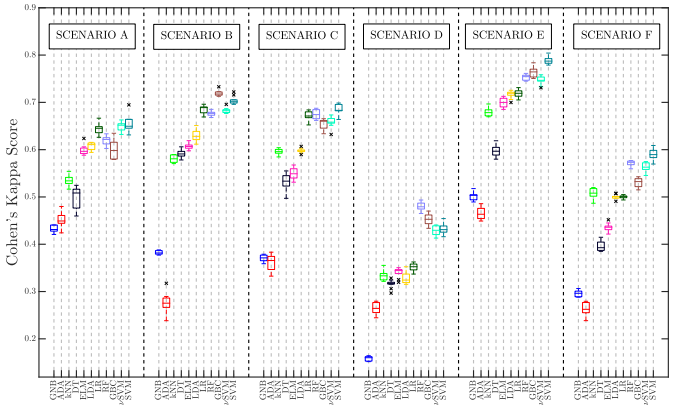

Fig. 5. Boxplot of the Kappa scores obtained by every model for all scenarios in the EHUCOUNT dataset, with FS. Score values falling outside the range $[Q_1 - 1.5 \cdot \text{IQR}, Q_3 + 1.5 \cdot \text{IQR}]$ (outliers) are highlighted as ×, with Q_1, Q_3, and IQR denoting the lower quartile, the upper quartile, and the interquartile range, respectively.

TABLE VI
NORMALIZED CONFUSION MATRIX CORRESPONDING TO SCENARIO D WITH SVM AND FS

Scenario D, SVM		Predicted class					
		0	1	2	3	4	5
True class	0	80.8%	8.1%	2.7%	2.3%	3.5%	2.6%
	1	11.0%	60.0%	13.2%	12.0%	2.1%	1.7%
	2	6.6%	19.0%	38.1%	17.4%	10.2%	8.7%
	3	6.2%	16.3%	15.1%	43.5%	11.2%	7.7%
	4	3.1%	4.4%	12.2%	13.7%	54.5%	12.1%
	5	2.7%	5.0%	15.7%	16.0%	18.2%	42.4%

TABLE VII
NORMALIZED CONFUSION MATRIX CORRESPONDING TO SCENARIO E WITH SVM AND FS

Scenario E, SVM		Predicted class				
		0	1	2	3	4
True class	0	99.6%	0.4%	0.0%	0.0%	0.0%
	1	0.8%	90.6%	4.6%	1.6%	2.4%
	2	0.6%	6.9%	83.4%	2.5%	6.6%
	3	0.7%	3.8%	5.1%	69.6%	20.8%
	4	0.1%	1.5%	8.2%	16.9%	73.3%

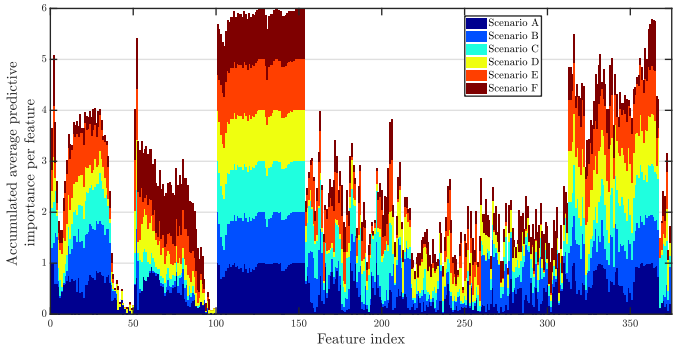

Fig. 6. Accumulative feature importance for the scenarios under study.

to each other when no FS is used. When discarding irrelevant predictors, the improvement of the generalization performance experienced by SVM-based models makes them stand out from the rest of classifiers. This is further buttressed by the boxplots in Fig. 5, where the statistical distribution of the obtained Kappa scores is depicted for all scenarios and models with FS: it is clear from this plot that the performance of GBC and/or SVM models is superior to the rest of the benchmark except for scenario D, where RF dominates with statistical significance in terms of Kappa score.

