

Received 2 June 2023, accepted 3 July 2023, date of publication 11 July 2023, date of current version 18 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3294429

RESEARCH ARTICLE

eHealth CSI: A Wi-Fi CSI Dataset of Human Activities

IANDRA GALDINO, JULIO C. H. SOTO^{ID}, EGBERTO CABALLERO^{ID},
VINICIUS FERREIRA^{ID}, (Member, IEEE), TAIANE COELHO RAMOS, CÉLIO ALBUQUERQUE^{ID},
AND DÉBORA C. MUCHALUAT-SAADE^{ID}, (Member, IEEE)

MídiaCom Laboratory, Institute of Computing, Universidade Federal Fluminense (UFF), Niterói 24210-240, Brazil

Corresponding author: Iandra Galdino (igaldino@midiacon.uff.br)

This study was financed in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), CAPES Programa Institucional de Internacionalização (Print), Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Instituto Nacional de Ciência e Tecnologia em Medicina Assistida por Computação Científica (INCT-MACC).

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of the Brazilian Ministry of Health under CAAE No. 54359221.4.0000.5243.

ABSTRACT Recent studies corroborate Wi-Fi Channel State Information (CSI) usability to monitor patients remotely and obtain health information non-invasively and with low-cost. In addition to monitoring vital signs, this technology can also be applied to presence detection, subject recognition, position and movement identification, among other uses. Despite its wide range of potential applications, there is a lack of CSI datasets that cover multiple activities and include participants' phenotype information to help develop, test, and compare new solutions. This study highlights the importance of building a robust public Wi-Fi CSI dataset. Therefore, we present *eHealth CSI*, a CSI dataset that includes a variety of Wi-Fi CSI data from more than 100 people in various activities in a controlled room. We also include CSI data collected in the same room without the presence of participants. This dataset was made publicly available online to other researchers under request. In this work, we introduce CSI technology and describe the data collection setup. In addition to CSI data, our dataset includes participants' phenotype information and heartbeat rate monitoring data using a smartwatch.

INDEX TERMS Wi-Fi CSI, dataset, vital sign monitoring.

I. INTRODUCTION

Wi-Fi devices are currently available in almost every environment. Electromagnetic waves of Wi-Fi signals can pass through some materials without requiring a direct line-of-sight for data transmission. However, the characteristics of the signal can be affected by environmental changes, including the presence and movement of humans, resulting in alterations in the received signal [1], [2]. These changes can be identified in Channel State Information (CSI) data. CSI provides channel status information available at a receiver's physical layer (PHY), such as amplitude, phase, and Received Signal Strength Indicator (RSSI) for each subcarrier involved in a multicarrier transmission [3], [4]. Current Wi-Fi standards, such as 802.11n/ac, use orthogonal frequency

division multiplexing (OFDM) modulation at the physical layer. OFDM is a technique that divides the transmission frequency band into several sub-bands, also called subcarriers. Each subcarrier can provide detailed information on the state of the channel [5]. CSI represents the Channel Frequency Response (CFR) for each subcarrier between the pairs of transmitting and receiving antennas.

Wi-Fi CSI based technologies have attracted much attention from academia and industry due to their potential to provide necessary information for several applications of contactless technologies. CSI can capture how the human body interferes with the electromagnetic signal in time, frequency, and spatial domains. This information can be used for various applications such as human presence detection, motion detection, human identification, fall detection, gesture recognition, and human localization [3], [4], [6], [7]. Many studies compared different wireless technologies [3], [6], [8]

The associate editor coordinating the review of this manuscript and approving it for publication was Mansoor Ahmed^{ID}.

applied to sensing, such as behavior recognition [4], [7], and indoor localization [9], [10].

One of the most prominent applications of Wi-Fi CSI technology is healthcare [11]. Due to the Covid-19 pandemic, we have faced an increasing number of patients who demand healthcare. It is a highly contagious and sometimes lethal disease that requires patient monitoring as contactless as possible. Several proposals have been studied in the literature for patient monitoring to address this demand [11]. In [12] for example, the authors proposed using frequency modulated carrier wave (FMCW) technology to detect human activities through radio frequency signals. Another possible solution to the contactless monitoring of patients is the use of radio frequency identification (RFID) technology [13]. RFID is a promising approach, but it depends on RFID tags to connect to patients. Therefore, searching for a low-cost and non-invasive approach has shown that Wi-Fi signals can be used to track human activities and movements, and CSI analysis can detect vital signs such as respiration, heartbeat, and others [2], [5], [6], [7], [14].

CSI promises to contribute to several applications, such as monitoring systems, surveillance, and person identity recognition. However, the lack of high-quality CSI datasets restricts the development of new CSI applications. To the best of our knowledge, there is no extensive and free Wi-Fi CSI database that provides large volumes of data and includes participants' information. The datasets available in the literature usually offer limited information, such as a low number of participant positions, short collecting time, and a low number of participants, including a very small amount of females.

Given the limitations of the available datasets in the literature, we proposed an experiment design and collection of Wi-Fi CSI data considering different possible applications. For this, our collection scenario was configured based on the room's dimensions in which the CSI data would be collected. Thus, we specified the position of each equipment and the distance between them and a participant. Then we proceeded to configure the CSI data collection parameters such as duration of each collecting, participant position, and sample frequency. Subsequently, the positions and protocols to be followed by the participant for data collection were defined. We have invited participants to collaborate voluntarily with this research among the academic community. Finally, we built a dataset repository and made it publicly available on a website.

In summary, the main objectives of this work are:

- Providing a dataset specifically designed for the development of new applications of CSI technology.
- Providing a detailed description of the experiment design, including equipment specification and protocols. We also aim to report the challenges faced during the data collection process and the adopted solutions.
- A comprehensive characterization of the dataset and a discussion of limitations and key factors to allow its straightforward usage.

- Providing a set of empty room data collected to allow room characterization.
- Providing processed data to exemplify the usability of the eHealth CSI Dataset.
- Providing a set of data collected from a smartwatch.
- Providing a set of information on the health and physical characteristics of participants in an anonymous format.
- Finally, we present a website that makes all the collected data available to researchers worldwide.

The remainder of the text is organized as follows. In Section II, we present an overview of the CSI technology and the data model. In Section III, we present the publicly available Wi-Fi CSI datasets currently reported in the literature. Our database including the collection scenario, setup, protocol, and participants are described in Section IV. In addition, Section VI describes the website where our dataset is available to the community. Finally, Section VIII presents our final remarks.

