CSI-IANet: An Inception Attention Network for Human-Human Interaction Recognition Based on CSI Signal

M. HUMAYUN KABIR¹, (Member, IEEE), M. HAFIZUR RAHMAN², (Member, IEEE), and WONJAE SHIN¹, (Senior Member, IEEE)

¹Department of Electrical and Computer Engineering, Ajou University, Suwon, Republic of Korea (e-mail: humayun_kabir@ieee.org; wjshin@ajou.ac.kr)
²Department of Electrical and Electronic Engineering, Islamic University, Kushtia, Bangladesh (e-mail: hafizur.iu@ieee.org)

Corresponding author: Wonjae Shin (e-mail: wjshin@ajou.ac.kr).

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (No. 2021-0-00467) grant, and the Basic Science Research Programs under the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (No. 2019R1C1C1006806, No. 2021R1A4A1030775).

ABSTRACT In recent years, Wi-Fi infrastructures have become ubiquitous, providing device-free passive-sensing features. Wi-Fi signals can be affected by their reflection, refraction, and absorption by moving objects in their path. The channel state information (CSI), a signal property indicator, of the Wi-Fi signal can be analyzed for human activity recognition (HAR). Deep learning-based HAR models can enhance performance and accuracy without sacrificing computational efficiency. However, to save computational power, an inception network, which uses a variety of techniques to boost speed and accuracy, can be adopted. In contrast, the concept of spatial attention can be applied to obtain refined features. In this paper, we propose a human–human interaction (HHI) classifier, CSI-IANet, which uses a modified inception CNN with a spatial-attention mechanism. The CSI-IANet consists of three steps: i) data processing, ii) feature extraction, and iii) recognition. The data processing layer first uses the second-order Butterworth low-pass filter to denoise the CSI signal and then segment it before feeding it to the model. The feature extraction layer uses a multilayer modified inception CNN with an attention mechanism that uses spatial attention in an intense structure to extract features from captured CSI signals. Finally, the refined features are exploited by the recognition section to determine HHIIs correctly. To validate the performance of the proposed CSI-IANet, a publicly available HHI CSI dataset with a total of 4800 trials of 12 interactions was used. The performance of the proposed model was compared to those of existing state-of-the-art methods. The experimental results show that CSI-IANet achieved an average accuracy of 91.30%, which is better than that of the existing best method by 5%.

INDEX TERMS Channel State Information (CSI), Convolution Neural Network (CNN), Deep Learning, Inception Module, Spatial Attention.

I. INTRODUCTION

HUMAN-human activity recognition (HAR) is a fast-paced and demanding research area. The human ability to recognize another person’s activities is of great interest in the fields of machine learning and Wi-Fi vision. Several applications, including surveillance cameras, human–computer interactions, and robots for human behavior characterization, require multiple activity detection systems as a consequence of this study. Traditional activity-recognition systems use image sensors [1], wearable sensors [2], RFID [3], RADAR [4], and other special-purpose devices. There are some limitations that affect their performance: The image sensor-based activity recognition methods produce false positive results owing to the deviation of the line of sight, illumination condition, and view-angle and run the risk of privacy leakage. For wearable sensing, users need to put on the sensing devices while monitoring, which can be uncomfortable. Radar-based approaches are expensive irrespective of the coverage range.

