Received October 12, 2020, accepted October 26, 2020, date of publication October 29, 2020, date of current version November 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3034849

An End-to-End Deep Learning Framework for Recognizing Human-to-Human Interactions Using Wi-Fi Signals

RAMI ALAZRAI[1], (Member, IEEE), MOHAMMAD HABABEH[1], BAHA' A. ALSAIFY[2], MOSTAFA Z. ALI[3], (Senior Member, IEEE), AND MOHAMMAD I. DAOUD[1]

[1]Department of Computer Engineering, School of Electrical Engineering and Information Technology, German Jordanian University, Amman 11180, Jordan
[2]Department of Network Engineering and Security, Jordan University of Science and Technology, Irbid 22110, Jordan
[3]Department of Computer Information Systems, Jordan University of Science and Technology, Irbid 22110, Jordan

Corresponding author: Rami Alazrai (rami.azrai@gju.edu.jo)

This work was supported by the Seed-Grant Program at the German Jordanian University under Grant SEEIT 02/2020.

ABSTRACT Channel state information (CSI)-based human activity recognition plays an essential role in various application domains, such as security, healthcare, and Internet of Things. Most existing CSI-based activity recognition approaches rely on manually designed features that are classified using traditional classification methods. Furthermore, the use of deep learning methods for CSI-based activity recognition is still at its infancy with most of the existing approaches focus on recognizing single-human activities. The current study explores the feasibility of utilizing deep learning methods to recognize human-to-human interactions (HHIs) using CSI signals. Particularly, we introduce an end-to-end deep learning framework that comprises three phases, which are the input, feature extraction, and recognition phases. The input phase converts the raw CSI signals into CSI images that comprise time, frequency, and spatial information. In the feature extraction phase, a novel convolutional neural network (CNN) is designed to automatically extract deep features from the CSI images. Finally, the extracted features are fed to the recognition phase to identify the class of the HHI associated with each CSI image. The performance of our proposed framework is assessed using a publicly available CSI dataset that was acquired from 40 different pairs of subjects while performing 13 HHIs. Our proposed framework achieved an average recognition accuracy of 86.3% across all HHIs. Moreover, the experiments indicate that our proposed framework enabled significant improvements over the results achieved using three state-of-the-art pre-trained CNNs as well as the results obtained using four different conventional classifiers that employs traditional handcrafted features.

INDEX TERMS Two-person interaction, channel state information (CSI), human activity recognition, convolutional neural networks (CNNs), Wi-Fi, deep learning.

I. INTRODUCTION

Human activity recognition has many applications in several domains [1], such as human-computer interaction, security and surveillance, and healthcare. Traditional human activity recognition approaches employ different sensing technologies, such as cameras [2], wearable sensors [3], and radars [4]. Despite the favorable results achieved by the traditional approaches, there are several factors that can limit their performance. For example, camera-based approaches can be affected by the illumination condition, the camera view-angle, and the level of occlusions [5], [6]. Besides, camera-based approaches are considered privacy invasion systems [7]. The approaches that rely on wearable sensors, such as gyroscopes and accelerometers, require the individuals to wear sensing devices all the time to monitor their motions, which might be obstructive and impractical for the users [5], [8], [9]. Radar-based approaches utilize special devices to acquire the signals, such as universal software radio peripheral [10], [11], that have a limited range of coverage.

Recently, researchers have shown that human activities can affect the characteristics of the pervasive Wi-Fi signals in indoor environments [12]–[15]. Therefore, human

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

activities can be recognized by analyzing the variations in the Wi-Fi signals surrounding the users. In fact, the use of Wi-Fi signals can alleviate the limitations associated with traditional sensing technologies [1], [16]. This is due to the following attractive properties of the Wi-Fi signals. First, Wi-Fi signals have a better range of coverage compared with traditional sensing technologies, such as cameras, wearable sensors, and radars. Second, Wi-Fi signals have a noninvasive nature that preserves the privacy of the users. Third, human activity recognition approaches that rely on Wi-Fi signals are device-free approaches that do not require the users to wear any sensing devices.

Literature reveals that the channel state information (CSI), which is a fine-grained metric that captures the variations in the amplitude and phase information associated with different subcarrier frequencies of a Wi-Fi channel, has been widely used to develop human activity recognition systems [5], [16], [17]. The majority of the existing CSI-based human activity recognition approaches rely on extracting handcrafted features from the CSI signals using various signal processing techniques [18]. These handcrafted features are processed using a classifier, such as hidden Markov model (HMM) and support vector machine (SVM), to identify the human activity associated with the CSI signals. Notwithstanding the favorable results achieved using the handcrafted features, the task of manually designing new features that describe the information encapsulated in the time, frequency, and spatial domains of the CSI signals is considered challenging [7], [16], [19].

To avoid the process of manually designing features, researchers have utilized deep learning (DL) methods, such as the convolutional neural networks (CNNs), to learn deep features from the input signals. In fact, DL methods have been successfully employed in several fields, such as signal and image classification [20] and computer vision [21], [22]. Motivated by the great success of DL methods in different fields, researchers have recently started to explore the feasibility of utilizing DL methods to develop CSI-based human activity recognition approaches [7], [16], [17], [19], [23]–[25]. The results obtained by these approaches indicate that the use of DL methods has significantly improved the recognition accuracy compared to other human activity recognition approaches that utilize manually designed features [7], [16], [19].

In spite of the remarkable results achieved by the existing CSI-based human activity recognition approaches, these approaches were mainly focused on recognizing single-human activities that are performed by one individual. This can limit the potentials of using these approaches in real-world scenarios that involve more than one human [7], [16], [17]. In this regard, previous studies have indicated that the problem of recognizing human-to-human interactions (HHIs), which involve two interacting humans (e.g., handshaking and hugging interactions), is considered more challenging than the problem of recognizing single-human activities (e.g., walking and falling activities) [26], [27]. This stems from the following factors. First, HHIs involve causal relationships and interdependencies among the moving body-parts of the two interacting humans. Second, HHIs comprise large inter- and intra-personal variations in the performed interactions. Third, different HHIs may involve similar movements that are performed by the two interacting humans.

In light of this, the current study proposes an end-to-end deep learning framework (E2EDLF) for recognizing HHIs using CSI signals. The proposed framework comprises three phases, which are the input, feature extraction, and recognition phases. The input phase converts the raw CSI signals into a set of two-dimensional (2D) gray-scale CSI images that comprise the time, frequency, and spatial information encapsulated in the raw CSI data. The feature extraction and recognition phases are implemented using a novel CNN architecture that comprises three blocks of layers. Particularly, in the feature extraction phase, the first two blocks of layers within our proposed CNN architecture are utilized to automatically extract data-driven features from the CSI images. Specifically, the first block of layers extracts joint time-frequency features from the CSI signals associated with each transmit-receive pair of antennas, while the second block of layers extracts spatial features from the different pairs of transmit-receive antennas. In the recognition phase, the joint time, frequency, and spatial features, which are extracted at the feature extraction phase, are fed to the third block of layers within our proposed CNN architecture to recognize the performed HHI within each CSI image.

The performance of our proposed E2EDLF is assessed using a publicly available CSI dataset [28] that was introduced by our research group. This dataset contains the raw CSI signals recorded for 40 different pairs of subjects while performing 13 HHIs. Moreover, we compare the results obtained by our proposed E2EDLF with the results obtained using three state-of-the-art pre-trained CNNs. Besides, we compare the results achieved by our proposed E2EDLF with the results achieved by traditional handcrafted features that are extracted from the CSI signals and classified using four different conventional classifiers, including a multi-class support vector machine (mcSVM) classifier, k-NN classifier, naive Bayes classifier, and decision tree classifier. In fact, the results indicate that the performance of our proposed E2EDLF outperforms the performances achieved using the pre-trained CNNs and the traditional handcrafted features, respectively. Moreover, the results provided in our study demonstrate the feasibility of recognizing HHIs by analyzing the CSI signals using DL technology.

The main contributions of the current study can be summarized as follows:

- For the first time, this study investigates the possibility of recognizing HHIs by analyzing CSI signals.
- We propose a novel E2EDLF for recognizing HHIs that can extract features from the time, frequency, and spatial domains of the CSI signals. To the best of our knowledge, this is the first study that explores the feasibility

of utilizing CNNs to learn features from the time, frequency, and spatial domains of the CSI signals with the goal of distinguishing between HHIs.
- Extensive experiments are performed using our publicly available CSI dataset to demonstrate the capability of the proposed framework for recognizing HHIs.

The remainder of this paper is structured as follows. Section II provides a review about the previous studies that were conducted in the field of CSI-based human activity recognition. Section III provides a background knowledge about the CSI of a Wi-Fi system. Section IV presents our proposed E2EDLF for recognizing HHIs. Section V describes the publicly available CSI dataset employed in the current study, presents the experimental results, and discusses the performance of the proposed framework. Finally, the conclusion is provided in Section VI.

II. RELATED WORK

Over the past few years, researchers have proposed numerous CSI-based approaches for recognizing human activities. These approaches can be generally grouped into two categories, namely fine-grained and coarse-grained human activity recognition approaches.

A. FINE-GRAINED ACTIVITY RECOGNITION APPROACHES

The approaches within this category focus on recognizing primitive movements of human body-parts that are in the range of millimeters [7], [18], [29], such as keystroke recognition, vital sign monitoring, and gesture recognition.

In this regard, researchers have recently explored the possibility of recognizing keystrokes using CSI signals. For example, Ali *et al.* [30] proposed a CSI-based keystroke recognition system called WiKey. Particularly, WiKey can distinguish between the CSI variants that correspond to different hands and fingers movements of a user while pressing various buttons on the keyboard. Li *et al.* [31] introduced a CSI-based approach that can infer keystrokes on a mobile device.

Another group of researchers has utilized the CSI signals to track human's vital signs. For instance, Liu *et al.* [32] proposed a CSI-based system for sleep monitoring called Wi-Sleep. The Wi-Sleep system analyzes the CSI values to extract sleep-related information, such as the respiration of the user and sleeping postures. Liu *et al.* [33] developed a system that can track human vital signs, such as heart rates and breathing, using CSI signals. Niu *et al.* [29] utilized the CSI signals to detect the human respiration rate. Zhao *et al.* [34] employed the CSI signals to detect the heartbeats of individuals. The detected heartbeats are used to infer the user's emotional state.

Others studies were focused on recognizing hand and finger gestures using CSI signals. In this vein, Abdelnasser *et al.* [35] proposed WiGest, which employs the CSI signals to recognize hand gestures. Li *et al.* [36] introduced WiFinger, which is a CSI-based system that can recognize finger gestures. Pu *et al.* [10] presented WiSee, which utilizes the CSI signals to recognize hand gestures.

B. COARSE-GRAINED ACTIVITY RECOGNITION APPROACHES

The approaches within this category focus on recognizing single-human activities that involve movements of different body-parts, such as walking, running, and falling. Coarse-grained human activity recognition approaches can be generally organized into two groups as per the employed classification schemes [18]: conventional approaches and deep leaning-based approaches.

Conventional approaches rely on manually designing and extracting features from the time and frequency domains of the CSI signals. Then, the manually extracted features are used to construct standard classification models, such as the SVM classifier, to recognize human activities. For instance, Wang *et al.* [1] developed CARM, which is a CSI-based system that can recognize nine daily human activities. The proposed system comprises two models, namely the CSI-speed model and the CSI-activity model, that are used to quantify the correlation between the CSI dynamics and a particular human activity. Palpana *et al.* [37] proposed a CSI-based fall detection system called FallDeFi. The FallDeFi system utilizes the short-time Fourier transform to extract time-frequency features from the CSI measurements. The extracted features were used to construct a SVM classifier that can recognize the following four types of falls: loss of balance, tripping, loss of consciousness, and slipping. Wang *et al.* [38] designed a CSI-based location-oriented activity recognition system called E-eyes. Particularly, the E-eyes system utilizes a moving variance thresholding method to distinguish between walking activity and nine in-place daily human activities. Then, human activities are recognized using a matching algorithm that computes the similarity between the CSI measurements and a set of pre-constructed activity profiles. Xiao *et al.* [39] proposed SEARE, which is a CSI-based system that can recognize exercise activities. Specifically, the SEARE system extracts features from the time and frequency domains of the CSI measurements, and then utilizes the dynamic time warping technique to quantify the distance between feature vectors to recognize the performed exercise.