II. CSI SYSTEM OVERVIEW

A smart and responsive system, capable of detecting, measuring, and collecting information about the environment and using that information, can be designed using radio frequency (RF) technologies, combined with signal processing and machine learning techniques [7]. One specific communication link that can be used for this purpose is Wi-Fi.

The orthogonal frequency division multiplexing (OFDM) technique is the baseline technology used in IEEE 802.11g/n/ac specification [15], for the physical layer of Wi-Fi systems considering both 2.4GHz and 5GHz frequency bands. OFDM is a multicarrier modulation technique that uses a predefined number of orthogonal subcarriers [16]. Thus, the information is transmitted independently over different subcarriers and OFDM symbols. OFDM features make it a good solution for multipath channels and Multiple-Input Multiple-Output (MIMO) systems [5].

To collect CSI, the Wi-Fi transmitter sends Long Training Fields (LTFs) that contain predefined information. The Wi-Fi receiver estimates the CSI using the received signal and the predefined LTFs. The amount of data collected depends on the channel bandwidth, which determines the number of subcarriers and the number of antennas.

Considering a MIMO Wi-Fi system operating under the IEEE 802.11n specification with m transmitting antennas and n receiving antennas, the signal containing the estimated CSI of each data stream can be mathematically expressed as in Equation 1. $h_{i,j}$ represents the CSI between the i -th transmission antenna and the j -th receiving antenna.

$$H = \begin{pmatrix} h_{1,1} & h_{1,2} & \dots & h_{1,n} \\ h_{2,1} & h_{2,2} & \dots & h_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{m,1} & h_{m,2} & \dots & h_{m,n} \end{pmatrix}. \quad (1)$$

Let c be the number of subcarriers used to estimate CSI, the state information of the channel established between a pair of

antennas (i, j) can be mathematically represented by a vector \mathbf{h} with c elements. Here, we use \mathbf{h} to represent a generic $\mathbf{h}_{i,j}$, in Equation 2, without losing generality.

$$\mathbf{h} = [h_1, h_2, \dots, h_c]^T. \quad (2)$$

CSI data provide information about the environment that can be used to estimate changes and phenomena that occur over time, such as the detection of the presence of humans. Thanks to the use of OFDM techniques and since the transmissions over each subcarrier are independent in OFDM, each data stream can be seen as an independent CSI sensor, which can improve the resolution of the collected information.

III. RELATED WORK

Research interests in using Wi-Fi CSI in human detection, identification, detection, and monitoring applications have increased significantly in the last few years. Some Wi-Fi CSI databases have been created to enable the development of these applications and are available in the literature [17], [18], [19], [20], [21], [22], [23].

Brinke et al. [17] collected data from nine participants over three days. Two of them repeated the collection process over the next three days. The collection comprises 5 or 6 activities (clap, walk, wave, jump, sit, and fall) with 50 trials each. They used 6 antenna pairs (3x2 MIMO), and 30 subcarriers were recorded. The server was a Raspberry Pi with an external hard drive and a 2.4 GHz network was used. The participants had very different physical characteristics. However, this information is missing in the dataset due to participants' denial to share confidential information. All collected data are available online.

Baha et al. [18] made another Wi-Fi CSI dataset available online. That dataset comprises data collected in five positions (falling from a sitting position, falling from a standing position, walking, sitting down, standing up, and picking a pen from the ground) performed by 30 different participants (28 males and 2 females) in three different indoor environments. The collection performed in the first two environments was of a line-of-sight (LOS) nature, whereas the collection performed in the third environment was of a non-line-of-sight (NLOS) nature. In addition, information about participants is available, such as gender, age, weight, and height. For this work, an Intel 5300 NIC was configured to operate in the 2.4 GHz frequency band, wireless channel number 3, channel bandwidth of 20 MHz, and sampling rate of 320 packets/second.

Furthermore, Alazrai et al. [19] considered 66 participants (63 males and 3 females) to perform twelve human-to-human interactions, and have set 40 different pairs of participants. A publicly available CSI tool was used to record Wi-Fi signals transmitted from a commercial standard access point, namely the Sagemcom 2704 access point, to a desktop computer with an Intel 5300 network interface card. The access point was configured to operate at a frequency band of 2.4GHz, wireless channel number 6, channel bandwidth of 20MHz,

and a modulation coding scheme of index 8 was used. The data collected are also available online.

Guo et al. [20] collected data from indoor environments commonly present in daily life: an empty room, a meeting room, and an office. They have also collected activity data at home, in the corridor, and in the laboratory. They collected data from sixteen activities of 10 participants, five male, and five female. Despite having collected data from what they called an empty room, it is worth mentioning that their definition of an empty room differs from the definition used in this paper. For them, it refers to collecting data from a participant in a room without objects other than the equipment used to perform the experiments. In our work, an empty room is one without any people, but containing the same furniture and equipment used during experiments with participants. They used a transmitter with one antenna and a receiver with three antennas. For all conducted experiments in [20], there was always a participant in the room.

Schäfer et al. [21] focused on 8 human activities: empty, lying, sitting, standing, sitting-down, standing-up, walking, and falling. It is worth mentioning that, although they have classified empty as an activity, the meaning of empty is not clear. They have used machine learning strategies in their work, such as long-term short memory (LSTM) and support vector machines (SVM). The Nexmon open source tool was used for CSI extraction on inexpensive hardware (Raspberry Pi 3B+, Pi 4B, and Asus RT-AC86U routers). The frequency bandwidth 20 MHz, 40 MHz, and 80 MHz were used, and data collection was done with a dual-band router Fritzbox (2.4 GHz and 5 GHz). The authors also provided the acquired dataset.

In addition to the already mentioned Wi-Fi CSI datasets focused on human sensing and detection, we also found other datasets created with the aim of locating an object in a room. Such methods work by finding a mapping between a location-dependent feature (called a fingerprint) and the location of the device. In the radio map acquisition phase, the area in which the device needs to be located is surveyed. That is, several locations in space are selected (called Reference Points (RPs)), and the fingerprints recorded at those RPs are stored - together with the coordinates of the RPs - in a database. Gassner et al. [22], for example, provided an automated way to acquire a radio map, using readily available hardware. The method consists of a software-defined radio (SDR) mounted on a wheeled robot. The source-code controlling the robot and the software-defined radio are open-source.