According to several studies, indoor human activities can be detected by examining the characteristics of Wi-Fi signals that are influenced by their activity [5]. Therefore, human activities can be recognized by analyzing the pattern of Wi-Fi
signals that are affected by the object in the propagation path. Wi-Fi signals offer a wider range of coverage than traditional RF-based sensing technologies. In addition, Wi-Fi signals are noninvasive, which protects users’ privacy, and human activity identification techniques based on Wi-Fi signals are device-free and do not require users to put on sensors. Hence, Wi-Fi signals can be utilized to replace traditional sensing technologies in activity recognition because of these advantages. Wi-Fi signals may travel through doors, furnishings, and windows. Wi-Fi signals can be analyzed by exploiting the signal properties in two ways: the received signal strength indicator (RSSI) and channel state information (CSI). The RSSI signal is currently the most widely utilized signal in indoor positioning [6], tracking [7], and radio tomographic imaging (RTM) [8]. However, RSSI is unable to function well in complicated scenarios owing to multi-way fading and time/dynamic properties. For other types of wireless signals (e.g., CSI signals), the amplitude of the signal transmission channel and the response to each subcarrier step can be expressed as a complex matrix. The quality of a channel can be evaluated by calculating the amplitude and frequency at the receiver end for each channel using a complex number. The signal power attenuation induced by the multipath effect is thus demonstrated by the amplitude of the CSI signal. When compared to RSSI, this is a fine-grained signal property representation of the wireless connection. This has a broad range of applications, including respiration detection [9], gesture recognition [10], and human behavior identification [11], and has shown excellent results. Consequently, the focus of this study is on HAR using Wi-Fi signals based on CSI.

Most existing CSI signal classification methods use statistical features that are extracted manually from the CSI signal. These handcrafted features are then analyzed using traditional machine learning classifiers, such as the hidden markov model (HMM) [12], random forest, and support vector machine (SVM) to classify CSI signals. Despite the positive results obtained with handcrafted features, extracting new features to characterize the information irrespective of time, frequency, and spatial domains is considered difficult. Deep convolutional neural networks may be used to learn deep features from input signals without having to construct them explicitly. The CSI captures the variations in the amplitude and phase information associated with different subcarrier frequencies of a Wi-Fi channel. Multi-path effects and the presence of moving objects in the signal propagation route affect the amplitude and phase information of the CSI signals. Changes in the amplitude of the CSI signals were more stable than changes in the phase information. Hence, we focused on the amplitude of CSI to build the model.

Despite the impressive performance of current CSI-based HAR methods, these methods were primarily focused on recognizing single-person actions conducted by a single individual [13]–[15]. These methods may not be applicable for detecting multi-person activity in real-world scenarios. Previous research has shown that the difficulty of identifying human–human interactions (HHIs), which involve multiple interacting people (e.g., high five and pushing interactions), is more challenging than identifying single-human activities (e.g., running and sitting activities) [16]. A three-layer CNN [17] is proposed, which employs publicly available CSI data [18], converting it into a 2D grayscale image to recognize the HHIs. This approach did not use any denoising, and the same time lost certain important features while converting the grayscale image.

To address these issues, a CNN design is proposed that employs both the inception module and the attention mechanism and is called CSI based inception attention network (CSI-IANet). It is an inception CNN with an attention mechanism that uses spatial attention in an intense structure. This network is utilized for the recognition of HHIs with CSI signals without converting it into other representations.

To summarize, the contributions of this paper are shown as follows:

1) To develop a CNN-based inception attention network (CSI-IANet) utilizing a spatial attention module.
2) To validate the effectiveness of the proposed model using publicly available datasets.
3) To verify the performance of the proposed model with that of other state-of-the-art models.

The remainder of this paper is organized as follows. In section II, we review related works of CSI signal-based HAR method. Section III presents details of the public HHI Datasets. Section IV describes the details of the system modeling including data processing, features extraction, recognition and methodology. The experimental results and discussion are presented in Section V and finally Section VI concludes the paper with a discussion on the future work.

II. RELATED WORKS

Sensing, recognition, and detection of humans are the driving factors for building a ubiquitous and pervasive indoor environment that can sense the environment and can act accordingly. Three types of approaches, vision-based, wearable-sensor-based, and RF-based approaches, are mainly applied for sensing, recognition, and detection [19]. Among the existing solutions, RF-based approaches are preferable because of their contactless and non-line-of-sight characteristics. The wireless signals transmitted from the transmitter propagate in the environment, which is reflected, refracted, and absorbed by the object and human presence before being received by the receiver. By analyzing the pattern of the received signal, it is possible to sense, recognize, and detect the target object. RF-based techniques, including RFID [3], Bluetooth [20], UWB [21] and Wi-Fi [22] are frequently used in this regard. The ubiquitously available infrastructure and the adaptation of the MIMO OFDM technique in Wi-Fi keeps it one step further than other RF-based techniques. The Wi-Fi signal can be analyzed using the two channel property indicators: received signal strength indicator (RSSI) and channel state information (CSI). The existing literature can be divided into two categories: RSSI-based and CSI-based technologies.
A. RSSI-BASED TECHNOLOGY