In contrast to the conventional approaches, DL-based approaches can automatically extract latent features from the CSI measurements, which can minimize the necessity to manually design the features. Recently, researchers have started to explore the possibility of utilizing DL methods to develop CSI-based human activity recognition approaches [18]. In this vein, Yousefi *et al.* [16] proposed a CSI-based DL approach that utilizes a long short-term memory (LSTM) network to recognize six daily human activities. Feng et al [7] presented a DL approach for human activity recognition that is based on LSTM networks. The approach can automatically extract time and frequency features from the raw CSI signals to recognize three types of human activities. Sheng *et al.* [19] presented a DL approach for activity recognition that can automatically learn temporal-spatial features from the CSI data. The approach integrates the spatial features extracted using a CNN into the temporal model which is realized using a bidirectional LSTM network. In another study, Gao et al [25] converted the CSI measurements associated with multiple channels into radio images and employed a sparse auto-encoder to extract deep optimized features from the radio images and recognize human activities.

The aforementioned studies indicate that the use of DL methods have obtained remarkable performance improvements compared with conventional classification methods that utilize manually designed features [16], [19]. Nonetheless, the use of DL methods to analyze the CSI signals and recognize human activities is still at its early stages with most of the existing approaches focus on recognizing single-human activities. Having that said, our work contributes to the continuing studies in the area of CSI-based human activity recognition by presenting a novel DL framework that can automatically extract effective features from the time, frequency, and spatial domains of the CSI signals and recognize thirteen HHIs with high accuracy.

III. BACKGROUND OF CHANNEL STATE INFORMATION

Commercial off-the-shelf Wi-Fi devices that run according to the IEEE 802.11n standard utilize the multiple-input multiple-output (MIMO) technology with the orthogonal frequency-division multiplexing (OFDM) scheme to send and receive different Wi-Fi signals over multiple transmit-receive antenna pairs [18]. Specifically, the OFDM scheme divides the bandwidth of a MIMO channel into a set of orthogonal subcarrier frequencies that are transmitted in parallel [12]. The propagation of wireless signals between a transmit-receive antenna pair is characterized by the CSI metric [16], which represents the channel frequency response (CFR) measured for a transmit-receive antenna pair and a particular OFDM subcarrier frequency [5]. In particular, a Wi-Fi system that utilizes the MIMO-OFDM scheme can be modeled as follows [12], [16]:

$$B_s(i) = H_s(i)A_s(i) + N, \tag{1}$$

where $s \in [1, \cdots, N_S]$ represents the index of the OFDM subcarrier frequency, N_S is the number of the OFDM subcarrier frequencies, i represents the index of the transmitted and received packets, $A_s(i)$ and $B_s(i)$ are the i^{th} transmitted and received packets associated with the OFDM subcarrier frequency s, respectively, N_T and N_R represent the number of transmitting and receiving antennas, respectively, N represents noise, and $H_s(i)$ is a complex-valued matrix of dimensions $N_T \times N_R$ that comprises the CSI measurements of the MIMO channel for the OFDM subcarrier frequency s. The structure of the CSI matrix $H_s(i)$ is shown below:

$$H_s(i) = \begin{bmatrix} h_{(s,i)}^{(T_1,R_1)} & \cdots & h_{(s,i)}^{(T_1,R_{N_R})} \\ \vdots & h_{(s,i)}^{(T_x,R_y)} & \vdots \\ h_{(s,i)}^{(T_{N_T},R_1)} & \cdots & h_{(s,i)}^{(T_{N_T},R_{N_R})} \end{bmatrix}, \tag{2}$$

where $h_{(s,i)}^{(T_x,R_y)}$ represents the CFR value measured for the OFDM subcarrier frequency s at the i^{th} packet between the x^{th} transmitting antenna, denoted as T_x where $x \in [1, \cdots, N_T]$, and the y^{th} receiving antenna, denoted as R_y where $y \in [1, \cdots, N_R]$. In this work, the employed CSI dataset was acquired using the Linux 802.11n CSI tool [40] which allows the recording of $N_S = 30$ OFMD subcarrier frequencies for each transmit-receive antenna pair. Moreover, the number of transmitting and receiving antennas of the equipment used to collect our dataset are $N_T = 2$ and $N_R = 3$, respectively.

IV. OUR PROPOSED E2EDLF FOR RECOGNIZING HHIs

This section presents our proposed CSI-based E2EDLF for recognizing HHIs. The proposed E2EDLF comprises three phases, namely the input, feature extraction, and recognition phases. In the input phase, the raw CSI data are converted into a set of 2D gray-scale CSI images. Section IV-A provides detailed description of the conversion procedure employed in the input phase. The feature extraction and recognition phases are implemented using a novel CNN architecture that comprises three blocks of layers. Particularly, in the feature extraction phase, the first two blocks of layers within our proposed CNN architecture are utilized to automatically analyze and extract salient features from the CSI images obtained in the input phase. Section IV-B provides detailed description of the feature extraction phase of the proposed E2EDLF. In the recognition phase, the features extracted at the feature extraction phase are fed to the third block of layers within our proposed CNN architecture to recognize the class of the HHI associated with each CSI image. Section IV-C provides detailed description of the recognition phase of the proposed E2EDLF. The structure of our proposed CSI-based E2EDLF for recognizing HHIs is shown in Fig. 1(B).

A. THE INPUT PHASE

The raw CSI data can be viewed as a four-dimensional (4D) tensor that characterizes the variations of the CFR values measured for a Wi-Fi system over the time domain (i.e., packet index), frequency domain (i.e., OFDM subcarrier frequencies), and the spatial domain (i.e., pairs of transmit-receive antennas). Figure 2(A) shows the structure of the recorded raw CSI signals included in the publicly available dataset included in this study [28]. The amplitude and phase information comprised within the raw CSI signals are affected by several factors, including the multi-path effects and the existence of moving objects and humans in the signal propagation path [18]. In this regard, literature reveals that the amplitude information of the CSI signals has been widely used to recognize human activities [16]. This is due to the fact that the changes in the amplitude of the CSI signals are relatively more stable than the deteriorations in the phase information [16], [41]. Therefore, in this study, we employ the amplitude of the CSI values to design an E2EDLF for HHIs recognition.

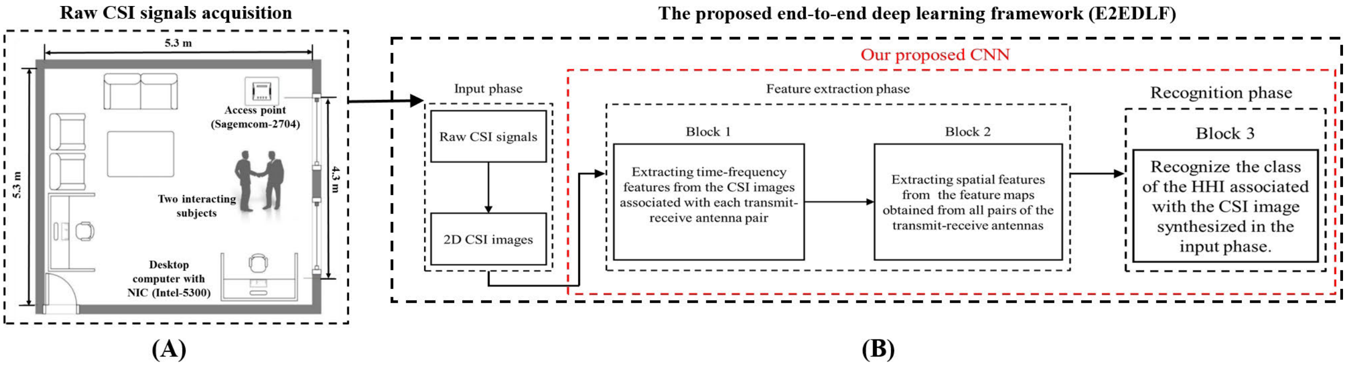
Raw CSI signals acquisition
5.3 m
5.3 m
4.3 m
Access point (Sagemcom-2704)
Two interacting subjects
Desktop computer with NIC (Intel-5300)
The proposed end-to-end deep learning framework (E2EDLF)
Our proposed CNN
Input phase
Raw CSI signals
2D CSI images
Feature extraction phase
Block 1
Extracting time-frequency features from the CSI images associated with each transmit-receive antenna pair
Block 2
Extracting spatial features from the feature maps obtained from all pairs of the transmit-receive antennas
Recognition phase
Block 3
Recognize the class of the HHI associated with the CSI image synthesized in the input phase.
(A)
(B)

FIGURE 1. The structure of the proposed E2EDLF: (A) the layout of the room used to record the raw CSI signals included in the CSI dataset [28], and (B) the three phases comprised within our proposed E2EDLF along with the three blocks of layers that are comprised within the CNN used to implement the feature extraction and recognition phases.

The objective of the input phase is to convert the original 4D raw CSI signals into a set of 2D CSI images that preserve the time, frequency, and spatial information comprised within the original raw CSI signals. To construct the CSI images, we compute the amplitude (i.e., the magnitude) of the raw CSI signals acquired in each recorded trial included in the CSI dataset [28]. Each trial in the dataset comprises the CSI signals recorded for a pair of subjects while performing a particular HHI. Furthermore, the computed amplitude of the CSI signals included in each trial are arranged into a 2D matrix of dimensions $M \times I$, where $M = N_P \times N_S$, $N_P = N_T \times N_R$ represents the number of transmit-receive antenna pairs in the Wi-Fi system, and I represents the number of packets recorded in a particular trial.

A sliding window is utilized to divide the CSI signals comprised within the computed 2D matrix of each trial into a set of overlapped segments. The size of each segment is set to $W = 256$ packets and the overlap between each two consecutive segments is set to $W/2$. Particularly, the CSI signals associated with the OFDM subcarrier frequencies of each transmit-receive antenna pair (T_x, R_y) are divided into overlapped segments. We refer to each segment as $CSI_{(s,\omega)}^{(T_x,R_y)}$, where $s \in [1, N_S]$ represents the index of the OFDM subcarrier frequency and ω represents the index of the packet located at the center of the current position of the utilized sliding window. The size of the CSI segment $CSI_{(s,\omega)}^{(T_x,R_y)}$ is $N_S \times W$. Then, the CSI segments obtained at each window position are normalized to be in the range of $[0, 255]$ and converted into 2D gray-scale sub-images. We denote each of the 2D gray-scale sub-images obtained at a particular position of the utilized sliding window as $\widetilde{CSI}_{(s,\omega)}^{(T_x,R_y)}$. Finally, at each position of the utilized sliding window, we vertically combine the sub-images $\widetilde{CSI}_{(s,\omega)}^{(T_x,R_y)}$ constructed for all $x \in [1, N_T]$ and $y \in [1, N_R]$ to construct a new image, denoted as CSI_ω, with dimensions $M \times W$. Figure 2 illustrates the construction procedure of the image CSI_ω using the raw CSI signals of one trial in the CSI dataset that was recorded for a pair of subjects while performing the handshaking interaction.

B. THE FEATURE EXTRACTION PHASE

The 2D CSI images generated in the input phase characterize the variations in the amplitude of the CSI signals that are comprised within each window position in the time, frequency, and spatial domains. This implies the necessity to analyze the changes in the CSI signals in the time and frequency domains for each transmit-receive antenna pair as well as across different pairs of transmit-receive antennas. Therefore, the objective of the feature extraction phase is to automatically learn latent features from each CSI_ω that can be used to recognize different HHIs.

In this work, the feature extraction phase is implemented using the first two blocks of layers within our proposed CNN architecture as depicted in Fig. 1(B). The first block, denoted by block 1, consists of three layers: convolutional layer ($L_{1,1}$), batch normalization layer ($L_{1,2}$), and rectified linear unit layer ($L_{1,3}$). The objective of the first block of layers is to extract time-frequency features from the CSI images associated with each transmit-receive antenna pair. The second block, denoted by block 2, consists of three layers: convolutional layer ($L_{2,1}$), batch normalization layer ($L_{2,2}$), and rectified linear unit layer ($L_{2,3}$). The objective of the second block of layers is to extract spatial features from all pairs of transmit-receive antennas.