At this point a closer look can be taken at the results by analyzing how prediction errors distribute over different classes (i.e., number of people). This analysis is exposed in the confusion matrix shown in Table VI, where the distribution of predictions is computed with respect to the true value of test samples for scenario D (namely, that yielding the worst scores in Table V) and the SVM model with FS. Values are normalized over the $20 \cdot 4 = 80$ test data subsets for this (model, scenario) combination. As expected, the most predictable class corresponds to 0 people traversing the scenario, but the $100 - 80.8 = 19.2\%$ of misclassified examples spread over the rest of labels unveils the non-negligible impact that the existing activity in the area can have in certain environments. For the sake of realism in the obtained results no countermeasures were implemented to reduce the effects of noncontrolled phenomena during data recording, e.g., other moving people in the surroundings or concurrent WiFi activity within the same spectrum band. When detecting a higher number of people, the classification error results to be more disperse over labels, yet most of the misclassified samples gather around the true label, e.g., when detecting four people 54.5% of the test samples were correctly predicted, while 45.5% of the entire set of test samples correspond to traces wrongly predicted to be caused by 5 (12.1%), 3 (13.7%), 2 (12.2%), 1 (4.4%), and 0 (3.1%) people in the environment.

Similar conclusions hold for other scenarios, as exemplified in Table VII for scenario E (the one with best scores as per Table V) with the same ML model and FS. In this latter case misclassification increases when detecting a higher number of persons; interestingly, the almost perfect prediction score (99.6%) for the case when no people is in the scenario is symptomatic of a good approach for *presence detection*, which is of inherent utility for, e.g., security applications in critical infrastructures. All in all, good discrimination properties among classes (i.e., number of people) can be concluded for the evaluated models when the scores in Tables VI and VII are compared to the lower accuracy bound of random guessing (i.e., 16.6% and 20% for scenarios D and E, respectively).

Finally, we end the discussion by elaborating on Fig. 6. This plot shows the averaged importance of the feature set in Table I for the analyzed scenarios. The importance of every feature is computed as defined at the beginning of Section V-A for each of the 80 train/test folds performed for every scenario, averaged and stacked on top of each other in this figure. Based on this procedure, the colored stacked plot in Fig. 6 provides two interesting insights: 1) it discriminates the relative importance among scenarios, so that the relevance of each predictor can be visually assessed across all scenarios and 2) a high amplitude of the stacked importance values of a certain predictor (up to a maximum of 6) indicates that the feature in question is highly informative for all scenarios, which serves as an empirical response to the research questions posed in the introduction.

This being said, one can see the high importance of the time-windowed mean of the CSI amplitudes (indexes 101–153) in all scenarios (the accumulated importance is close to 6). In this line of reasoning, the time-windowed interquartile range of CSI amplitudes also play a relevant role in averaged terms. Paying attention to the time-correlations-based descriptors, we can observe that are most of the eigenvalues from the Pearson correlation matrices of both amplitude and phase are also interesting features for a crowd counting application. It is also worth noticing that not only the maximum eigenvalues, employed in the previous works because of concentrating most of the correlation matrix energy, but also the rest of eigenvalues result to be informative descriptors. Additionally, it can be seen that there is no correlation between the accumulated importance of amplitude and phase-based eigenvalues. In consequence, it is clear that phase-based descriptors include a higher diversity that could help the classification performance.

B. IoT Implementation Aspects

Although the measurements provided in the dataset have been obtained with laboratory equipment, they are insightful for an IoT implementation, taking into account the following aspects. First, different computation architectures should be considered in the deployment of the people counting framework according to the bandwidth and computational resources of the different commodity WiFi devices that can form a WiFi IoT network (i.e., access points or end-points). That is, depending on the learning model and the number of features employed in the detection process, one should select the computation architecture that meets best the specifications and functionalities of the IoT nodes. As can be seen in Table V, SVM-based models with FS provide in average terms higher performance scores at the expense of an important computational effort when compared to other models. In this sense, when WiFi IoT nodes cannot afford the computational cost of the models, they should transfer raw CSI data or the selected feature set to the cloud so as to perform complex processing therein. Additionally, moving WiFI IoT nodes should also follow a centralized Cloud computing scheme due to the need of periodically retraining the models and adapt their learned knowledge to changing scenarios.