In the past decade, RSSI has been employed in studies on human positioning, human surveillance systems, and human activity analysis. RSSI is a device-bound technique that utilizes radio frequency sensing devices and uses the signal strength obtained under the direct influence of shadowing and multiway fading. The existence of a human between the wireless links reduces the strength of Wi-Fi signal, so the discrepancy between the signal intensity broadcasted and received can be computed. Although this is a fundamental and simple strategy, it is challenging to record changes in the signals in real time. Moore et al. [23] suggested a human movement detection method that keeps track of the variations in the default signal strength considering fixed wireless transmitters and receivers. An RSSI-based environment tracking system was proposed by Kosba et al. [24], which monitors the variation in the environment when a human enters the area of interest. Yang et al. [25] introduced a hybrid approach to classify human intrusion patterns simultaneously. Booranaong et al. [26] introduced a human movement detection and tracking system based on the RSSI approach. It first captures and measures the RSSI signals due to human movement and then introduces a region selection technique for the identification of human motion.

Sigg et al. [27] presented a system for the recognition of human activity that considers the variation in RSSI signals. They recognized a number of human behaviors, including lying, moving, sitting, and crawling. Their technology obtained remarkable precision under various situations utilizing a universal software peripherals radio platform. A gesture identification system, WiGest, is proposed [28], which relies on the RSSI fluctuation induced by human hand gestures in test movements. WiGest identified different patterns of hand gestures and utilized one overhead and three overhead transmitters. The average accuracies for a single transmitter and three overhead transmitters were 87.5% and 96%, respectively. Gu et al. [29] demonstrated an HAR method using the Wi-Fi RSSI. From the RSSI signal, they manually extracted several representative features. Then, to recognize the simple activities of sitting, standing, and walking, a fusion method was developed. The average accuracy achieved ranged between 75% and 92.58%. However, the RSSI-based techniques suffer from the drawbacks of RSSI signal variation, which is caused by the varying environment. This may lead to an erroneous detection.

B. CSI BASED TECHNOLOGY

CSI has recently been used for indoor localization and activity classification because it provides a fine-grained representation of the wireless link compared with RSSI. Damodaran et al. [30] presented a device-free HAR and CSI fall detection system that identified five activities using long short-term memory. Linear discriminant analysis is used for feature extraction, and discrete wavelet transformation is used for noise removal during data preprocessing. This yields an average accuracy of approximately 95%. A reliable HAR framework Wi-Motion [31] is proposed, which uses amplitude and phase information from CSI to classify five common human activities. R-DEHM [32] is a modern method for robust duration estimation of human motion that employs CSI for motion detection to predict the presence, absence, and duration of human motion. Furthermore, CSI segmentation was used to estimate the motion duration, with an average accuracy of 94%. Chase [33] used all CSI subcarrier data to distinguish coarse movements such as standing, jogging, and moving hands. Unlike moving activities, hand movements use recurring patterns in a stable position. This method uses two ML techniques: k-nearest neighbor (kNN) and SVM. The Wi-Chase research claims that the performance can be updated utilizing additional CSI channel subcarriers with multiple access point (AP) and receiver links. The E-eyes [34] algorithm was presented to detect various indoor activities and walking directions. The E-eyes method calculated the correlation between known and unknown activities to identify unknown activity. Moreover, the E-eyes algorithm used CSI variance to distinguish between walking activity and in-place activity because walking activity causes more CSI variance than in-place activity. Subsequently, using Earth Mover’s Distance, in-place activities were detected based on similarities to known activities, and walking directions were identified using dynamic time warping. They also claim that the recognition accuracy increases for large packet transmission rates.