In block 1, $L_{1,1}$ is a 2D convolutional layer with neurons that are connected to subregions of the input image CSI_ω. This layer learns the features localized within these subregions while scanning the input image along the horizontal and vertical dimensions using a set of 2D filters. We refer to the number of 2D filters in $L_{1,1}$ as $N_F^{L_{1,1}}$. The height and width of each of the 2D filters in $L_{1,1}$ are denoted by $D_1^{L_{1,1}}$ and $D_2^{L_{1,1}}$, respectively. The 2D filters in $L_{1,1}$ are shifted along the horizontal and vertical directions to scan the input image CSI_ω. The shift amount performed along the horizontal and vertical directions is determined based on the value of the stride parameter associated with $L_{1,1}$, denoted as $S^{L_{1,1}} = [S_1^{L_{1,1}}, S_2^{L_{1,1}}]$, where $S_1^{L_{1,1}}$ and $S_2^{L_{1,1}}$ represent the value of the shift along the horizontal and vertical directions, respectively.

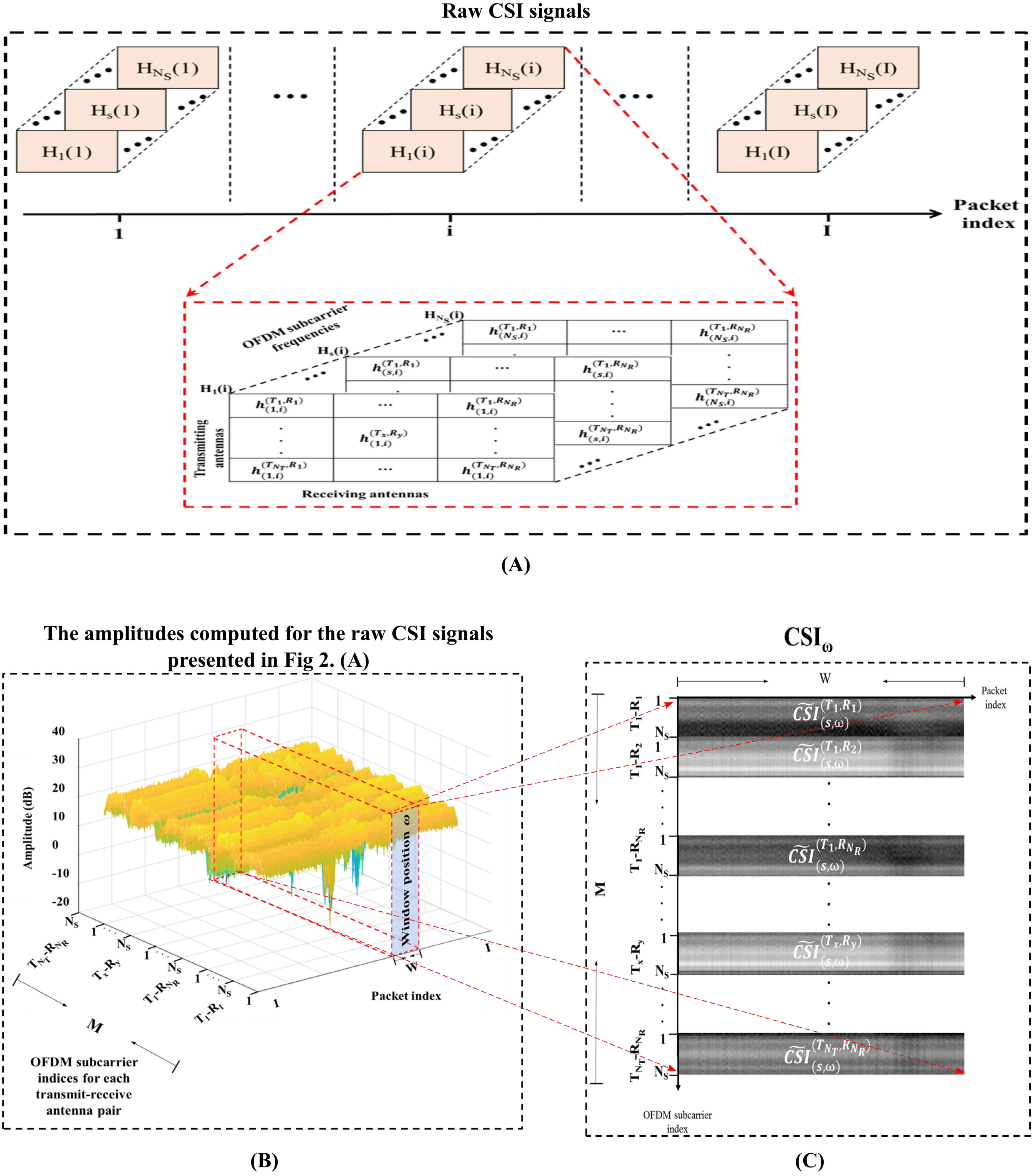
Raw CSI signals
$H_{N_S}(1)$
$H_s(1)$
$H_1(1)$
$H_{N_S}(i)$
$H_s(i)$
$H_1(i)$
$H_{N_S}(I)$
$H_s(I)$
$H_1(I)$
Packet index
1
i
I
OFDM subcarrier frequencies
Transmitting antennas
Receiving antennas
$h^{(T_1,R_1)}_{(1,i)}$
$h^{(T_1,R_{N_R})}_{(1,i)}$
$h^{(T_x,R_y)}_{(1,i)}$
$h^{(T_{N_T},R_1)}_{(1,i)}$
$h^{(T_{N_T},R_{N_R})}_{(1,i)}$
$h^{(T_1,R_1)}_{(s,i)}$
$h^{(T_1,R_{N_R})}_{(s,i)}$
$h^{(T_{N_T},R_{N_R})}_{(s,i)}$
$h^{(T_1,R_1)}_{(N_S,i)}$
$h^{(T_1,R_{N_R})}_{(N_S,i)}$
$h^{(T_{N_T},R_{N_R})}_{(N_S,i)}$
(A)
The amplitudes computed for the raw CSI signals presented in Fig 2. (A)
Amplitude (dB)
40
30
20
10
0
-10
-20
Window position ω
W
Packet index
M
OFDM subcarrier indices for each transmit-receive antenna pair
T_{N_T}-R_{N_R}
T_x-R_y
T_1-R_{N_R}
T_1-R_1
N_S
1
I
(B)
CSI_ω
W
Packet index
M
T_1-R_1
T_1-R_2
T_1-R_{N_R}
T_x-R_y
T_{N_T}-R_{N_R}
$\widetilde{CSI}^{(T_1,R_1)}_{(s,\omega)}$
$\widetilde{CSI}^{(T_1,R_2)}_{(s,\omega)}$
$\widetilde{CSI}^{(T_1,R_{N_R})}_{(s,\omega)}$
$\widetilde{CSI}^{(T_x,R_y)}_{(s,\omega)}$
$\widetilde{CSI}^{(T_{N_T},R_{N_R})}_{(s,\omega)}$
OFDM subcarrier index
(C)

FIGURE 2. Graphical illustration of the procedure used to construct the 2D CSI images. (A) The structure of the recorded raw CSI signals included in the CSI dataset [28], where I represents the number of CSI packets recorded during each trial of a particular HHI. (B) A mesh plot that shows the amplitudes computed for the raw CSI signals presented in Fig.2(A). The red rectangular represents the employed sliding window at position ω. (C) The 2D CSI image constructed for the CSI values comprised within the sliding window at position ω, namely CSI_ω.

The height of the 2D filters and the stride parameter along the vertical direction in $L_{1,1}$ are selected to be $D_1^{L_{1,1}} = N_S$ and $S_2^{L_{1,1}} = N_S$. Furthermore, the width of the 2D filters, the stride parameter along the horizontal direction, and the number of 2D filters in $L_{1,1}$ are selected experimentally to be $D_2^{L_{1,1}} = 32$, $S_1^{L_{1,1}} = 16$, and $N_F^{L_{1,1}} = 160$, respectively.

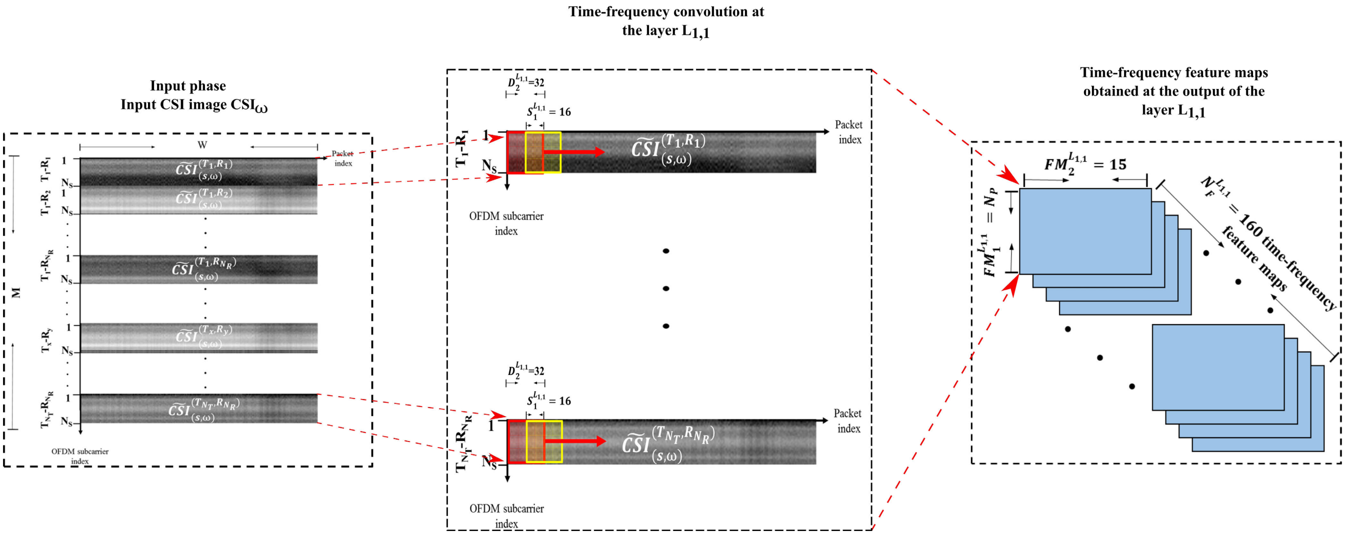
Input phase
Input CSI image CSI$_\omega$
W
Packet index
T_1-R_1
T_1-R_2
T_1-R_{N_R}
T_x-R_y
T_{N_T}-R_{N_R}
M
1
N_S
$\widetilde{CSI}^{(T_1,R_1)}_{(s,\omega)}$
$\widetilde{CSI}^{(T_1,R_2)}_{(s,\omega)}$
$\widetilde{CSI}^{(T_1,R_{N_R})}_{(s,\omega)}$
$\widetilde{CSI}^{(T_x,R_y)}_{(s,\omega)}$
$\widetilde{CSI}^{(T_{N_T},R_{N_R})}_{(s,\omega)}$
OFDM subcarrier index
Time-frequency convolution at the layer $L_{1,1}$
$D_2^{L_{1,1}}$=32
$S_1^{L_{1,1}}$ = 16
Packet index
OFDM subcarrier index
Time-frequency feature maps obtained at the output of the layer $L_{1,1}$
$FM_2^{L_{1,1}}$ = 15
$FM_1^{L_{1,1}}$ = N_P
$N_F^{L_{1,1}}$ = 160 time-frequency feature maps

FIGURE 3. **The time-frequency convolution applied at the layer $L_{1,1}$ to the input CSI image CSI_ω. The number of sub-images comprised within the input CSI image CSI_ω is equal to N_P, where each sub-image corresponds to a transmit-receive antenna pair. The red and yellow rectangles represent the current and next positions of a particular 2D filter, respectively.**

As described earlier, the input CSI image CSI_ω consists of N_P sub-images, where each sub-image is associated with a particular transmit-receive antenna pair and has a size of $N_S \times W$. In light of this, the aforementioned parameter selection scheme enables the analysis of each of the N_P sub-images comprised within the input image CSI_ω along the horizontal axis, which corresponds to the packet index (i.e., time domain), and the vertical axis, which represents the indices of the OFDM subcarrier frequencies associated with a particular transmit-receive antenna pair (i.e., frequency domain). Therefore, $L_{1,1}$ can be viewed as a time-frequency convolutional layer that analyzes the CSI sub-images associated with each transmit-receive antenna pair in the input CSI image CSI_ω and produces a set of time-frequency feature maps (FMs). The number of time-frequency FMs obtained at the output of layer $L_{1,1}$ is equal to the number of 2D filters employed in this layer. In addition, the height and width of each time-frequency FM are $FM_1^{L_{1,1}} = N_P$ and $FM_2^{L_{1,1}} = 15$, respectively. Figure 3 demonstrates the time-frequency convolution applied in the layer $L_{1,1}$ to the input CSI image CSI_ω.