Second, depending on the regular service/application of commodity WiFi end-points, data traffic constraints may hold, disabling a dedicated CSI data flow to the Cloud. In this case, an Edge Computing architecture should be rather implemented wherein final decisions could be shared with other remote IoT installations.

Lastly, this paper has assessed the performance and predictive potential of the ML models combined with FS for counting people through WiFi signals. In this context, the deployment of a collaborative WiFi IoT network could further improve the predictive capabilities of the whole system over a monitored area. Data from multiple WiFi IoT devices can provide an additional level of diversity that would allow contrasting single decisions for self-correction and transfer knowledge among IoT nodes.

VI. Conclusion

In this paper device-free people counting using WiFi signals in IoT environments has been addressed in depth from different research angles.

1) First, a thorough survey of the advances made in this field has been presented, with a clear focus on the plethora of contributions resorting to ML models to undertake this multiclass detection problem. The comprehensive literature review performed in this paper has unveiled opportunities and trends in regards to the universality of the engineered features and the narrow spectrum of models used in the previous works. These noted facts are a key for the construction of good predictive models for detecting people from WiFi traces captured by IoT sensors.
2) Next, an architectural design for an IoT system has been presented, underscoring the technical and computational implications of deploying an ML-based people counting functionality at different levels.
3) With the aim to encourage the research community to delve into this problem, a new dataset coined as `EHUCOUNT` has been presented. The data repository, which has been made available in [40], contains raw data traces embedding amplitude and phase information of WiFi OFDM subcarriers, which have been captured over physical spaces of very different topology and characteristics. This data substrate is expected not only to foster new advances on device-free people counting, but also to help validating new proposals made in this research field.
4) Finally, an experimental benchmark has been discussed with 11 different ML models and 374 features extracted from the `EHUCOUNT` dataset. Enlightening results have been achieved in regards to the selection of the classifier and the predictive importance empirically quantified after embedded FS for each of the considered scenarios. On one hand, the choice of the ML model has been proven to be crucial and should not be overlooked anyhow in future studies. On the other hand, features have rendered varying levels of importance across the considered scenarios; however, a subset of predictors has been found to be recurrently chosen in all setups, a fact that underpins the need for gauging the universality of the predictive power of each feature over different scenarios. Some implementation notes for a practical IoT deployment of the ML models completes the experimental part of this paper.

Several future research lines stem from the existing literature and the work presented in this paper. The observation around the universality of features highlighted in the last point above calls for the adoption of inductive learning (also known as transfer learning) so as to make people counting ML models leverage the information captured in different scenarios. In essence transfer learning gathers all ML techniques aimed at exploiting the knowledge learned in a certain predictive task in another, interrelated task. The adoption of transfer learning might reduce the initial training data requirements when the IoT WiFi sensors are deployed in a new scenario.

In this hypothesized setup an IoT architecture based on the centralized Cloud computing paradigm as per Fig. 3 should be preferred, since transfer learning techniques are known to be computationally demanding. Another research line worth to be tackled from this paper onward is the renowned capability of deep learning (DL) models to automatically learn not only the patterns between features and the class to be predicted, but also the features themselves. DL classifiers have been recently utilized in the context of image-based people counting [49], [50], but to the best of the authors' knowledge, they remain unexplored with WiFi-based IoT sensing data. Furthermore, recent advances in evolutionary feature extraction wrappers (e.g., genetic programming [51]) and on-line learning in non-stationary environments are also on the spotlight for the near future.

REFERENCES

[1] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," *Future Gener. Comput. Syst.*, vol. 29, no. 7, pp. 1645–1660, 2013.

[2] F. Xia, L. T. Yang, L. Wang, and A. Vinel, "Internet of Things," *Int. J. Commun. Syst.*, vol. 25, no. 9, pp. 1101–1102, 2012.

[3] J. A. Stankovic, "Research directions for the Internet of Things," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 3–9, Feb. 2014.