Khorov et al. [23] presented FIND (a tool for Fine INDOor localization). They have used a 4 channel receiver based on NI USRP-2955, capable of capturing real Wi-Fi frames in the 80 MHz band. Retrieve CSI from the 242 subcarriers. It is worth mentioning that the collected data are available online for researchers who want to develop studies related to the direction of arrival (DoA). Thus, there is no collection regarding a participant inside the room, just collections considering different positions of the transmitter.

TABLE 1. Information available in related work.

Ref.	Number of Partic.	Health Info	Positions	Empty room	Ground Truth
[17]	9	N/A	6	No	No
[18]	30	Yes	5	No	No
[19]	66	Yes	12	No	No
[20]	10	Yes	16	No	No
[21]	N/A	N/A	8	No	No
[22]	N/A	N/A	N/A	No	No
[23]	N/A	N/A	N/A	Yes	No
eHealth CSI	118	Yes	17	Yes	Yes

In Table 1, we summarize some basic characteristics of the related literature, such as the number of participants, positions, and ground truth (G.T.) for health measures. Furthermore, we also compare with the information available in our proposed dataset, eHealth CSI.

Clearly labeled Wi-Fi CSI datasets with proper metadata, including basic information about the participants, ground truth health information measures from other equipment, and empty room measures to calibrate the system, are hard to find, as they are often not shared. The ones that are shared often lack documentation or metadata. Furthermore, collecting data from several participants over multiple days is often challenging due to time constraints, participants' availability, and ethics protocols. The eHealth CSI dataset aims to address this gap.

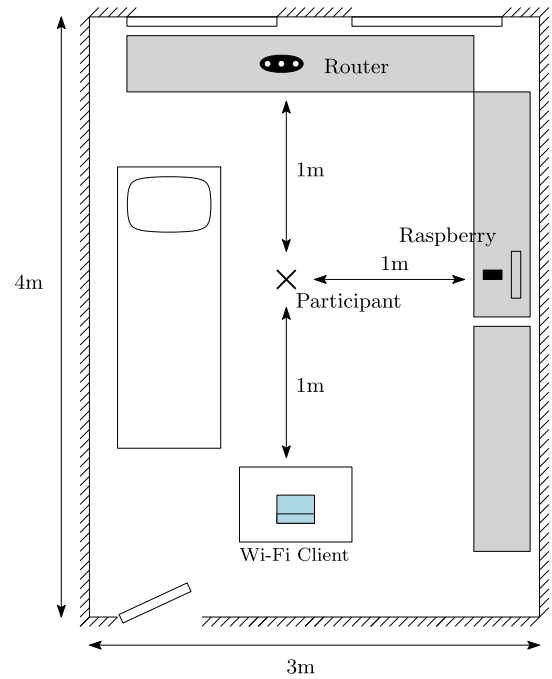
IV. EHEALTH CSI DATASET

With the growing attention to Wi-Fi CSI technology and the lack of a dataset with a large diversity and participants' information to help develop and test new CSI based proposals, it is necessary to build a robust and public dataset of Wi-Fi CSI signals. Therefore, we built a dataset named eHealth CSI dataset with Wi-Fi CSI signals and made it publicly available online. We have defined a data collection protocol with different positions and temporal duration. We recruited volunteers from the academic community. We also collected phenotype information from the participants, such as age, gender, height, and weight. To carry out the procedure, we had the approval of the Research Ethics Committee of the Brazilian Ministry of Health under the CAAE number 54359221.4.0000.5243 and also by the Fluminense Federal University. In the following, we present our work in detail.

A. COLLECTING SCENARIO

The first step of this work consisted of defining a scenario for data collection. The scenario must keep the same conditions for all participants to facilitate data comparison. A 3 x 4m room equipped with three tables and one bed was used. In this room, a Wi-Fi network was set up, with a Wi-Fi router, a laptop client and a Raspberry Pi 4B used as a probe.

The physical setup is depicted in Figure 1. All devices are 1 meter from the participant. The Wi-Fi router and client are on opposite sides of the participant, and the Raspberry Pi is

**FIGURE 1.** Data collection scenario.

equidistant from the router and client, while also 1m away from the participant.

It is worth mentioning that the bed was always in the collection room and for the positions that used the bed, it was placed 1m away from the Raspberry Pi, centered on the participant spot.

B. COLLECTING SETUP

This section presents the configuration parameters of the devices described in Section IV-A. We start with the configuration and parameters of the Wi-Fi network. The Wi-Fi router operates in the 5GHz band, uses the channel 36 within the ISM (industrial, scientific, medical) frequency band [24], and a 80 MHz bandwidth. The 80MHz bandwidth was used to offer a larger amount of CSI data compared to the 40 MHz bandwidth, resulting in greater CSI granularity.

The Wi-Fi client (laptop) connects to the Wi-Fi router on channel 36 in the 5GHz band. The router provides the IP dynamically to the client, and it stays connected, thus forming a Wi-Fi network. The laptop sends a *ping* to the router with a transmission interval of 136ms. We use this transmission interval to estimate events like vital signs (breathing rate and heartbeat rate). Vital signs are observed between 0.2Hz and 3.5Hz. Thus, we need a transmission rate capable of generating an adequate sampling frequency to reconstruct the vital signs and avoid aliasing.

The Raspberry Pi 4B uses the NEXMON firmware [25] to collect CSI data. The settings are the same as for the Wi-Fi network, channel 36 and an 80 MHz bandwidth. The Raspberry Pi has only one transmission/reception antenna. These settings enable the collection of CSI data from the *ping*,

executed by the client. The CSI data from *ping* transmissions are stored in pcap files.

In addition to the information collected from the Wi-Fi signal, all participants wore a smartwatch to collect the heart frequency simultaneously with the CSI collection. The smartwatch model used was Samsung Galaxy Watch 4. The smartwatch data is also available in the dataset, so it can be used as ground truth information to be compared to the estimated measures obtained from the Wi-Fi CSI information.

C. POSITIONS PROTOCOL

In this section, we present the protocol that was followed during the acquisition process. We defined 17 different positions/procedures that we asked the participants to perform during the CSI acquisition. Among them, we can cite: sitting, standing, lying, and walking. Each position is described in Table 2, and an image is placed as a reference.