The time-frequency FMs generated at the output of layer $L_{1,1}$ are propagated to the next layer in block 1, which is layer $L_{1,2}$. Layer $L_{1,2}$ normalizes the FMs to simplify the training of the CNN and reduces the potential occurrence of overfitting [42]. The performed normalization at the layer $L_{1,2}$ does not affect the number and size of the FMs obtained at the output of layer $L_{1,1}$. Therefore, the height and width of each FM generated at the output of layer $L_{1,2}$ are $FM_1^{L_{1,2}} = N_P$ and $FM_2^{L_{1,2}} = 15$, respectively. The FMs produced at the output of layer $L_{1,2}$ are propagated to the last layer in block 1, which is layer $L_{1,3}$. Layer $L_{1,3}$ performs a threshold operation to each value in the FMs obtained from layer $L_{1,2}$, where any value less than zero is set to zero [42]. Similar to layer $L_{1,2}$, the performed threshold operation at layer $L_{1,3}$ does not affect the number and size of the FMs obtained at the output of layer $L_{1,2}$. Therefore, the height and width of each FM generated at the output of layer $L_{1,3}$ are $FM_1^{L_{1,3}} = N_P$ and $FM_2^{L_{1,3}} = 15$, respectively.

Figure 4 shows the structure of the FMs produced at the output of layer $L_{1,3}$. Particularly, the number of rows in each FM is equal to N_P. We refer to each row in each FM as $sub - map_p$, where $p \in [1, N_P]$. Each sub-map contains the features extracted from a particular sub-image in the input CSI image CSI_ω. Specifically, the sub-maps within each FM are arranged from top to bottom according to the following order: the top sub-map, denoted as $sub - map_1$, contains the time-frequency features extracted from the CSI signals associated with the transmit-receive antenna pair $T_1 - R_1$, while the bottom sub-map, denoted as $sub - map_{N_P}$, contains the time-frequency features extracted from the CSI signals associated with the transmit-receive antenna pair $T_{N_T} - R_{N_R}$. This implies that the FMs generated at the output of the first block of layers characterize the time-frequency variations of the CSI signals associated with each individual transmit-receive antenna pair without taking into consideration the variations in the CSI signals across different pairs of transmit-receive antennas.

To analyze the variations of the CSI signals across different pairs of transmit-receive antennas, we passed on the FMs generated at the output of layer $L_{1,3}$ to the first layer in block 2, namely $L_{2,1}$. In particular, layer $L_{2,1}$ is a 2D convolutional layer with neurons that are connected to subregions of the FMs generated at the output of layer $L_{1,3}$. This layer learns the features localized within these subregions while scanning the FMs along the horizontal and vertical dimensions using a set of 2D filters. We refer to the number of 2D filters in $L_{2,1}$ as $N_F^{L_{2,1}}$. The height and width of every 2D filter in $L_{2,1}$ are denoted by $D_1^{L_{2,1}}$ and $D_2^{L_{2,1}}$, respectively. The 2D filters in $L_{2,1}$ are shifted along the horizontal and vertical

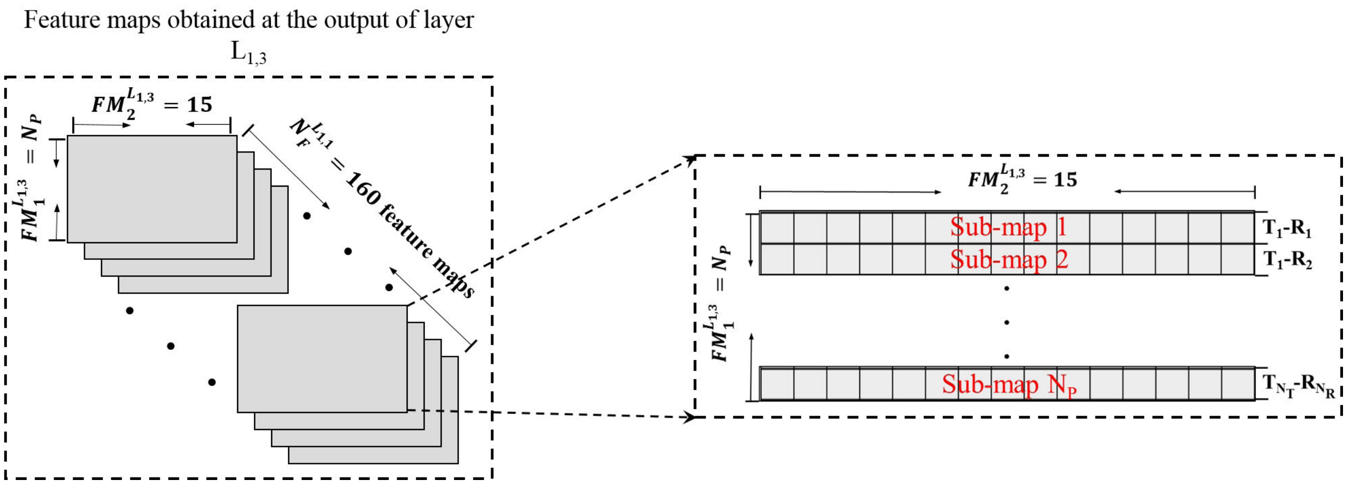

FIGURE 4. Graphical illustration of the structure of the FMs generated at the output of the layer $L_{1,3}$.

directions to scan the FMs. The shift amount performed along the horizontal and vertical directions is determined based on the value of the stride parameter associated with $L_{2,1}$, denoted as $S^{L_{2,1}} = [S_1^{L_{2,1}}, S_2^{L_{2,1}}]$, where $S_1^{L_{2,1}}$ and $S_2^{L_{2,1}}$ represent the values of the shifts along the horizontal and vertical directions, respectively.

The height of the 2D filters and the stride parameter along the vertical direction in $L_{2,1}$ are selected to be $D_1^{L_{2,1}} = N_P$ and $S_2^{L_{2,1}} = 0$. Furthermore, the width of the 2D filters, the stride parameter along the horizontal direction, and the number of 2D filters in $L_{2,1}$ are selected experimentally as follows: $D_2^{L_{2,1}} = 32$, $S_1^{L_{2,1}} = 2$, and $N_F^{L_{2,1}} = 160$. As described earlier, each of the FMs obtained at the output of layer $L_{3,1}$ consists of N_P sub-maps, each of which is associated with a particular transmit-receive antenna pair. Hence, the selected values of the parameters associated with layer $L_{2,1}$ enable the analysis of all sub-maps comprised within each FM, which are associated with different pairs of transmit-receive antennas, over time. Thus, layer $L_{2,1}$ can be viewed as a spatial convolutional layer that analyzes all the sub-maps associated with all the pairs of transmit-receive antennas in the FMs obtained at the output of the layer $L_{1,3}$ to generate a new set of FMs. The number of FMs obtained at the output of layer $L_{2,1}$ is equal to the number of 2D filters employed in this layer. In addition, the height and width of each FM generated at the output of layer $L_{2,1}$ are $FM_1^{L_{2,1}} = 1$ and $FM_2^{L_{2,1}} = 6$, respectively. Figure 5 demonstrates the spatial convolution applied in layer $L_{2,1}$ to each of the time-frequency FMs obtained at the output of layer $L_{1,3}$.

The FMs generated at the output of layer $L_{2,1}$ are passed on to the next layer in block 2, namely layer $L_{2,2}$. Layer $L_{2,2}$ normalizes the FMs and propagates its values to the next layer, namely layer $L_{2,3}$. In layer $L_{2,3}$, a threshold operation is applied to the values of the normalized FMs, which are obtained at the output of layer $L_{2,2}$, by setting the negative values to zero. The performed normalization and threshold operation in layers $L_{2,2}$ and $L_{2,3}$, respectively, do not affect the number and size of the FMs generated at the output of layer $L_{2,1}$. Hence, the size of the FMs generated at the output of layer $L_{2,3}$ is 1×6. The FMs generated at the output of layer $L_{2,3}$ represent the features extracted from the input CSI image CSI_ω. These FMs are propagated to the recognition phase to recognize the class of the HHI associated with the input CSI image CSI_ω.

C. THE RECOGNITION PHASE

In this phase, the FMs learned at the feature extraction phase are further analyzed to recognize the class of the HHI associated with the input CSI image CSI_ω, where the number of HHI classes considered in the current study is thirteen classes, as described in SectionV-A. The recognition phase is implemented using the third block of layers within our proposed CNN architecture, denoted as block 3, which comprises four layers, namely the flatten layer ($L_{3,1}$), fully connected layer ($L_{3,2}$), softmax layer ($L_{3,3}$), and classification layer ($L_{3,4}$). Figure 6 illustrates the structure of the recognition phase.

In layer $L_{3,1}$, the FMs obtained at the output of layer $L_{2,3}$ are rearranged into a column vector of dimensions $D_1^{L_{3,1}} \times 1$, where $D_1^{L_{3,1}} = FM_1^{L_{2,1}} \times FM_2^{L_{2,1}} \times N_F^{L_{2,1}}$. The column vector obtained at the output of layer $L_{3,1}$ is propagated to the next layer of the recognition phase, namely layer $L_{3,2}$. Layer $L_{3,2}$ consists of neurons that are connected to all features in the column vector at the output of layer $L_{3,1}$. The number of neurons in layer $L_{3,2}$ is selected to be the same as the number of HHI classes, which is equal to 13. Moreover, layer $L_{3,2}$ has a set of parameters, namely a weight matrix and a bias vector, that are learned during the training of the CNN. After that, the outputs of layer $L_{3,2}$ are passed on to the next layer of the recognition phase, namely layer $L_{3,3}$. Layer $L_{3,3}$ normalizes the outputs of layer $L_{3,2}$, such that all the values obtained at the output of layer $L_{3,3}$ are greater than zero and their sum is equal to one. Each of the thirteen