[4] L. Da Xu, W. He, and S. Li, "Internet of Things in industries: A survey," *IEEE Trans. Ind. Informat.*, vol. 10, no. 4, pp. 2233–2243, Nov. 2014.

[5] K. R. Joshi, D. Bharadia, M. Kotaru, and S. Katti, "WiDeo: Fine-grained device-free motion tracing using RF backscatter," in *Proc. NSDI*, Oakland, CA, USA, 2015, pp. 189–204.

[6] M. Seifeldin, A. Saeed, A. E. Kosba, A. El-Keyi, and M. Youssef, "Nuzzer: A large-scale device-free passive localization system for wireless environments," *IEEE Trans. Mobile Comput.*, vol. 12, no. 7, pp. 1321–1334, Jul. 2013.

[7] H. Abdelnasser, K. A. Harras, and M. Youssef, "Ubibreathe: A ubiquitous non-invasive WiFi-based breathing estimator," in *Proc. 16th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, Hangzhou, China, 2015, pp. 277–286.

[8] J. Liu *et al.*, "Tracking vital signs during sleep leveraging off-the-shelf WiFi," in *Proc. 16th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, Hangzhou, China, 2015, pp. 267–276.

[9] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, "Whole-home gesture recognition using wireless signals," in *Proc. 19th Annu. Int. Conf. Mobile Comput. Netw.*, Miami, FL, USA, 2013, pp. 27–38.

[10] Y. Wang, K. Wu, and L. M. Ni, "WiFall: Device-free fall detection by wireless networks," *IEEE Trans. Mobile Comput.*, vol. 16, no. 2, pp. 581–594, Feb. 2017.

[11] M. Raja and S. Sigg, "Applicability of RF-based methods for emotion recognition: A survey," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops (PerCom Workshops)*, Sydney, NSW, Australia, 2016, pp. 1–6.

[12] F. J. Oppermann, C. A. Boano, and K. Römer, "A decade of wireless sensing applications: Survey and taxonomy," in *The Art of Wireless Sensor Networks*. New York, NY, USA: Springer, 2014, pp. 11–50.

[13] M. Youssef, M. Mah, and A. Agrawala, "Challenges: Device-free passive localization for wireless environments," in *Proc. 13th Annu. ACM Int. Conf. Mobile Comput. Netw. (MobiCom)*, Montreal, QC, Canada, 2007, pp. 222–229.

[14] S. Savazzi *et al.*, "Device-free radio vision for assisted living: Leveraging wireless channel quality information for human sensing," *IEEE Signal Process. Mag.*, vol. 33, no. 2, pp. 45–58, Mar. 2016.

[15] K. Witrisal *et al.*, "High-accuracy localization for assisted living: 5G systems will turn multipath channels from foe to friend," *IEEE Signal Process. Mag.*, vol. 33, no. 2, pp. 59–70, Mar. 2016.

[16] C. Wu *et al.*, "Non-invasive detection of moving and stationary human with WiFi," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 11, pp. 2329–2342, Nov. 2015.

[17] E. Cianca, M. De Sanctis, and S. Di Domenico, "Radios as sensors," *IEEE Internet Things J.*, vol. 4, no. 2, pp. 363–373, Apr. 2017.

[18] C. Xu *et al.*, "SCPL: Indoor device-free multi-subject counting and localization using radio signal strength," in *Proc. ACM/IEEE Int. Conf. Inf. Process. Sensor Netw. (IPSN)*, Philadelphia, PA, USA, Apr. 2013, pp. 79–90.

[19] J. Xiao, K. Wu, Y. Yi, L. Wang, and L. M. Ni, "FIMD: Fine-grained device-free motion detection," in *Proc. IEEE 18th Int. Conf. Parallel Distrib. Syst. (ICPADS)*, Singapore, Dec. 2012, pp. 229–235.

[20] K. Qian, C. Wu, Z. Yang, Y. Liu, and Z. Zhou, "PADS: Passive detection of moving targets with dynamic speed using PHY layer information," in *Proc. 20th IEEE Int. Conf. Parallel Distrib. Syst. (ICPADS)*, Hsinchu, Taiwan, Dec. 2014, pp. 1–8.