Wi-Fi CSI is used to identify vital indicators (respiration and heartbeat rates) in a smart healthcare system. Wang et al. [35] employed CSI phase information to discover the vital signs. The researchers employed multiple antennas at the AP end to increase the power of the reflected signal to detect heart rates and heart motions. Likewise, Liu et al. [36] used CSI to detect the respiratory rates. To enhance the signal quality, the AP and receiver were placed on opposite sides of the user in the test-bed scenario. They discovered that sleeping positions had an impact on the accuracy of respiration detection: when a person sleeps in the “Embryo,” “Block,” or “Yearner,” the back of the patient interrupts the Wi-Fi signal routes. As a result, the researchers concluded that users should move frequency-domain spectral measurements for detection. A gesture recognition system was proposed by Tian et al. [37] based on CSI signals. The main concept is to create a virtual antenna using the signals reflected by hand motions. To identify each hand action, they used an SVM. The proposed technique was tested with six hand gestures and was found to be 97% accurate on average.

A number of obstacles stand in the way of the creation of a reliable and efficient HHI recognition model. The first and most difficult task is to reduce noise from raw CSI that has been included in the received signal as a result of the carrier frequency offset (CFO). This is a typical issue caused by oscillator differences between the transmitter and receiver. The phase data of the received signal are changed by the CFO, making it impossible to determine whether the signal loss is due to CFO or human movement. This...
problem is handled by ignoring the phase received signal and focusing on the strength of the complex CSI, which includes adequate indication of a body’s movement. However, residual noise lowers the signal strength and can be compensated by using effective denoising algorithms. Another difficult task for optimal activity detection using CSI is feature processing. Certain implicit features that are useful for activity detection may be lost when features are extracted using handcrafted methods. Therefore, modern researchers have chosen deep learning with autonomous feature learning [38]–[40]. The end-to-end deep learning framework (E2EDLF) [17] consists of a three-layer CNN that can handle temporal and spatial features and utilizes publicly available CSI data [18]. They converted the raw 4-dimensional CSI data into 2D grayscale images to recognize HHIs for the first time and reported an 86.3% accuracy. This approach did not use any denoising and simultaneously lost some important features while converting CSI data into grayscale images.

Network design has become an essential aspect of the present research because how well a network is constructed depends on the performance of an application. Since the successful implementation of CNN, an extensive range of architectures has been developed, from a relatively simple LeNet to a complicated inception network. When prior models went deeper to improve performance and accuracy with time complexity, Inception net set a new standard in CNN classifiers and has been meticulously planned. It employs a variety of techniques to boost the speed and accuracy [41]. Recently, several researchers [42]–[44] have investigated another critical topic called attention to improve the performance of CNNs. Several prior studies on object identification have highlighted the importance of the attention process [45], [46]. It not only indicates where an object’s focusing points are, but also increases the interest representation. In this paper, we propose a CNN design that employs both the inception module and the attention mechanism, inspired by recent developments in deep learning. Here, we refer to the proposed model as CSI based inception attention network (CSI-IANet). It is an inception CNN with an attention mechanism that uses spatial attention in an intense structure. Instead of converting the raw CSI data to grayscale images, we directly utilized the raw CSI data, which preserved all the features. Moreover, we utilized a second-order Butterworth filter to denoise the raw CSI data. The proposed CSI-IANet shows better performance in terms of accuracy and number of interactions that are being recognized.