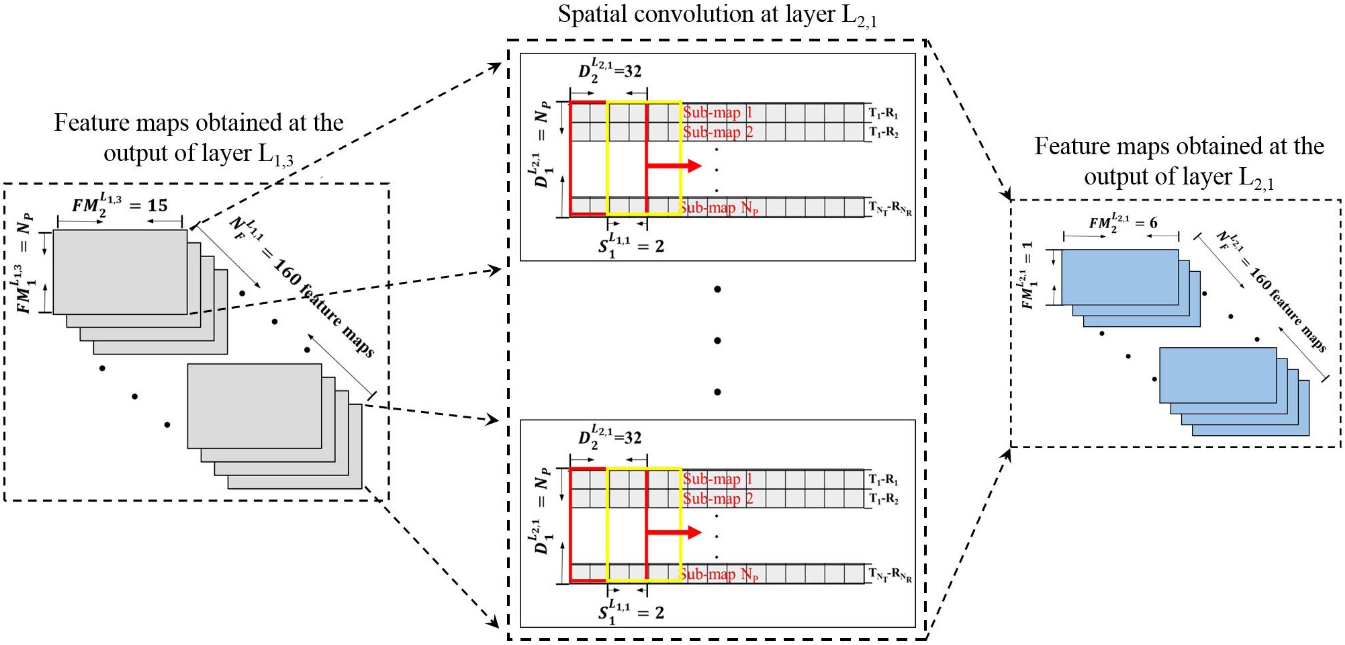
Spatial convolution at layer $L_{2,1}$
Feature maps obtained at the output of layer $L_{1,3}$
$FM_2^{L_{1,3}} = 15$
$FM_1^{L_{1,3}} = N_P$
$N_F^{L_{1,1}} = 160$ feature maps
$D_2^{L_{2,1}}=32$
$D_1^{L_{2,1}} = N_P$
Sub-map 1
Sub-map 2
Sub-map N_P
T_1-R_1
T_1-R_2
T_{N_T}-R_{N_R}
$S_1^{L_{1,1}} = 2$
Feature maps obtained at the output of layer $L_{2,1}$
$FM_2^{L_{2,1}} = 6$
$FM_1^{L_{2,1}} = 1$
$N_F^{L_{2,1}} = 160$ feature maps

FIGURE 5. The spatial convolution applied at layer $L_{2,1}$ to the FMs obtained at the output of layer $L_{1,3}$. The red and yellow rectangles represent the current and next positions of a particular 2D filter, respectively.

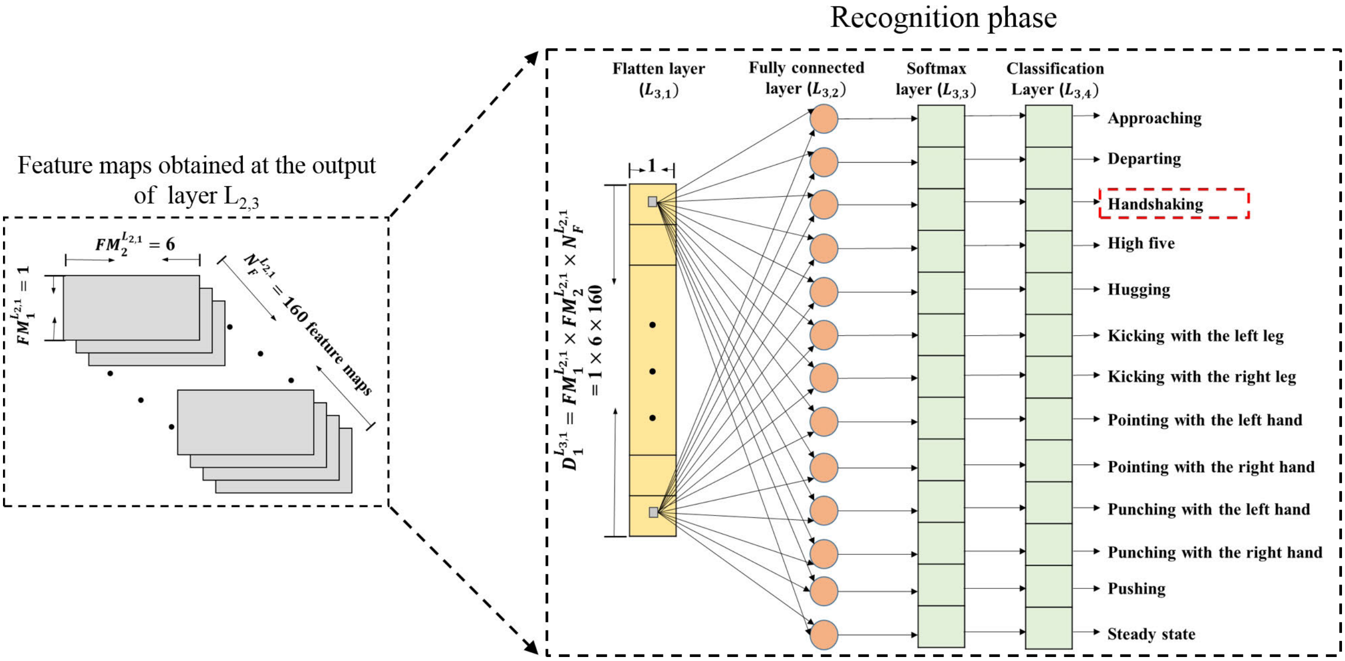
Recognition phase
Feature maps obtained at the output of layer $L_{2,3}$
$FM_2^{L_{2,1}} = 6$
$FM_1^{L_{2,1}} = 1$
$N_F^{L_{2,1}} = 160$ feature maps
Flatten layer ($L_{3,1}$)
Fully connected layer ($L_{3,2}$)
Softmax layer ($L_{3,3}$)
Classification Layer ($L_{3,4}$)
1
$D_1^{L_{3,1}} = FM_1^{L_{2,1}} \times FM_2^{L_{2,1}} \times N_F^{L_{2,1}} = 1 \times 6 \times 160$
Approaching
Departing
Handshaking
High five
Hugging
Kicking with the left leg
Kicking with the right leg
Pointing with the left hand
Pointing with the right hand
Punching with the left hand
Punching with the right hand
Pushing
Steady state

FIGURE 6. Graphical illustration of the structure of the recognition phase.

normalized values obtained at the output of layer $L_{3,3}$ represents the classification probability that the input CSI image CSI_ω belongs to one of the thirteen HHI classes. Finally, the normalized values obtained at the output of layer $L_{3,3}$ are passed on to the last layer in the recognition phase, namely layer $L_{3,4}$. Layer $L_{3,4}$ assigns the input CSI image CSI_ω to the HHI class that has the highest classification probability. Table 1 summarizes the details of the layers comprised within the proposed CNN architecture that is used to implement the feature extraction and recognition phase in our proposed E2EDLF.

V. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we present the publicly available CSI dataset of HHIs that was previously published by our research group [28] and used in this work to assess the performance

TABLE 1. **Summary of the details of the layers comprised within the proposed CNN architecture that is used to implement the feature extraction and recognition phases of our proposed E2EDLF.**

Phase	Block	Layer name	Layer type	Number and size of the feature maps obtained at the output of the layers within the feature extraction phase	Size of the activations at the output of the layers within the recognition phase
Feature extraction phase	Block 1	$L_{1,1}$	Convolution layer	160, 6 ×15	-
		$L_{1,2}$	Batch normalization layer	160, 6 ×15	-
		$L_{1,3}$	Rectified linear unit layer (ReLU)	160, 6 ×15	-
	Block 2	$L_{2,1}$	Convolution layer	160, 1 ×6	-
		$L_{2,2}$	Batch normalization layer	160, 1 ×6	-
		$L_{2,3}$	Rectified linear unit layer (ReLU)	160, 1 ×6	-
Recognition phase	Block 3	$L_{3,1}$	Flatten layer	-	960 × 1
		$L_{3,2}$	Fully connected layer	-	13 × 1
		$L_{3,3}$	Softmax layer	-	13 × 1
		$L_{3,4}$	Classification layer	-	13 × 1

of our proposed E2EDLF. Furthermore, we describe the procedure used to train and test our proposed E2EDLF. After that, we describe and discuss the results achieved by our proposed E2EDLF based on the CSI dataset. Moreover, we present and discuss the runtime of our proposed E2EDLF. Finally, we compare the results achieved by our proposed E2EDLF with the results achieved using different pre-trained CNNs and the results achieved using traditional handcrafted features that are extracted from the CSI signals and classified using a mcSVM classifier.

A. THE CSI DATASET OF HHI

A publicly available CSI dataset of HHIs [28] is used to validate the performance of our proposed E2EDLF. The dataset contains the CSI packets that were recorded for forty distinct pairs of subjects while performing different HHIs inside an office with dimensions 5.3 m × 5.3 m, as illustrated in Fig. 1(A). Each pair of subjects performed ten different trials of the following HHIs: approaching, departing, handshaking, high five, hugging, kicking with the left leg, kicking with the right leg, pointing with the left hand, pointing with the right hand, punching with the left hand, punching with the right hand, and pushing. In addition, each of the recorded trials comprises two interludes, namely the steady state and the interaction interludes. Specifically, throughout the steady state interlude, the pair of subjects were confronting each other without doing any action. During the interaction interval, the pair of subjects perform one of the aforementioned HHIs. Therefore, the total number of HHI classes incorporated in the CSI dataset is equal to thirteen classes, which include the steady state interaction as well as the twelve HHIs described above.

The publicly available Linux 802.11n CSI tool [40] was utilized to record the Wi-Fi signals transmitted from a commercial off-the-shelf access point (AP), namely the Sagemcom 2704, to a desktop computer that is equipped with an Intel 5300 network interface card (NIC). The constructed MIMO system comprises six different pairs of transmit-receive antennas (i.e., $N_P = 6$). Thus, for our MIMO-OFDM system, the number of CSI values contained within each packet is equal to 180 values. Detailed description of the CSI dataset is provided in [28].

B. TRAINING AND TESTING OUR PROPOSED E2EDLF

To train and test the proposed E2EDLF, we have employed a 10-fold cross-validation (CV) procedure. Particularly, for all pairs of subjects, we apply the procedure described in subsection IV-A to transform the CSI signals recorded during each trial in the CSI dataset into a set of labeled CSI images, where the label of each CSI image can be one of the thirteen HHI classes described in subsection V-A.

The labeled CSI images obtained from all pairs of subjects, all trials, and all HHI classes are divided into ten different folds. Particularly, nine folds of the CSI images associated with the thirteen HHI classes are randomly chosen and used to train the feature extraction and recognition phases of our proposed framework, while the remaining fold of the CSI images is used for testing. The 10-fold CV procedure is repeated ten times, and the overall recognition performance is computed for each of the thirteen HHI classes by averaging the results obtained from each repetition [1], [19], [24]. During each repetition of the 10-fold CV procedure, the stochastic gradient decent (SGD) algorithm was employed to learn the weights and biases of the convolutional layers of the feature extraction phase as well as the weights and biases of the fully connected layer in the recognition phase by minimizing the categorical cross-entropy loss function. The training process was run for 50 epochs and the learning rate of the SGD algorithm was experimentally selected to be 0.001.

1) RESULTS OF OUR PROPOSED E2EDLF

The proposed E2EDLF achieved an average recognition accuracy of 86.3% across the thirteen HHI classes. Figure 7 shows the confusion matrix of our proposed framework computed over the ten repetitions of the employed 10-fold CV procedure. The average recognition accuracies computed for each of the thirteen HHI classes, which are shown along the main diagonal of the confusion matrix presented in Fig. 7, are substantially higher than the random classification rate, which is equal to 7.7% (i.e., the reciprocal of the number of HHI classes).