[21] F. Viani, A. Polo, E. Giarola, M. Salucci, and A. Massa, "Principal component analysis of CSI for the robust wireless detection of passive targets," in *Proc. Int. Appl. Comput. Electromagnet. Soc. Symp. Italy (ACES)*, Florence, Italy, Mar. 2017, pp. 1–2.

[22] L. Gong *et al.*, "WiFi-based real-time calibration-free passive human motion detection," *Sensors*, vol. 15, no. 12, pp. 32213–32229, Dec. 2015.

[23] L. Gong *et al.*, "An adaptive wireless passive human detection via fine-grained physical layer information," *Ad Hoc Netw.*, vol. 38, pp. 38–50, Mar. 2016.

[24] J. Lv, W. Yang, L. Gong, D. Man, and X. Du, "Robust WLAN-based indoor fine-grained intrusion detection," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, Washington, DC, USA, Dec. 2016, pp. 1–6.

[25] W. Xi *et al.*, "Electronic frog eye: Counting crowd using WiFi," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Toronto, ON, Canada, Apr. 2014, pp. 361–369.

[26] J. Zhang, B. Wei, W. Hu, and S. S. Kanhere, "WiFi-ID: Human identification using WiFi signal," in *Proc. Int. Conf. Distrib. Comput. Sensor Syst. (DCOSS)*, Washington, DC, USA, May 2016, pp. 75–82.

[27] Y. Zeng, P. H. Pathak, and P. Mohapatra, "Analyzing shopper's behavior through WiFi signals," in *Proc. 2nd Workshop Phys. Analytics (WPA)*, Florence, Italy, 2015, pp. 13–18.

[28] J. Lv, W. Yang, and D. Man, "Device-free passive identity identification via WiFi signals," *Sensors*, vol. 17, no. 11, pp. 1–17, 2017.

[29] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[30] S. Di Domenico, M. De Sanctis, E. Cianca, and G. Bianchi, "A trained-once crowd counting method using differential WiFi channel state information," in *Proc. 3rd Int. Workshop Phys. Analytics (WPA)*, Singapore, 2016, pp. 37–42.

[31] Z. Zhou, Z. Yang, C. Wu, L. Shangguan, and Y. Liu, "Omnidirectional coverage for device-free passive human detection," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 7, pp. 1819–1829, Jul. 2014.

[32] W. Liu, X. Gao, L. Wang, and D. Wang, "Bfp: Behavior-free passive motion detection using PHY information," *Wireless Pers. Commun.*, vol. 83, no. 2, pp. 1035–1055, 2015.

[33] S. Depatla, A. Muralidharan, and Y. Mostofi, "Occupancy estimation using only WiFi power measurements," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 7, pp. 1381–1393, Jul. 2015.

[34] S. Di Domenico, G. Pecoraro, E. Cianea, and M. De Sanctis, "Trained-once device-free crowd counting and occupancy estimation using WiFi: A doppler spectrum based approach," in *Proc. IEEE 12th Int. Conf. Wireless Mobile Comput. Netw. Commun. (WiMob)*, New York, NY, USA, Oct. 2016, pp. 1–8.

[35] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[36] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.

[37] W. Shi and S. Dustdar, "The promise of edge computing," *Computer*, vol. 49, no. 5, pp. 78–81, May 2016.

[38] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[39] J. Lee, B. Bagheri, and H.-A. Kao, "A cyber-physical systems architecture for industry 4.0-based manufacturing systems," *Manuf. Lett.*, vol. 3, pp. 18–23, Jan. 2015.

[40] I. Sobron, J. Del Ser, I. Eizmendi, and M. Velez. (2017). *EHUCOUNT Dataset*. Accessed: Nov. 30, 2017. [Online]. Available: www.ehu.eus/tsr_radio/index.php/research-areas/data-analytics-in-wireless-networks

[41] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Tool release: Gathering 802.11n traces with channel state information," *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 1, p. 53, Jan. 2011.