C. BACKGROUND OF CSI

CSI measures the channel features of a wireless communication system that integrates the effects of delay time, intensity reduction, and phase change [47]. A signal from the recipient is generally superimposed as scattering, diffraction, and reflectance events that occur in the passage of the signal channel. The fundamental objective of CSI is to adjust the communication system to the present channel circumstances. The multiantenna system ensures excellent dependability and high-speed connections. The entire wireless channel is split into several narrowband subcarriers in an orthogonal frequency-division multiplexing (OFDM) scheme. The communication system can be calculated as follows [47]:

$$y_i = H_i x_i + v, \quad i = 1, 2, 3, \ldots, N,$$  

where $H_i \in \mathbb{C}^{N_R \times N_T}$ denotes the CSI matrix of $i^{th}$ subcarrier, $v$ denotes the noise term, $N$ represents the number of OFDM subcarrier frequencies, and $y_i \in \mathbb{C}^{N_R \times 1}$ and $x_i \in \mathbb{C}^{N_T \times 1}$ is the $i^{th}$ received and transmitted signal.

$$H_i = \begin{bmatrix} h_i^{11} & h_i^{12} & \cdots & h_i^{1N_R} \\ h_i^{21} & h_i^{22} & \cdots & h_i^{2N_R} \\ \vdots & \vdots & \ddots & \vdots \\ h_i^{N_T 1} & h_i^{N_T 2} & \cdots & h_i^{N_T N_R} \end{bmatrix}$$  

where $h_i^{jk}$ is the CSI of the $i^{th}$ subcarrier for the link between the $j^{th}$ transmitted antenna and the $k^{th}$ receiving antenna. The $h_i^{jk}$ is a complex value, which can be represented as

$$h_i^{jk} = |h_i^{jk}| e^{j\angle h_i^{jk}}$$

where $|h_i^{jk}|$ and $\angle h_i^{jk}$ denote amplitude and phase respectively.

Therefore, one CSI measurement will contain $N$, CSI matrices with $N_{R_x} \times N_{R_y}$ dimensions, $N_{T_x}$ and $N_{R_y}$ denote the number of the transmit and receive antennas, respectively. The amplitude and phase information are included in the CSI measurements. The carrier frequency offset (CFO) frequently deteriorates phase information [13]. CSI has a somewhat steady amplitude and is commonly used for human identification [48]. In this study, we use CSI amplitude information to recognize HHIs.

III. DATASET DESCRIPTION

In this study, we used a CSI dataset of HHIs [18], which is available online to train and measure the performance of our model. This dataset has 12 distinct interactions made by 40 different pairs of participants from 66 participants who were willing to experiment in an indoor space. Each pair of participants engaged in ten trials of 12 different interactions: approaching ($I_1$), departing ($I_2$), hand shaking ($I_3$), high-five ($I_4$), hugging ($I_5$), kicking with the left leg ($I_6$), kicking with the right leg ($I_7$), pointing with the left hand ($I_8$), pointing with the right hand ($I_9$), punching with the left hand ($I_{10}$), punching with the right hand ($I_{11}$), and pushing ($I_{12}$). Therefore, they recorded a total of 4800 trials of 12 interactions. There are two types of intervals in each of the 12 HHIs: steady-state and interaction intervals. Within the steady-state duration, the two participants faced each other and did nothing. Within the interaction duration, the pair of participants performed one of the twelve different HHIs. As a result, the thirteenth interaction was recorded.
The recorded Wi-Fi signals transferred from a commercial off-the-shelf access point (AP) named Sagemcom 2704, to a desktop PC equipped with an Intel 5300 NIC with the help of the publicly accessible CSI tool [49]. The AP was set up to operate in the 2.4GHz band, with wireless channel number 6, a channel bandwidth of 20MHz, and an index 8 modulation coding scheme. The AP has two internal transmit antennas ($N_T = 2$) whereas the NIC has three external receive antennas ($N_R = 3$) and the resulting system has 2x3 Wi-Fi streams. The CSI tool captures the CSI for 30 subcarriers (i.e., $N_{sc} = 30$) uniformly distributed across the channel bandwidth of 20MHz. As a result, each packet contains 2x3x30 CSI values. Fig. 1 shows the number of samples in each interaction class of the datasets used. It shows that the steady-state interaction class has the highest number of 22792 samples. Pointing with the left hand and pointing with the right hand had almost equal numbers of samples. Similarly, interactions kicking with the left leg and kicking with the right leg have a close number of samples. Because of these variations in the number of samples for each interaction, this dataset is imbalanced in nature.