The results presented in Fig. 7 show some confusion between the kicking with the left leg and kicking with the right leg interactions. Similarly, one can observe a confusion

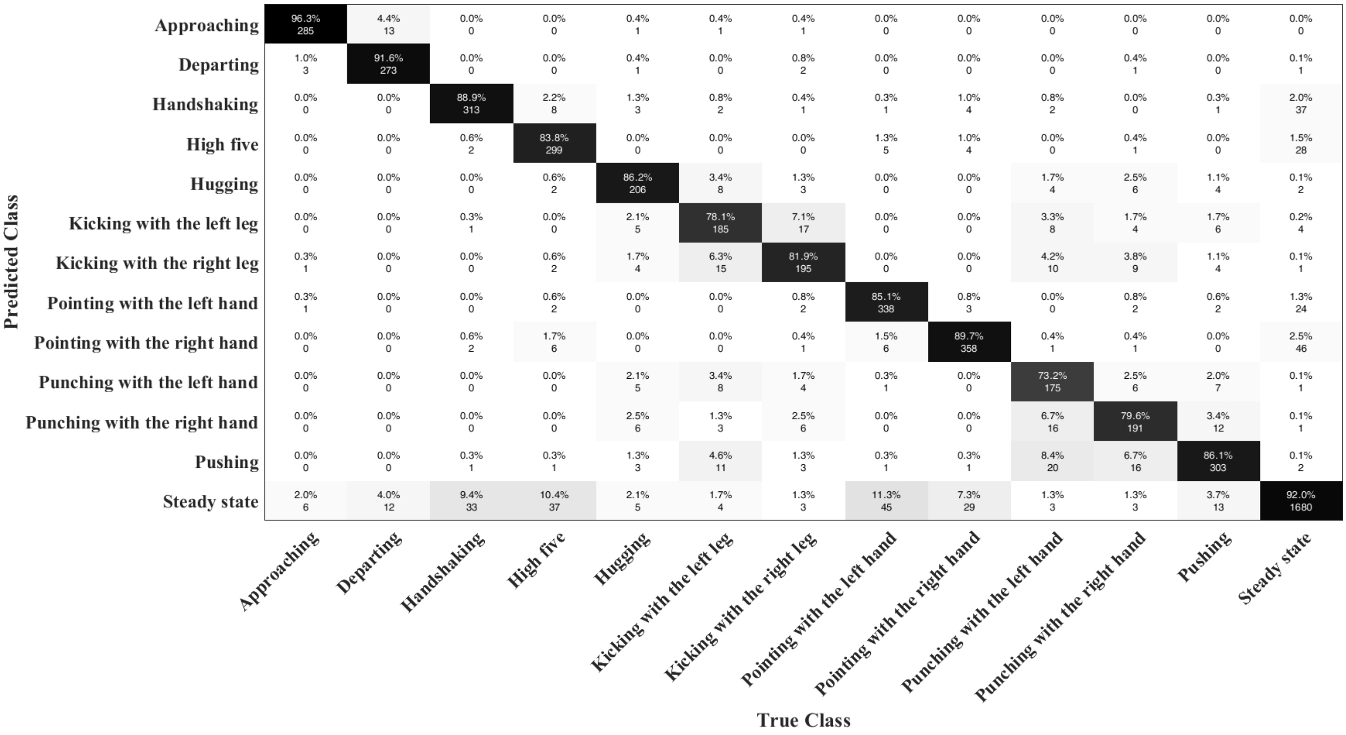

FIGURE 7. The confusion matrix computed for our proposed E2EDLF.

between the punching with the left hand and punching with the right hand interactions. These confusions can be attributed to the large similarity between the two kicking interactions and the two punching interactions. In addition, the confusion matrix indicates the existence of some confusion between the steady state interaction and some HHIs, such as the handshaking, high five, pointing with the left hand, and pointing with the right hand. This can be attributed to the relatively large similarity between the steady state interaction and the beginning and end of each of the previously mentioned HHIs. Particularly, at the beginning and end of the aforementioned HHIs, the pairs of subjects are standing still against each other, which is similar to the behavior performed by the subjects during the steady state interaction.

We also compute the $F_1 - Score$ for each of the thirteen HHI classes. The $F_1 - Score$ is a harmonic mean of the recall and precision that attains its best value at 1 and the worst value at 0. The $F_1 - Score$ can be used to evaluate the recognition performance when the numbers of samples associated with different classes are imbalanced [43]–[45]. In this regard, Fig. 8 shows the number of CSI images extracted from the recorded trials of each of the thirteen interactions across all pairs of subjects in our dataset. Figure 8 illustrates that the number of constructed CSI images varies substantially across the thirteen interactions. This is due to the variation in the lengths of the trials recorded for the different interactions in our dataset. To evaluate the impact of the CSI dataset imbalance on the recognition performance of our proposed E2EDLF, we have computed the $F_1 - Score$ for each of the thirteen HHI classes.

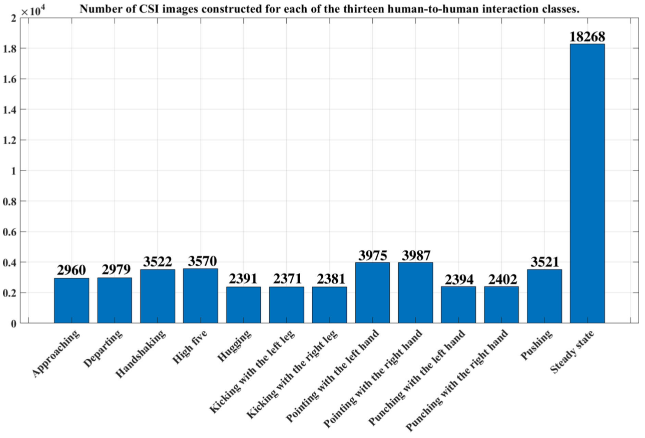

FIGURE 8. The number of CSI images extracted from the recorded trials associated with each of the thirteen HHIs across all pairs of subjects in our dataset.

The blue bars presented in Fig. 9 show the mean $F_1 - Score$ values computed for each of the thirteen HHI classes across the ten repetitions of the 10-fold CV procedure. The $F_1 - Score$ values obtained using our proposed framework for each of the thirteen HHI classes are higher than 0.8. In fact, the average $F_1 - Score$ value computed across all interactions is equal to 0.86.

To further analyze the recognition performance obtained using our proposed framework, we have computed the Cohen's kappa score [46], [47] for each of the thirteen HHI classes. The Cohen's kappa score is used to measure the agreement between the classes of the CSI images predicted

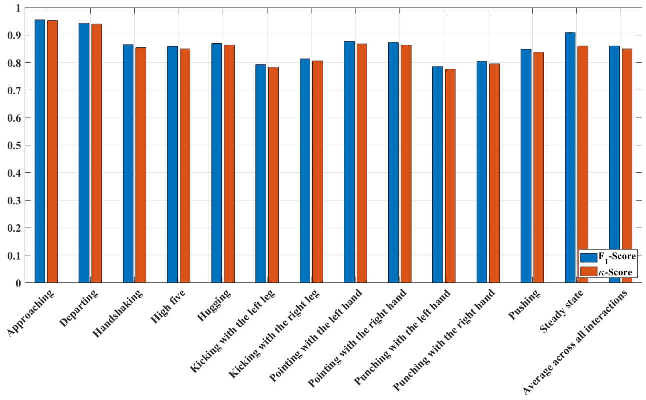

FIGURE 9. The $F_1 - Score$ and $\kappa - Score$ values obtained using our proposed E2EDLF for each of the thirteen HHI classes.

by the proposed E2EDLF and the matching true classes of these images after removing the agreements occurring by chance. Particularly, the Cohen's kappa score enables us to compare the recognition performance obtained by our proposed E2EDLF with the recognition performance obtained by random guessing according to the number of samples of each class. According to Landis *et al.* [48], the value of the Cohen's kappa score ($\kappa - Score$) can be interpreted as follows to determine the strength of agreement: ($\kappa - Score \leq 0$) poor agreement, ($0 < \kappa - Score \leq 0.2$) slight agreement, ($0.2 < \kappa - Score \leq 0.4$) fair agreement, ($0.4 < \kappa - Score \leq 0.6$) moderate agreement, ($0.6 < \kappa - Score \leq 0.8$) substantial agreement, and ($0.8 < \kappa - Score \leq 1$) almost perfect agreement.

The red bars in Fig. 9 show the $\kappa - Score$ values computed for each of the thirteen HHI classes over the ten repetitions of the 10-fold CV procedure. The $\kappa - Score$ values presented in Fig. 9 indicate that the strength of agreement of the recognition accuracies computed for the kicking with the left leg, kicking with the right leg, punching with the left hand, and punching with the right hand interactions are within the substantial agreement range. Furthermore, the $\kappa - Score$ values that are computed for the remaining interactions are within the perfect agreement range. In fact, the average $\kappa - Score$ value computed across all interactions using our proposed framework is equal to 0.85. Therefore, the strength of agreement of the recognition accuracy computed across all interactions is within the perfect agreement range. The results presented in Figs. 7 and 9 illustrate the ability of our proposed E2EDLF to accurately recognize HHIs.

2) RUNTIME OF OUR PROPOSED E2EDLF

The proposed E2EDLF was executed on a workstation with an Intel Xeon Silver-4110 2.1GHz 16 cores CPU, 64 GB RAM, and Nvidia Quadro P6000 GPU. The runtime of our proposed framework is quantified in terms of the following three different metrics: (1) The average ± standard deviation value of the time required to train the proposed framework computed over the ten repetitions of the employed 10-fold CV procedure, and we refer to this metric as the training time. (2) The average ± standard deviation value of the time required to construct the input CSI image associated with a particular window position at the input phase computed across the thirteen HHI classes, and we refer to this metric as the CSI image construction time. (3) The average ± standard deviation value of the time required to recognize the class of an input CSI image at the recognition phase computed across the ten repetitions of the 10-fold CV procedure, and we refer to this metric as the CSI image recognition time.

The average± standard deviation values of the training time, CSI image construction time, and CSI image recognition time computed for our proposed framework were 934.27 ± 3.56 s, 0.00051 ± 0.000042 s, 0.00022 ± 0.000018 s, respectively. Despite the relatively large training time required to train our proposed framework, the training process is performed offline and the trained framework is used to recognize the testing CSI images online. The average time required to construct and recognize an input CSI image using our trained E2EDLF is equal to 0.00073 s. We refer to the average time required to construct and recognize an input CSI image as the framework response time. The proportion between the response time of our proposed E2EDLF and the length of each window position, where the later time is computed by dividing the number of packets in each window position (which is equal to 256 packets) by the number of packets received per each second (which is equal to 320 packets/s), is approximately 0.091%.

The previously described runtime analysis shows the ability of our proposed E2EDLF to recognize the class of the input CSI image associated with a particular window position before moving to the next window position. This indicates the suitability of using our proposed E2EDLF for real-time CSI-based HHI recognition.

3) COMPARISON WITH THE RESULTS OBTAINED USING OTHER STATE-OF-THE-ART PRE-TRAINED CNNs

In this section, we compare the results obtained by our proposed E2EDLF with the results obtained using three state-of-the-art pre-trained CNNs, namely the GoogleNet [49], ResNet-18 [50], and SqueezeNet [51]. Particularly, in this study, we have used the implementation provided in the MATLAB DL toolbox [52] for each of the three pre-trained CNNs. Furthermore, we have utilized the concept of transfer-learning [53] to tune the three pre-trained CNNs using the CSI images extracted from the CSI dataset. A zero-padding procedure is applied to adjust the size of each input CSI image, which is equal to 180 × 256 × 1 in our proposed E2EDLF, to match the sizes of the input layers of the GoogleNet, ResNet-18, and SqueezeNet, which are equal to 224×224×3, 224 × 224 × 3, and 227 × 227 × 3, respectively. Furthermore, the number of neurons in the last fully connected layer in each one of these three pre-trained CNNs was set to 13 (i.e., the number of HHI classes in the CSI dataset). Moreover, the initial learning rate was set to 0.001, the number of epochs was set to 15, and the SGD with momentum

TABLE 2. The results obtained for each of the employed three pre-trained CNNs.