Before starting, the participant was invited to enter the room without any device that could generate interference with the settings used in the Wi-Fi network. The participant was also required to fill out and sign an authorization term in which he agreed to be part of the research. This procedure is mandatory according to the ethics committee of the Ministry of Health of the Brazilian Government. The participant was then asked to fill out a form indicating some characteristics regarding their physical and clinical status (age, height, weight, preexisting disease, and others). After that, a smartwatch was placed on the participant's wrist, and the participant was left alone in the collecting room to start the collection procedure.

The protocol follows these guidelines:


















- Each position lasts 60s.
- The participant must not move beyond the position instructions.
- Some positions require natural breathing during 60s of collection, and others alternate breathing. The alternate breathing comprises 20s of natural breathing, followed by 10s of holding breath. This alternate breathing cycle is performed twice during the 60s sampling.
- The positions are performed in the location indicated with an X on the floor within the room.

The protocol was explained to each participant. A slide presentation containing instructions for each position and a stopwatch to indicate when to stop and/or change the position was shown during the whole collecting process.

A systematic analysis of the collected data, using signal processing techniques and artificial intelligence, is intended to establish a correlation between the collected CSI data and the metadata, such as the participant's position, phenotype data, and vital signs. As the smartwatch collects the heartbeat rate, it is possible to measure the performance of the CSI data analysis for vital signs detection and future development of remote patient monitoring techniques.

It is worth mentioning that in addition to the participant's data, we also collected CSI data from the empty room before

TABLE 2. Participants' positions protocol.

Position Number	Reference Image	Description
1		Sitting position facing the collector and Wi-Fi devices on each side of the participant.
2		Sitting position in front of the device alternating breathing.
3		Alternating the position of sitting and standing in front of the Raspberry.
4		Position facing away from the collector with the Wi-Fi devices on each side of the participant.
5		Position facing away from the device and alternating breathing.
6		Standing position facing the collector and Wi-Fi devices on each side of the participant.
7		Standing position facing the device alternating breathing.
8		Standing position facing away from the collector and the Wi-Fi devices on each side.
9		Standing position facing away from the device, and alternating breathing.
10		Lying position on the bed with stomach up and sideways to the collector.
11		Lying position on the bed with stomach up and alternating breathing.
12		Lying position on the bed face down and sideways to the collector.
13		Lying position on the bed face down and alternating breathing.
14		Alternating between positions 6 and 10.
15		Walking position (walking in place) facing the collector.
16		Running position (running in place) in front of the collector.
17		Sweeping position (the act of sweeping) in the indicated area.

each participant's data collection. The empty room CSI data can be used as a groundbase of the wireless channel conditions and help to study the impact of human presence in the room as we did in [26].

D. PARTICIPANTS

In this section, we describe the procedure carried out with the participants since their invitation informing the conditions to make their information available to the scientific community anonymously. We started the process with an invitation to collaborate with the project, where some basic information about our research was presented to the participant. Before starting the collection process, the participant was informed about the risks involved and asked to sign a term of consent.

To date, we have collected the CSI signal from 118 participants. Individuals of different ages, gender, weight, and height were invited to collaborate to allow us to acquire a dataset with as much phenotype diversity as possible. The collected signals have been stored anonymously in a database and were made available to the scientific community upon request, subject to controlled and secure access. It should be noted that the data collection procedures to which participants were submitted do not include any type of physical intervention from our team.

The choice of collecting CSI signals from women and men was made to obtain a diverse sample between biological genders and later serve as data to discriminate between them. Another important aspect is age. We recruited people over 18 years old with no maximum age restriction to observe different CSI patterns at different ages. People with different body mass index (BMI) rates were invited to participate too.

To the best of our knowledge, there are no indications in the literature of an adequate phenotype profile for CSI data collecting, or any contraindication, given that all the equipment used is approved by the Brazilian Telecommunications Agency (Anatel), and fulfills all the requirements of Resolution No. 680/2017 on Restricted Radiation Radiocommunication Equipment.

Even knowing that there is no contraindication, we decided to restrict the recruitment of participants to individuals over 18 years old. The exclusion criteria eliminate individuals with infectious diseases and individuals with heart or lung diseases, as these activities would entail a greater risk than the benefit of the research. We also did not consider the collaboration of people with motor disabilities that prevent the execution of the collection protocol.

Participants were recruited through invitations sent by e-mail, instant messaging applications, social networks, and in the University's social environments. The invitation to participate was made in ways that do not allow the identification of the participant or the visualization of their contact data (e.g. email, telephone) by third parties. Any individual invitation sent by email had only one sender and one recipient, or was sent in the form of a hidden list. It should be noted that participants were not paid to collaborate with this research.

In the dataset, there are currently 88 male participants and 30 female participants. The ages of the participants ranged from 18 to 64 years, with an average of 22.38 years and a standard deviation of 11.85 years. The age distribution of the participants is depicted in Figure 2.

The participant's height ranged from 152 to 198cm, with an average of 173.42cm and a standard deviation of 8.89cm. Figure 3 shows the height distribution of the participants.

The participant's weight ranged from 40 to 116 kg, with an average of 72.79 kg and a standard deviation of 15.96 kg. Figure 4 shows the weight distribution of the participants. Given the height and weight of the participants, the BMI ranged from 14.34 to 42.97, with an average of 24.10 and a standard deviation of 4.54.

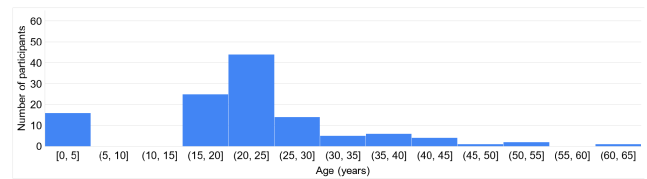


FIGURE 2. Participants' age distribution.

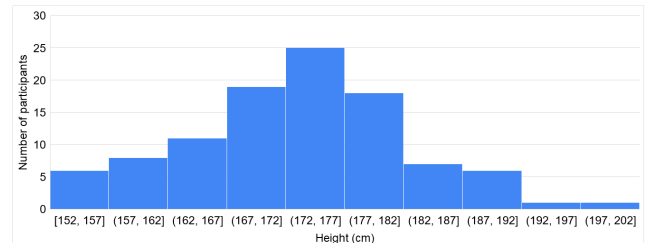


FIGURE 3. Participants' height distribution.