IV. SYSTEM MODELING

The proposed CSI signal classifier works in four sections: CSI data collection, data processing, feature extraction and recognition as shown in Fig. 2. The commercial off-the-shelf, Wi-Fi device was used as the transmitter to collect the CSI data. An Intel 5300 NIC interfacing with a personal computer was used as a receiver to collect CSI signals. Here, the online available public data set was utilized. Detailed descriptions of the datasets are included in previous III data description section. Noise may induce while propagation; thus, we used the second-order low pass Butterworth filter [50] to remove the noise. Next, a three-layer CNN with inception and spatial attention module is used to capture the features from the CSI data. The features are then classified into 13 different classes in recognition section.

A. DATA PROCESSING

In this section, the data pre-processing task is presented. Pre-processing was performed in two steps: 1. Denoising Filter and 2. Segmentation. Detailed descriptions of denoising filter and segmentation are provided in Sections IV-A1 and IV-A2 respectively.

1) DENOISING FILTER

The raw Wi-Fi CSI data obtained from the publicly available CSI dataset are four-dimensional tensors. These tensors describe the time (packet index), frequency (OFDM subcarrier frequencies), and spatial variations of the carrier frequency response values observed for a Wi-Fi system (i.e., pairs of transmit-receive antennas). High-frequency noise, outliers, and artifacts are induced in the raw Wi-Fi CSI data, which may decrease the recognition rate of the classifier. Therefore, it is necessary to eliminate this unwanted noise. Here, a second-order low-pass Butterworth filter is utilized to remove high-frequency noise. This filter can remove a significant amount of noise from CSI data. Fig. 3 shows the raw and filtered CSI signal of 1st subcarrier among the 30 subcarriers for the 1st transmit-receive antenna pairs of 13 HHIs. The shaded area indicates the steady state before and after performing any interactions. After denoising the four-dimensional filtered CSI data is converted into 2D matrix of dimension $D \times I$ that retains the time, frequency, and spatial
data. Where, \( D = N_p \times N_s \) (i.e. Number of subcarries, \( N_s = 30 \)), \( N_p = N_{Tx} \times N_{Rx} \), the number of transmit-receive antenna pairs \( (N_{Tx} = 2, N_{Rx} = 3) \) in the test-bed, where \( I \) denotes the number of packets recorded during a given trial.

2) SEGMENTATION
Segmentation is the process of partitioning signals into smaller segments, also called windows. This helps to resolve certain limitations due to data pre-processing issues. The first issue is that the recorded trials of data of different subjects have different lengths, which may limit the recognition process. Another issue is that the large length of recorded data requires high computational power, which consumes more time. To overcome these limitations, the window size is set at 256, and 50% of the window is overlapped. Moreover, the overlap window reduces noise caused by data truncation during the windowing process and improves efficiency by increasing the number of data points. Fig. 4 shows the segmentation process of CSI signals.

B. FEATURE EXTRACTION
In this paper, a convolutional neural network (CNN) design that employs both the inception module and the spatial attention mechanism is proposed. This CNN was utilized to extract both the temporal and spatial features. Here, the proposed model is termed an inception attention network (CSI-IANet). It is an inception CNN with an attention mechanism that uses both temporal and spatial features in an intense structure. The architecture of the model is shown in Fig. 5. It has four layers. First three layer used for extracting temporal and spatial features. Inception and spatial attention module are utilized in two layers to produce more refined features. Each layer uses different size of filter, pooling and stride. For normalization and activation batch normalization (Batch Norm) and Rectified Linear Unit (ReLU) is used respectively. A brief description of each component of the proposed CSI-IANet model is given here.