Interaction	Pre-trained CNNs								
	GoogleNet			ResNet-18			SqueezeNet		
	Recognition accuracy (%)	$F_1 - Score$	$\kappa - Score$	Recognition accuracy (%)	$F_1 - Score$	$\kappa - Score$	Recognition accuracy (%)	$F_1 - Score$	$\kappa - Score$
Approaching	92.7	0.95	0.93	91.6	0.94	0.92	94.5	0.92	0.93
Departing	93.4	0.95	0.94	90.3	0.95	0.92	92.9	0.92	0.92
Handshaking	79.3	0.85	0.80	84.6	0.86	0.84	82.7	0.85	0.83
High five	75.5	0.78	0.74	79.3	0.80	0.78	75.8	0.83	0.77
Hugging	63.5	0.75	0.66	77.4	0.81	0.77	69.5	0.79	0.72
Kicking with the left leg	53.5	0.54	0.47	67.5	0.71	0.66	66.3	0.66	0.64
Kicking with the right leg	49.8	0.49	0.44	60.3	0.74	0.63	62.1	0.66	0.62
Pointing with the left hand	78.0	0.74	0.74	81.9	0.82	0.80	78.4	0.81	0.78
Pointing with the right hand	77.3	0.77	0.75	79.9	0.81	0.78	78.9	0.82	0.79
Punching with the left hand	59.1	0.60	0.57	60.4	0.67	0.60	60.4	0.71	0.61
Punching with the right hand	58.6	0.70	0.61	64.9	0.70	0.65	71.6	0.71	0.70
Pushing	68.0	0.79	0.71	75.6	0.73	0.72	74.2	0.75	0.72
Steady state	88.4	0.85	0.81	88.2	0.84	0.80	90.3	0.86	0.82
Average across all interactions	72.1	0.75	0.70	77.1	0.80	0.76	76.7	0.79	0.76

algorithm was used to tune each of the three pre-trained CNNs. To facilitate the comparison with our proposed framework, we have computed the recognition performance for each of the three pre-trained CNNs using the same training and testing sets of CSI images that were employed to evaluate our proposed framework, where these training and testing sets of CSI images were obtained using the 10-fold CV procedure described in subsection V-B.

Table 2 shows the recognition accuracy, $F_1 - Score$, and $\kappa - Score$ values obtained for each of the thirteen HHI classes using each each one of the three pre-trained CNNs. In particular, the average recognition accuracies computed across all HHI classes for the GoogleNet, ResNet-18, and SqueezeNet are 72.1%, 77.1%, and 76.7%, respectively. The average $F_1 - Score$ values computed across all interactions for the GoogleNet, ResNet-18, and SqueezeNet are 0.75, 0.80, and 0.79, respectively. Moreover, the average $\kappa - Score$ values computed across all interactions for the GoogleNet, ResNet-18, and SqueezeNet are 0.70, 0.76, and 0.76, respectively. This implies that the strength of agreement obtained for the average recognition accuracies computed across all interactions for each of the three pre-trained CNNs are within the substantial agreement range. The results presented in Figs. 7 and 9 indicate that the recognition performance obtained by our proposed framework outperforms the recognition results obtained using each of the three pre-trained CNNs, which are depicted in Table 2.

We also compute the runtime for each of the three pre-trained CNNs in terms of the training time and CSI image recognition time, as described in subsection V-B2. Table 3 shows the runtime computed for each of the three pre-trained CNNs. The runtime computed for our proposed framework, which is presented in subsection V-B2, indicates that our proposed framework required less training time and CSI image recognition time compared with the training time and CSI image recognition time required by each of the three pre-trained CNNs. This can be attributed to fact that our proposed E2EDLF comprises a relatively smaller number of layers compared with the number of layers contained within each of the three pre-trained CNNs. This implies that the number of free parameters in our proposed framework is considerably less than the free parameters in each of the three pre-trained CNNs. As a consequence, the runtime analysis reported in the current study suggests the feasibility of utilizing our proposed framework for developing real-time systems that can accurately recognize HHIs.

TABLE 3. The runtime computed for each of the employed three pre-trained CNNs.

Pre-trained CNN	Training time (s)	CSI image recognition time (s)
GoogleNet	7314.1 ± 175.1	0.00125 ± 0.000015
ResNet-18	5343.9 ± 537.2	0.0016 ± 0.00001
SqueezeNet	10984.4 ± 102.9	0.00097 ± 0.000027

4) COMPARISON WITH THE RESULTS ACHIEVED USING HANDCRAFTED FEATURES AND CONVENTIONAL CLASSIFIERS

This section presents a comparison between the results achieved by our proposed E2EDLF and the results achieved using traditional handcrafted features that are extracted from the CSI signals. Specifically, the sliding window approach, which was described in subsection IV-A, is used to divide the CSI signals into a set of overlapped segments, where each segment contains 256 packets and the overlap between any two consecutive window positions is equal to 128 packets. At each window position, we extract a set of commonly used handcrafted features that are computed from the time- and

TABLE 4. The results obtained using the handcrafted features and the four conventional classifiers described in subsection V-B4.

Interaction	Conventional classifiers											
	mcSVM			k-NN			Naïve Bayes			Decision tree		
	Recognition accuracy (%)	$F_1 - Score$	$\kappa - Score$	Recognition accuracy (%)	$F_1 - Score$	$\kappa - Score$	Recognition accuracy (%)	$F_1 - Score$	$\kappa - Score$	Recognition accuracy (%)	$F_1 - Score$	$\kappa - Score$
Approaching	70.37	0.70	0.68	61.10	0.53	0.54	35.84	0.37	0.33	45.16	0.42	0.40
Departing	64.40	0.66	0.64	51.29	0.53	0.50	44.90	0.34	0.35	39.45	0.40	0.36
Handshaking	68.95	0.74	0.70	55.36	0.49	0.49	9.58	0.13	0.07	28.51	0.27	0.23
High five	61.72	0.73	0.65	48.77	0.50	0.46	18.23	0.15	0.11	27.20	0.27	0.23
Hugging	55.90	0.66	0.59	39.98	0.48	0.41	13.49	0.21	0.13	27.44	0.28	0.24
Kicking with the left leg	45.65	0.46	0.43	29.33	0.37	0.30	11.49	0.21	0.12	20.78	0.21	0.17
Kicking with the right leg	47.96	0.52	0.47	28.26	0.43	0.32	13.33	0.20	0.13	21.31	0.20	0.17
Pointing with the left hand	54.23	0.70	0.59	38.98	0.51	0.41	19.58	0.14	0.09	22.90	0.22	0.17
Pointing with the right hand	59.04	0.68	0.61	35.33	0.51	0.38	4.81	0.20	0.05	22.56	0.22	0.17
Punching with the left hand	39.95	0.48	0.41	28.67	0.33	0.28	14.82	0.21	0.14	19.42	0.21	0.17
Punching with the right hand	41.63	0.46	0.41	26.64	0.37	0.28	22.96	0.22	0.19	21.78	0.23	0.19
Pushing	55.81	0.55	0.53	36.64	0.48	0.38	20.36	0.33	0.22	26.46	0.30	0.24
Steady state	91.86	0.80	0.76	91.05	0.75	0.69	87.43	0.72	0.63	74.42	0.75	0.59
Average across all interactions	58.27	0.63	0.57	43.95	0.48	0.42	24.37	0.26	0.20	30.57	0.31	0.26

TABLE 5. Summary of the average recognition accuracies computed across all interactions using each of the four conventional classifiers described in subsection V-B4, each of the three pre-trained CNNs described in subsection V-B3, and our proposed E2EDLF.

Type of the employed machine learning method	Classifier	Average recognition accuracy across all interactions (%)
Conventional machine learning methods	mcSVM	58.27
	k-NN	43.95
	Naïve Bayes	24.37
	Decision tree	30.57
Deep learning methods	Pre-trained GoogleNet	72.1
	Pre-trained ResNet-18	77.1
	Pre-trained SqueezeNet	76.7
	Our proposed E2EDLF	**86.3**

frequency-domains of the CSI signals [7], [54], including the mean, minimum value, standard deviation, maximum value, skewness, kurtosis, entropy, fast Fourier transform (FFT) peak, energy, and domain frequency ratio. The extracted features at each window position are combined to form a feature vector. The constructed feature vectors are used to train and test four conventional classifiers, including a mcSVM classifier with the radial basis function kernel [55], k-NN classifier [56] with $k = 5$, naive Bayes classifier [56], and decision tree classifier [56], to recognize the HHI class associated with each feature vector. To evaluate the performance of each one of the four conventional classifiers, we have utilized the 10-fold CV procedure described in subsection V-B. Table 4 shows the recognition accuracy, $F_1 - Score$, and $\kappa - Score$ values obtained using each one of the four conventional classifiers for each of the thirteen interactions. Specifically, the average recognition accuracy computed across all interactions using the mcSVM, k-NN, naive Bayes, and decision tree classifiers are 58.3%, 43.9%, 24.3%, and 30.5%, respectively. Moreover, the average $F_1 - Score$ / $\kappa - Score$ values computed across all interactions using the mcSVM, k-NN, naive Bayes, and decision tree classifiers are 0.63/0.57, 0.48/0.42, 0.26/0.2, and 0.31/0.26, respectively.

Table 5 shows the average recognition accuracies computed across all interactions using each of the four conventional classifiers in this subsection, each of the three pre-trained CNNs described in subsection V-B3, and our proposed E2EDLF. The results presented in Table 5 indicate that the performance of our proposed E2EDLF outperforms significantly the performances obtained using the handcrafted features combined with each one of the four conventional classifiers as well as the performances achieved using each of the three pre-trained CNNs.

VI. CONCLUSION

In this paper, we explored the feasibility of recognizing HHIs based on the CSI signals. Particularly, we presented a new E2EDLF that analyzes the time, frequency, and spatial domains of the CSI signals to recognize the class of the performed HHI. A publicly available CSI dataset of HHI was utilized to validate the performance of our proposed E2EDLF. Moreover, we have compared the results of our proposed E2EDLF with the results achieved using three well-known pre-trained CNNs and the performance obtained using commonly used handcrafted features that were classified using four different conventional classifiers. The experimental results depicted in this study illustrate the ability of our proposed E2EDLF to accurately recognize HHIs based on CSI signals analysis. Furthermore, the recognition accuracies achieved by our proposed E2EDLF are considerably higher than the accuracies achieved using the pre-trained CNNs and the achieved using the traditional handcrafted features.

In the future, we aim to extend our proposed E2EDLF to recognize group activities that involve more than two interacting persons. Furthermore, we intend to explore the potential of applying our proposed E2EDLF to recognize HHIs that are performed in a non-line-of-sight configuration. In addition, we plan to investigate the use of our proposed E2EDLF to recognize different types of fine-grained single-human activities, such as hand gestures and sign language. Moreover, we plan to investigate the possibility of developing CNN architectures that can directly analyze the raw CSI signals without converting it into another representation.

REFERENCES

[1] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, ''Understanding and modeling of WiFi signal based human activity recognition,'' in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, 2015, pp. 65–76.

[2] J. K. Aggarwal and M. S. Ryoo, ''Human activity analysis: A review,'' *ACM Comput. Surv.*, vol. 43, no. 3, pp. 1–43, 2011.

[3] O. D. Lara and M. A. Labrador, ''A survey on human activity recognition using wearable sensors,'' *IEEE Commun. Surveys Tuts.*, vol. 15, no. 3, pp. 1192–1209, 3rd Quart., 2013.

[4] X. Li, Y. He, and X. Jing, ''A survey of deep learning-based human activity recognition in radar,'' *Remote Sens.*, vol. 11, no. 9, p. 1068, May 2019.

[5] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, ''Device-free human activity recognition using commercial WiFi devices,'' *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1118–1131, May 2017.

[6] R. Alazrai, M. Momani, and M. Daoud, ''Fall detection for elderly from partially observed depth-map video sequences based on view-invariant human activity representation,'' *Appl. Sci.*, vol. 7, no. 4, p. 316, Mar. 2017.