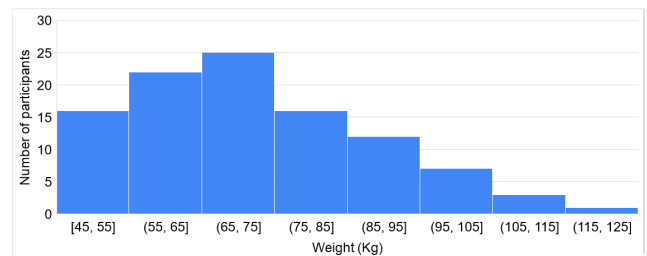


FIGURE 4. Participants' weight distribution.

The dataset also contains information about the participant's health record, such as respiratory and cardiovascular disease, smoking status, and whether he/she has any prostheses.

V. APPLICATIONS

The dataset offers information that can be used for different applications. Using the CSI data, the smartwatch data, and the participants' phenotype data, we can categorize at least three sorts of applications:

- Identification of human activities.
- Health monitoring.
- Identification of physical characteristics of the participants.

Using the set of positions and data from the empty room, it is possible to perform a CSI analysis and determine the presence, motion, and activity of humans. In the literature, several studies emphasize using CSI as a technology accessible for monitoring human activities [3], [4], [6], [7].

In addition to respiration rate and heartbeat rate, some studies proposed new sensing modalities of human activities such as detecting changes in position, micro-movements, tremors, and falls. The WiSleep [27] proposal, for example, focused on extracting rhythmic patterns from the CSI associated with respiration and abrupt changes due to body movement.

The WiSleep proposal was further extended in [28]. Compared to existing works, the system proposed in [28] can track abnormal respiration (e.g., sleep apnea). It can also provide information on respiration when the person is in different sleeping positions.

Phenotype data, such as age, gender, weight, and height, can also be used in applications that focus on biometric identifications and characteristics of participants. The aforementioned applications and many others can take advantage of the dataset provided in this work to enable the development of this and other surveys that may arise over time.

In this paper, we provide two application examples. One is the extraction of vital signs. The heartbeat rate and respiration rate estimates are provided in a dashboard used to access the dataset. Another application is detection of human presence.

VI. VITAL SIGN DASHBOARD DESCRIPTION

Data collected from all 118 participants are available online on our homepage [29]. We provide labeled Wi-Fi CSI dataset applicable to different research areas for free so that researchers can use them to carry out more effective studies and have comparable results. In this section, we present a detailed description of our homepage and the dataset, which is available to download.

The Wi-Fi CSI data are stored in a database linked to each participant's personal information. A number identifies each participant to maintain privacy.

The dashboard summarizes information about the data that can be downloaded. Some available metrics include the number of male and female participants and the average age. Figure 5 shows the main screen of the dashboard.

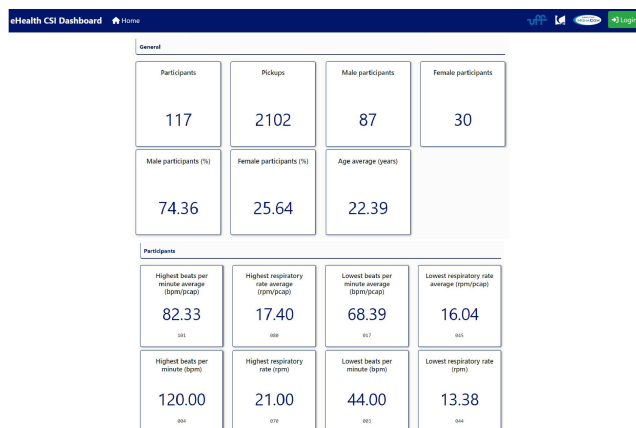


FIGURE 5. Dashboard main screen.

For those who want access to the available dataset, it is necessary to fill out an access request form. Once the user receives its credentials, it is possible to download the database. It is worth mentioning that the available dataset consists of raw data (i.e. without signal processing).

In addition to the raw data, we show some applications that can use the collected data. The dashboard also presents the vital signs of the participants (breathing rate and

heartbeat rate) for each position in the collection protocol. Two graphs are assigned to each participant, one for heartbeat rate and the other for respiratory rate. In Figure 6, we present the graphs for one participant.

The graph's X-axis corresponds to each position defined by the collection protocol, and the Y-axis corresponds to the measure of the vital sign. The heartbeats per minute graph has 2 points per position. The filled dots represent the data detected through Wi-Fi CSI analysis, and the opaque dots represent the data collected via the smartwatch.

Furthermore, the available information can be sorted according to the user's choice. The user has the option to add filters for one or several parameters. This option allows comparisons and analysis of selected groups of participants with pre-defined characteristics. It is possible to filter participant collections using his/her identification number, position, gender, height, weight, age, body mass index, beats per minute, and respiratory rate.

There are four different ways to present the information in the dashboard: listing, average, curves by participants, and report. If the user chooses the listing option, it will present a list of all participants, a heartbeat rate graph, and a breathing rate graph assigned to each participant. If he chooses the average option, it will present the average of the vital signs of each participant. The curve per participant option will display all the heartbeat rate graphs and the respiration rate graphs related to each participant. The report option presents a report with all the information about participants.

VII. PRESENCE DETECTION

In this section, we present another possible application of the proposed dataset. We explain the general framework of WiFi-based human presence detection and present the signal processing tools used to evaluate the system's performance. Finally, we analyze the performance of the presented dataset on human presence detection to exemplify its applicability. An initial work on presence detection using a subset of our dataset was previously published in [26].

A. DATA TREATMENT

Several treatment techniques were applied to the collected data to improve the accuracy of the obtained results as much as possible. This section addresses step-by-step the procedure used to detect human presence, such as signal preprocessing, the distance between two time series, the definition of unbalanced and balanced datasets, and classification analysis.

1) PRE-PROCESSING

The collected CSI signal contains noise and outliers. To improve the signal-to-noise ratio, we applied some filtering techniques. The Hampel filter was used to remove noise. The Moving Average filter was used to remove outlier measures and smooth the ripple. After that, we obtain the amplitude of each signal. As each collection has 234 subcarriers, we have the amplitude of those subcarriers throughout 60s that collection in each position lasted.