[42] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. Amsterdam, The Netherlands: Morgan Kaufmann, 2016.
[43] C. C. Aggarwal, *Data Classification: Algorithms and Applications*. Boca Raton, FL, USA: CRC Press, 2014.
[44] G.-B. Huang, D. H. Wang, and Y. Lan, "Extreme learning machines: A survey," *Int. J. Mach. Learn. Cybern.*, vol. 2, no. 2, pp. 107–122, 2011.
[45] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.
[46] D. G. Altman, *Practical Statistics for Medical Research*. London, U.K.: CRC Press, 1990.
[47] J. Weston *et al.*, "Feature selection for SVMs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 668–674.
[48] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, 1945.
[49] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 833–841.
[50] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 589–597.
[51] B. Tran, B. Xue, and M. Zhang, "Genetic programming for feature construction and selection in classification on high-dimensional data," *Memetic Comput.*, vol. 8, no. 1, pp. 3–15, 2016.

Iker Sobron (M'10) received the Electronics Engineering degree from the University of the Basque Country (UPV/EHU), Leioa, Spain, in 2006, the Physics degree from the National University of Distance Education, Madrid, Spain, in 2010, and the Ph.D. degree in electronics engineering from the University of Mondragon (MU), Mondragón, Spain, in 2011.

From 2011 to 2012, he was with the Department of Electronics and Computing, MU, as a Lecturer and a Researcher in the signal theory and communications area. Since 2013, he has been with the Department of Communications Engineering, UPV/EHU, where he has been Post-Doctoral Researcher with the TSR Group and a Substitute Lecturer. From 2013 and 2014, he was Post-Doctoral Visiting Researcher (SMT Lab) with the Department of Electronics and Computing Engineering, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil. His current research interests include next-generation wireless communications (IoT systems, 5G networks, cognitive radio, and spectrum sharing), machine learning, sensor and distributed networks, and detection and estimation theory.

Javier Del Ser (M'07–SM'12) received the Ph.D. degree (*cum laude*) in telecommunication engineering from the University of Navarra, Pamplona, Spain, in 2006 and the Ph.D. degree (*summa cum laude*) in computational intelligence from the University of Alcalá, Alcalá de Henares, Spain, in 2013.

He is currently a Research Professor of data analytics and optimization with TECNALIA, Donostia-San Sebastián, Spain, a Visiting Fellow with the Basque Center for Applied Mathematics, Bilbao, Spain, and an Adjunct Professor with the University of the Basque Country (UPV/EHU), Leioa, Spain. He has authored or co-authored over 220 papers and conference contributions, co-supervises over 10 Ph.D. theses, has edited 4 books and co-invented 6 patents. His current research interests include use of descriptive, prescriptive, and predictive data mining and optimization in a diverse range of application and sectors such as energy, transport, telecommunications, industry, and security.

Iñaki Eizmendi (M'14) received the M.S. and Ph.D. degrees in telecommunications engineering with the University of the Basque Country (UPV/EHU), Leioa, Spain, in 1994 and 2012, respectively.

He was a Research and Development Engineer with several companies. From 2003, he has been with the TSR (radiocommunications and signal processing) Research Group, Communications Engineering Department, UPV/EHU, where he is an Assistant Professor. His current research interest includes wireless communications systems including new digital broadcasting technologies (DVB-T2 and ATSC 3.0), 5G, and IoT.

Manuel Vélez (M'03) received the M.S. and Ph.D. degrees in telecommunication engineering from the University of the Basque Country (UPV/EHU), Leioa, Spain, in 1993 and 2008, respectively.

In 1995, he joined the University of the Basque Country, where he is currently an Associate Professor with the Department of Communications Engineering. In 2008, he was a Visiting Post-Doctoral Researcher with the Department of Information Technology, University of Turku, Turku, Finland. In 2017, he was a Visiting Professor with the Department of Computer and Information Science, University of Konstanz, Konstanz, Germany. He has been involved in several research projects concerning broadcasting and wireless systems for 20 years and has supervised several Ph.D. dissertations in this field. He has co-authored over 100 papers in international journals and conferences. His current research interests include new broadcasting technologies, wireless communications systems, software defined radio developments, and system testing and data analytics for spectrum and mobility management.