[7] C. Feng, S. Arshad, S. Zhou, D. Cao, and Y. Liu, ''Wi-multi: A three-phase system for multiple human activity recognition with commercial WiFi devices,'' *IEEE Internet Things J.*, vol. 6, no. 4, pp. 7293–7304, Aug. 2019.

[8] R. Alazrai, A. Zmily, and Y. Mowafi, ''Fall detection for elderly using anatomical-plane-based representation,'' in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2014, pp. 5916–5919.

[9] R. Alazrai, Y. Mowafi, and E. Hamad, ''A fall prediction methodology for elderly based on a depth camera,'' in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 4990–4993.

[10] 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.*, 2013, pp. 27–38.

[11] J. Lien, ''Soli: Ubiquitous gesture sensing with millimeter wave radar,'' *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–19, 2016.

[12] Y. Ma, G. Zhou, and S. Wang, ''WiFi sensing with channel state information: A survey,'' *ACM Comput. Surv.*, vol. 52, no. 3, pp. 1–36, Jul. 2019.

[13] W. Jiang, D. Koutsonikolas, W. Xu, L. Su, C. Miao, F. Ma, S. Yao, Y. Wang, Y. Yuan, H. Xue, C. Song, and X. Ma, ''Towards environment independent device free human activity recognition,'' in *Proc. 24th Annu. Int. Conf. Mobile Comput. Netw.*, 2018, pp. 289–304.

[14] A. Virmani and M. Shahzad, ''Position and orientation agnostic gesture recognition using WiFi,'' in *Proc. 15th Annu. Int. Conf. Mobile Syst., Appl., Services*, Jun. 2017, pp. 252–264.

[15] R. H. Venkatnarayan, G. Page, and M. Shahzad, ''Multi-user gesture recognition using WiFi,'' in *Proc. 16th Annu. Int. Conf. Mobile Syst., Appl., Services*, Jun. 2018, pp. 401–413.

[16] 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.

[17] Z. Chen, L. Zhang, C. Jiang, Z. Cao, and W. Cui, ''WiFi CSI based passive human activity recognition using attention based BLSTM,'' *IEEE Trans. Mobile Comput.*, vol. 18, no. 11, pp. 2714–2724, Nov. 2019.

[18] H. F. Thariq Ahmed, H. Ahmad, and A. C. V., ''Device free human gesture recognition using Wi-Fi CSI: A survey,'' *Eng. Appl. Artif. Intell.*, vol. 87, Jan. 2020, Art. no. 103281.

[19] B. Sheng, F. Xiao, L. Sha, and L. Sun, ''Deep Spatial–Temporal model based cross-scene action recognition using commodity WiFi,'' *IEEE Internet Things J.*, vol. 7, no. 4, pp. 3592–3601, Apr. 2020.

[20] R. Alazrai, M. Abuhijleh, H. Alwanni, and M. I. Daoud, ''A deep learning framework for decoding motor imagery tasks of the same hand using eeg signals,'' *IEEE Access*, vol. 7, pp. 109612–109627, 2019.

[21] Y. LeCun, Y. Bengio, and G. Hinton, ''Deep learning,'' *Nature*, vol. 521, no. 7553, p. 436, 2015.

[22] J. Schmidhuber, ''Deep learning in neural networks: An overview,'' *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

[23] F. Wang, W. Gong, J. Liu, and K. Wu, ''Channel selective activity recognition with WiFi: A deep learning approach exploring wideband information,'' *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 1, pp. 181–192, Jan. 2020.

[24] D. A. Khan, S. Razak, B. Raj, and R. Singh, ''Human behaviour recognition using WiFi channel state information,'' in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 7625–7629.

[25] Q. Gao, J. Wang, X. Ma, X. Feng, and H. Wang, ''Csi-based device-free wireless localization and activity recognition using radio image features,'' *IEEE Trans. Veh. Technol.*, vol. 66, no. 11, p. 10346–10356, Dec. 2017.

[26] T. Gu, L. Wang, H. Chen, X. Tao, and J. Lu, ''Recognizing multiuser activities using wireless body sensor networks,'' *IEEE Trans. Mobile Comput.*, vol. 10, no. 11, pp. 1618–1631, Nov. 2011.

[27] R. Alazrai, Y. Mowafi, and C. S. George Lee, ''Anatomical-plane-based representation for human–human interactions analysis,'' *Pattern Recognit.*, vol. 48, no. 8, pp. 2346–2363, Aug. 2015.

[28] 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.

[29] K. Niu, F. Zhang, J. Xiong, X. Li, E. Yi, and D. Zhang, ''Boosting fine-grained activity sensing by embracing wireless multipath effects,'' in *Proc. 14th Int. Conf. Emerg. Netw. Exp. Technol.*, Dec. 2018, pp. 139–151.

[30] K. Ali, A. X. Liu, W. Wang, and M. Shahzad, ''Recognizing keystrokes using WiFi devices,'' *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1175–1190, May 2017.

[31] M. Li, Y. Meng, J. Liu, H. Zhu, X. Liang, Y. Liu, and N. Ruan, ''When CSI meets public WiFi: Inferring your mobile phone password via WiFi signals,'' in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2016, pp. 1068–1079.

[32] 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.

[33] J. Liu, Y. Wang, Y. Chen, J. Yang, X. Chen, and J. Cheng, ''Tracking vital signs during sleep leveraging off-the-shelf WiFi,'' in *Proc. 16th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2015, pp. 267–276.

[34] M. Zhao, F. Adib, and D. Katabi, ''Emotion recognition using wireless signals,'' in *Proc. 22nd Annu. Int. Conf. Mobile Comput. Netw.*, Oct. 2016, pp. 95–108.

[35] H. Abdelnasser, K. Harras, and M. Youssef, ''A ubiquitous WiFi-based fine-grained gesture recognition system,'' *IEEE Trans. Mobile Comput.*, vol. 18, no. 11, pp. 2474–2487, Nov. 2019.

[36] H. Li, W. Yang, J. Wang, Y. Xu, and L. Huang, ''WiFinger: Talk to your smart devices with finger-grained gesture,'' in *Proc. ACM Int. Joint Conf. Pervas. Ubiquitous Comput.*, Sep. 2016, pp. 250–261.

[37] S. Palipana, D. Rojas, P. Agrawal, and D. Pesch, ''FallDeFi: Ubiquitous fall detection using commodity Wi-Fi devices,'' *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 4, pp. 1–25, Dec. 2018.

[38] Y. Wang, J. Liu, Y. Chen, M. Gruteser, J. Yang, and H. Liu, ''E-eyes: Device-free location-oriented activity identification using fine-grained WiFi signatures,'' in *Proc. 20th Annu. Int. Conf. Mobile Comput. Netw.*, 2014, pp. 617–628.

[39] F. Xiao, J. Chen, X. Xie, L. Gui, L. Sun, and R. Wang, ''SEARE: A system for exercise activity recognition and quality evaluation based on green sensing,'' *IEEE Trans. Emerg. Topics Comput.*, vol. 8, no. 3, pp. 752–761, Jul. 2020.

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

[41] 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.

[42] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, ''Deep learning with convolutional neural networks for EEG decoding and visualization,'' 2017, *arXiv:1703.05051*. [Online]. Available: http://arxiv.org/abs/1703.05051

[43] R. Alazrai, M. Momani, H. A. Khudair, and M. I. Daoud, ''EEG-based tonic cold pain recognition system using wavelet transform,'' *Neural Comput. Appl.*, vol. 31, no. 7, pp. 3187–3200, Oct. 2017.

[44] R. Alazrai, R. Homoud, H. Alwanni, and M. Daoud, ''EEG-based emotion recognition using quadratic time-frequency distribution,'' *Sensors*, vol. 18, no. 8, p. 2739, Aug. 2018.

[45] R. Alazrai, A. Al-Saqqaf, F. Al-Hawari, H. Alwanni, and M. I. Daoud, ''A time-frequency distribution-based approach for decoding visually imagined objects using eeg signals,'' *IEEE Access*, vol. 8, p. 138,955–138972, 2020.

[46] J. Sim and C. C. Wright, ''The kappa statistic in reliability studies: Use, interpretation, and sample size requirements,'' *Phys. Therapy*, vol. 85, no. 3, pp. 257–268, Mar. 2005.

[47] J. Cohen, ''A coefficient of agreement for nominal scales,'' *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960.

[48] J. R. Landis and G. G. Koch, ''The measurement of observer agreement for categorical data,'' *Biometrics*, vol. 15, pp. 159–174, Dec. 1977.

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

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

[51] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and 0.5MB model size," 2016, *arXiv:1602.07360*. [Online]. Available: http://arxiv.org/abs/1602.07360

[52] I. The MathWorks. (2020). *Deep Learn. Toolbox*. [Online]. Available: https://www.mathworks.com/help/deeplearning/

[53] A. S. Qureshi, A. Khan, A. Zameer, and A. Usman, "Wind power prediction using deep neural network based meta regression and transfer learning," *Appl. Soft Comput.*, vol. 58, pp. 742–755, Sep. 2017.

[54] Y. Zeng, P. H. Pathak, and P. Mohapatra, "WiWho: WiFi-based person identification in smart spaces," in *Proc. 15th ACM/IEEE Int. Conf. Inf. Process. Sensor Netw. (IPSN)*, Apr. 2016, pp. 1–12.

[55] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, Apr. 2011.

[56] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, NY, USA: Wiley, 2001.

RAMI ALAZRAI (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from Purdue University, West Lafayette, IN, USA, in 2013. In June 2013, he joined the School of Electrical Engineering and Information Technology, German Jordanian University (GJU), where he is currently an Associate Professor. His research interests include machine learning, image and signal processing, semantic video analysis, human activity recognition, emotion recognition, human-to-human interaction representation and analysis, and elderly fall detection.

MOHAMMAD HABABEH received the B.S. degree in computer engineering from German Jordanian University (GJU), Amman, Jordan, in 2019. He is currently pursuing the master's degree in embedded systems with the University of Freiburg. During his bachelor's program, he spent an exchange semester at Siegen University, Germany, and an exchange year at Hochschule Bonn-Rhein-Sieg, Germany. He is currently a Research Assistant with the University of Freiburg. His research interests include signal processing, image and video processing, pattern recognition, computer vision, human activity recognition, machine learning, and deep learning.

BAHA' A. ALSAIFY received the B.S. degree in computer engineering from the Jordan University of Science and Technology, Irbid, Jordan, in 2007, and the M.S. and Ph.D. degrees in computer engineering from the University of Arkansas, Fayetteville, AR, USA, in 2009 and 2012, respectively. Then, he joined the Computer Engineering Department, Yarmouk University, Irbid. In 2015, he joined the Network Engineering and Security Department, Jordan University of Science and Technology, where he is currently an Assistant Professor. His research interests include radio frequency identification (RFID), computer security, pattern recognition, and computer networks.

MOSTAFA Z. ALI (Senior Member, IEEE) received the bachelor's degree in applied mathematics from the Jordan University of Science and Technology (JUST), Irbid, Jordan, in 2000, the master's degree in computer science from the University of Michigan-Dearborn, MI, USA, in 2003, and the Ph.D. degree in computer science/artificial intelligence from Wayne State University, MI, USA, in 2008. He is currently a Professor with the Department of Computer Information Systems, Jordan University of Science and Technology, Irbid. His research interests include evolutionary computation, machine learning, deep learning, virtual/augmented reality, and data mining. He is a member of the IEEE Computer Society, the American Association of Artificial Intelligence (AAAI), and the ACM. He is an Associate Editor of the *Swarm and Evolutionary Computation (SWEVO)* (Elsevier).

MOHAMMAD I. DAOUD received the Ph.D. degree in electrical and computer engineering from the University of Western Ontario, London, Canada, in 2009. After his graduate studies, he worked at the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, Canada, as a Postdoctoral Research Fellow, where he held an NSERC Postdoctoral Fellowship. In September 2011, he joined the Department of Computer Engineering, German Jordanian University, Amman, Jordan, where he is currently a Full Professor. His research interests include image and signal processing, sensors, machine learning, and parallel computing.