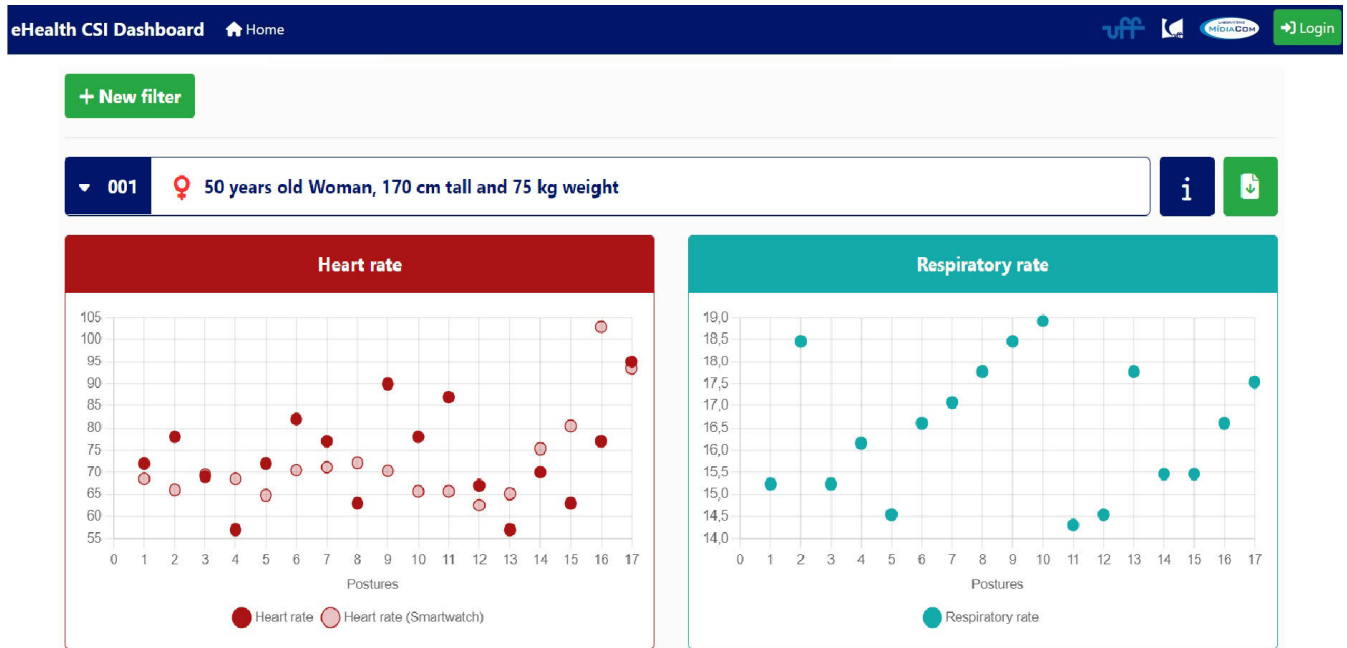


FIGURE 6. Processed CSI information display screen for one participant.

2) DISTANCE BETWEEN SIGNALS

The chosen technique was Dynamic Time Warping (DTW), an algorithm that compares the similarity and calculates the distance between two time series. The data collected from an empty room was considered a reference in this work. The DTW between the reference signal and the signal corresponding to each instance was then calculated, and this similarity was used as a feature.

The reference empty room was chosen randomly since they all presented the same pattern. Then, it was compared with each other collection using DTW.

Each DTW feature was calculated between the equivalent subcarrier of each collection, that is, the first reference empty room subcarrier versus the first collection subcarrier of a filled room or empty room. After comparison, each collection returned 17 instances with 234 features, as we have 17 different positions. Thus, we have 1,700 instances, as we considered 100 participants for this analysis. We reserved the other available participants' data for further testing as will be detailed later on this section. The data for empty rooms were processed in the same way.

3) UNBALANCED AND BALANCED DATASETS

The dataset comprises 1,700 instances representing data from filled rooms. Our dataset was initially unbalanced. It had a total of 1,700 instances, with 100 instances only representing 100 empty rooms, one for each participant. The majority class was the filled room, while the minority class was the empty room.

To avoid a majority class, we collected more data from empty rooms, reaching a total of 1,700 instances for building a balanced dataset.

4) CLASSIFICATION ANALYSIS

We used different algorithms such as Naive Bayes, J48, Support Vector Machine (SVM) and Random Forest [30], [31] to classify empty and filled rooms. For Naive Bayes, the default parameters were used without a priori class specification. For the J48 algorithm, an entropy criterion and a tree with a depth of 3 were used. In SVM, we used a linear kernel, thus following a linear model. For Random Forest, 100 estimators were left, which is the default parameter in the Python library used for implementation. In this analysis, all algorithms received 70% of the data for training and 30% for testing.

B. OBTAINED RESULTS

In this section, we present the results obtained from the data of 100 participants available in eHealth CSI dataset. We consider the room to be filled, regardless of the position and/or the activities performed.

For the obtained results, all datasets were balanced. Table 3 summarizes the results achieved for each classification model: SVM, J48, Naive Bayes, and Random Forest. As can be seen, with classifiers and a balanced dataset, we obtained an accuracy of 99.9%, for SVM, 93.43% when applying the Naive Bayes algorithm, 94.90% in the case of J48, and 99.9% in the case of Random Forest.

The results obtained show good accuracy for the human presence classification when using a balanced dataset.

TABLE 3. Results for balanced data (%).

Algorithm	Accuracy	Precision	Recall	F-Measure
SVM	99.90	100.0	99.80	99.90
J48	94.90	92.57	97.65	95.04
Naive Bayes	93.43	99.78	87.06	92.98
Random Forest	99.90	100.0	99.80	99.90

TABLE 4. Results for test.

Algorithm	Accuracy	Precision	Recall	F-Measure
SVM	76.47	72.13	86.27	78.57
J48	78.92	81.72	74.51	77.95
Naive Bayes	83.33	100.0	66.67	80.00
Random Forest	91.18	98.34	83.82	90.48

These results are obtained using all 234 features calculated for each subcarrier. As we can see, the application of the data set to human presence detection is satisfactory. It attests the effectiveness of the work carried out as an application using CSI data. We also carried out a test of the application made. Here, we consider information that was not part of the initial training set. We saved data from 18 people to evaluate the trained model on never seen data. Table 4 summarizes the results achieved. As can be seen, the results declined considerably, with the best result being 91.18% using Random Forest. This decrease in accuracy is due to submitting the human presence detection proposal to a real test. It also opens the way to improve the proposal and open up new approaches to presence detection.

VIII. CONCLUSION

In this work, we present a public Wi-Fi CSI dataset, named eHealth CSI, of human activities with 17 different positions performed by 118 participants, in an indoor environment. In addition, we also provide vital sign data from all participants collected by a smartwatch during experiments, as well as empty room Wi-Fi CSI data. The aim was to provide a huge and diverse dataset to the research community to promote the development of research related to detecting human activities, presence, vital signs, and others, using Wi-Fi signals.

As future work, we will continue to run experiments with new participants and the size of our dataset will increase as the number of collection increases. We also intend to work on different applications using our dataset, such as human identity detection, human activities recognition, apnea detection, among others.

ETHICS STATEMENT

The experimental procedure performed has been previously approved by the Ethics Committee of the Brazilian Ministry of Health under the CAAE number 54359221.4.0000.5243 and also by the Fluminense Federal University. Before performing any of the experiments, each participant was asked to sign a consent form in which they were informed that their personal information would not be

disclosed and that they have the right to stop participating in any of the experiments at any time if they choose to do so.

REFERENCES

- [1] Y. Gu, J. Zhan, Y. Ji, J. Li, F. Ren, and S. Gao, "MoSense: An RF-based motion detection system via off-the-shelf WiFi devices," *IEEE Internet Things J.*, vol. 4, no. 6, pp. 2326–2341, Dec. 2017.
- [2] Y. Gu, T. Liu, J. Li, F. Ren, Z. Liu, X. Wang, and P. Li, "EmoSense: Data-driven emotion sensing via off-the-shelf WiFi devices," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–6.
- [3] Z. Wang, K. Jiang, Y. Hou, W. Dou, C. Zhang, Z. Huang, and Y. Guo, "A survey on human behavior recognition using channel state information," *IEEE Access*, vol. 7, pp. 155986–156024, 2019.
- [4] J. Liu, H. Liu, Y. Chen, Y. Wang, and C. Wang, "Wireless sensing for human activity: A survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1629–1645, 3rd Quart., 2020.
- [5] S. Lee, Y. D. Park, Y. J. Suh, and S. Jeon, "Design and implementation of monitoring system for breathing and heart rate pattern using WiFi signals," in *Proc. IEEE Annu. Consum. Commun. Netw. Conf.*, Jan. 2018, pp. 1–7.
- [6] Y. Ma, G. Zhou, and S. Wang, "WiFi sensing with channel state information: A survey," *ACM Comput. Surv.*, vol. 52, no. 3, 2019.
- [7] S. Yousefi, H. Narui, S. Dayal, S. Ermon, and S. Valaee, "A survey on behavior recognition using WiFi channel state information," *IEEE Commun. Mag.*, vol. 55, no. 10, pp. 98–104, Oct. 2017.
- [8] A. Uchiyama, S. Saruwatari, T. Maekawa, K. Ohara, and T. Higashino, "Context recognition by wireless sensing: A comprehensive survey," *J. Inf. Process.*, vol. 29, pp. 46–57, Jan. 2021.
- [9] J. Xiao, Z. Zhou, Y. Yi, and L. M. Ni, "A survey on wireless indoor localization from the device perspective," *ACM Comput. Surv.*, vol. 49, no. 2, pp. 1–31, Jun. 2017.
- [10] Z. Yang, Z. Zhou, and Y. Liu, "From RSSI to CSI: Indoor localization via channel response," *ACM Comput. Surv.*, vol. 46, no. 2, pp. 1–32, Nov. 2013.
- [11] J. C. H. Soto, I. Galdino, E. Caballero, V. Ferreira, D. Muchaluat-Saade, and C. Albuquerque, "A survey on vital signs monitoring based on Wi-Fi CSI data," *Comput. Commun.*, vol. 195, pp. 99–110, Nov. 2022.
- [12] M. Zhao, F. Adib, and D. Katabi, "Emotion recognition using wireless signals," *Commun. ACM*, vol. 61, no. 9, pp. 91–100, Aug. 2018.
- [13] S. F. Khan, "Health care monitoring system in Internet of Things (IoT) by using RFID," in *Proc. 6th Int. Conf. Ind. Technol. Manage. (ICITM)*, Mar. 2017, pp. 198–204.
- [14] N. Damodaran, E. Haruni, M. Kokhharova, and J. Schäfer, "Device free human activity and fall recognition using WiFi channel state information (CSI)," *CCF Trans. Pervasive Comput. Interact.*, vol. 2, no. 1, pp. 1–17, Mar. 2020.
- [15] *IEEE Standard for Information Technology*, Standard IEEE 802.11 Working Group, IEEE 802.11ac-2013E, Tech. Rep., 2013. [Online]. Available: https://standards.ieee.org/standard/802_11ac-2013.html
- [16] S. Weinstein and P. Ebert, "Data transmission by frequency-division multiplexing using the discrete Fourier transform," *IEEE Trans. Commun. Technol.*, vol. COM-19, no. 5, pp. 628–634, Oct. 1971.
- [17] J. K. Brinke and N. Meratnia, "Dataset: Channel state information for different activities, participants and days," in *Proc. 2nd Workshop Data Acquisition Anal.*, Nov. 2019, pp. 61–64.
- [18] B. A. Alsaify, M. M. Almazari, R. Alazrai, and M. I. Daoud, "A dataset for Wi-Fi-based human activity recognition in line-of-sight and non-line-of-sight indoor environments," *Data Brief*, vol. 33, Dec. 2020, Art. no. 106534.
- [19] R. Alazrai, A. Awad, B. Alsaify, M. Hababeh, and M. I. Daoud, "A dataset for Wi-Fi-based human-to-human interaction recognition," *Data Brief*, vol. 31, Aug. 2020, Art. no. 105668.
- [20] L. Guo, L. Wang, C. Lin, J. Liu, B. Lu, J. Fang, Z. Liu, Z. Shan, J. Yang, and S. Guo, "WiAR: A public dataset for Wi-Fi-based activity recognition," *IEEE Access*, vol. 7, pp. 154935–154945, 2019.
- [21] J. Schäfer, B. R. Bariswal, M. Kokhharova, H. Adil, and J. Liebehenschel, "Human activity recognition using CSI information with nexmon," *Appl. Sci.*, vol. 11, no. 19, p. 8860, Sep. 2021.
- [22] A. Gassner, C. Musat, A. Rusu, and A. Burg, "OpenCSI: An open-source dataset for indoor localization using CSI-based fingerprinting," 2021, *arXiv:2104.07963*.

- [23] E. Khorov, A. Kureev, and V. Molodtsov, "FIND: An SDR-based tool for fine indoor localization," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, May 2021, pp. 1–2.
- [24] K. Pahlavan, T. H. Probert, and M. E. Chase, "Trends in local wireless networks," *IEEE Commun. Mag.*, vol. 33, no. 3, pp. 88–95, Mar. 1995.
- [25] M. Schulz, J. Link, F. Gringoli, and M. Hollick, "Shadow Wi-Fi: Teaching smartphones to transmit raw signals and to extract channel state information to implement practical covert channels over Wi-Fi," in *Proc. 16th Annu. Int. Conf. Mobile Syst., Appl., Services*, Jun. 2018, pp. 256–268.
- [26] J. C. H. Soto, I. Galdino, B. G. Gouveia, E. Caballero, V. Ferreira, D. Muchaluat-Saade, and C. Albuquerque, "Wi-Fi CSI-based human presence detection using DTW features and machine learning," in *Proc. IEEE Latin-Amer. Conf. Commun. (LATINCOM)*, Nov. 2022, pp. 1–6.
- [27] X. Liu, J. Cao, S. Tang, and J. Wen, "Wi-sleep: Contactless sleep monitoring via WiFi signals," in *Proc. IEEE Real-Time Syst. Symp.*, Dec. 2014, pp. 346–355.
- [28] X. Liu, J. Cao, S. Tang, J. Wen, and P. Guo, "Contactless respiration monitoring via off-the-shelf WiFi devices," *IEEE Trans. Mobile Comput.*, vol. 15, no. 10, pp. 2466–2479, Oct. 2016.
- [29] *CSI MidiaCom*. Accessed: Jan. 6, 2023. [Online]. Available: <http://csi.midiacom.uff.br/dashboard/>
- [30] B. Baharudin, L. H. Lee, and K. Khan, "A review of machine learning algorithms for text-documents classification," *J. Adv. Inf. Technol.*, vol. 1, no. 1, pp. 4–20, Feb. 2010.
- [31] F. Osisanwo, J. Akinsola, O. Awodele, J. Hinmikaiye, O. Olakanmi, and J. Akinjobi, "Supervised machine learning algorithms: Classification and comparison," *Int. J. Comput. Trends Technol.*, vol. 48, no. 3, pp. 128–138, Jun. 2017.



IANDRA GALDINO received the Engineering degree in telecommunications engineering from Universidade Federal Fluminense (UFF), Niterói, Rio de Janeiro, Brazil, in 2015, the M.Sc. degree in electrical engineering from the Federal University of Rio de Janeiro (COPPE/UFRJ), in 2017, and the Ph.D. degree in electrical engineering from COPPE/UFRJ and the CEDRIC/LAETITIA Research Laboratory, Conservatoire National des Arts et Métiers (CNAM), Paris, France. She is currently a Postdoctoral Fellow with the MidiaCom Laboratory, Institute of Computing, UFF. Her areas of interest include communication systems, multicarrier waveforms, and signal processing. Her current research interests include signal processing applied to health and waveform design for post 5G systems.



JULIO C. H. SOTO received the degree in systems engineering from Universidad Católica de Santa Maria (UCSM), in 2009, and the master's degree in informatics from the Federal University of Paraná (UFPR), in 2012. He is currently pursuing the Ph.D. degree in computing with the Institute of Computing, Universidade Federal Fluminense (UFF). He is also a member of the MidiaCom Laboratory, UFF. He has experience in the field of systems engineering, with an emphasis on computer networks and wireless systems. His research interests include wireless networks, security, cognitive radio, wireless body area networks, and wireless health networks.



EGBERTO CABALLERO received the bachelor's degree in telecommunications and electronics engineering and the first master's degree in telecommunication systems from Universidad de Oriente (UO), Cuba, in 2009 and 2012, respectively, and the second master's degree in computer systems from Universidade Federal Fluminense (UFF), Brazil, in 2020, where he is currently pursuing the Ph.D. degree in computer science with the Institute of Computing. He is currently a member of the MidiaCom Laboratory, Institute of Computing, UFF. His research interests include telecommunications and computer systems, especially computer networks, wireless networks, telecommunications networks, machine learning, and e-health.



VINICIUS FERREIRA (Member, IEEE) received the bachelor's degree in telecommunications engineering, the M.Sc. degree in electrical and telecommunications engineering, and the Ph.D. degree in computer science from Universidade Federal Fluminense (UFF), in 2013, 2016, and 2021, respectively. He is currently a Researcher of the Algoritmi Center, University of Minho and a Collaborator of the MidiaCom Laboratory, UFF. His research interests include computer networks, wireless networks, the Internet of Things, and e-health.



TAIANE COELHO RAMOS received the Ph.D. degree in computer science from the University of São Paulo (USP), in 2021. She is currently an Adjunct Professor with the Department of Computer Science, Universidade Federal Fluminense (UFF). She is also a member of the MidiaCom Research Laboratory. Her main research interest includes the use of machine learning for biological signal analysis.



CÉLIO ALBUQUERQUE received the B.S. and first M.S. degrees in electrical and electronics engineering from Universidade Federal do Rio de Janeiro, Brazil, in 1993 and 1995, respectively, and the second M.S. and Ph.D. degrees in information and computer science from the University of California at Irvine, in 1997 and 2000, respectively. From 2000 to 2003, he was a Networking Architect with Magis Networks, designed high-speed wireless medium access control. Since 2004, he has been an Associate Professor with the Computer Science Department, Universidade Federal Fluminense, Brazil. His research interests include wireless networks, network security, smart grid communications, internet architectures and protocols, and multicast and multimedia services.



DÉBORA C. MUCHALUAT-SAADE (Member, IEEE) received the Ph.D. degree in computer science from the Pontifical Catholic University of Rio de Janeiro (PUC-Rio), in 2003. She is currently a Full Professor with the Department of Computer Science, Universidade Federal Fluminense (UFF), Brazil. She is also one of the founders and the heads of the MidiaCom Research Laboratory. She has contributed to the design and development of the Ginga-NCL Middleware, used in the Brazilian digital TV standard and IPTV services. Her main research interests include multimedia systems, computer networks, the IoT, smart grids, and e-health.

...