1) INCEPTION MODULE
Recently, inception nets have set a new standard for CNN classifiers. It reduces the computational complexity and improves the performance and accuracy compared to the conventional multilayer-based approach of CNN. It also employs a variety of techniques to boost the speed and accuracy [41]. The inception module is usually slightly wider than the deeper. The proposed CSI-IANet used a three-step approach for the inception module, and instead of maximum pooling (MaxPool), it utilizes average pooling (AvgPool). The dotted portion in Fig. 5 shows the architecture of the inception layer. The inception module uses the features from the previous layer. The first step will perform a convolution with a filter size of 1x1 and stride value of 1. The second step first performs a convolution with a filter size of 1x1 and stride value of 1, and then apply another 3x3 convolution with stride value 2. The last step in this inception module consists of using an average pooling with 3x3 filters and a stride value of 2, followed by a 3x3 convolution applied with stride 2. Finally, all the outcomes of the three steps were concatenated and passed through the next layer.

2) SPATIAL MODULE
Nowadays, the concept of the attention module was introduced to improve the performance of CNNs [42]–[44]. Several prior studies on object identification have highlighted the importance of the attention process [45], [46]. It not only indicates where an object’s focusing points are, but it also increases the interest representation. Many recent studies have revealed that typical fully convolutional networks provide local feature representations that can lead to object misclassification [51], [52]. To model different descriptive relationships regarding local feature representations, a spatial attention matrix is developed, which represents the spatial interactions between features of every two neighbors. The spatial attention module (SAM) concentrates on “where” and “which” information is the most significant to a section of the data. The average pooling and max pooling procedures are used first to calculate spatial attention, and then they are added elementally to provide a series of resilient features. Finally, the concatenated descriptor uses a convolutional layer to build a spatial attention map, which highlights or weakens the information in the inputs. A schematic representation of the SAM is presented in Fig. 6.

Let us consider the input features \( F \in \mathbb{R}^{C \times H \times W} \) which are given to two pooling layers to generate two 2D maps: \( F_{max}^{s} \in \mathbb{R}^{1 \times H \times W} \) and \( F_{avg}^{s} \in \mathbb{R}^{1 \times H \times W} \) where, \( C \) is the number of input channels, \( H \) and \( W \) are the height and width of \( F \) respectively. Subsequently a convolution operation is performed with the help of a single convolution kernel with a size of 7x7 filter. Lastly, a sigmoid activation function is applied to the convolutional procedure to create a feature map. Finally, a sigmoid activation function was applied to the convolutional procedure to create a feature map. In the spatial dimension, the output feature map matches the input feature map. \( F' \) represents the result of the spatial attention map \( (SAM(F)) \) element-wise multiplied by \( F \) which is passed to the next step. The mathematical expression of \( SAM \) and final output \( F' \) can be expressed as follows:

\[
SAM(F) = (f^{T \times 7}([AvgPool(F); MaxPool(F)]))
\]  
(4)

\[
SAM(F) = \sigma(f^{T \times 7}([F_{avg}^{s}; F_{max}^{s}]])
\]  
(5)

where \( \sigma \) is the sigmoid activation function.

\[
F' = SAM(F) \otimes F
\]
(6)

C. RECOGNITION
The fourth layer of the proposed CSI-IANet acts as a recognition phase. It consists of five sublayers: flatten layer, dropout
FIGURE 3. Visualization of CSI raw and filtered signal for different interactions.
The methodological steps involved in the proposed recognition method are described in the block diagram in Fig. 7. This was done in two phases. In the first phase, pre-processing of raw CSI signal and data split was performed, and in the second phase, model training and evaluation were performed.

Three steps must be followed to design a statistical model for classification: i. Model building, ii. Training and model validation, and iii. Model evaluation. The quality of model development and training depends on the amount of data with sufficient variety. Moreover, the proper selection of the hyperparameters (i.e., the number of epochs, learning rate, batch size, activation function, etc.) also provoked model quality. This study was performed using a publicly available CSI dataset. The training set was used to select the hyperparameters of the proposed model, and a validation set was used to evaluate its performance. The proposed CSI-IANet model was trained for up to 100 epochs with 64 batch sizes. An early stop callback for validation loss with 10 epochs of patience was used to end the training if no improvements were identified. The learning rate is a hyperparameter that governs how much the weights of the network need to be altered with respect to the loss gradient. The model can learn to best estimate the function given the available resources in a certain number of training epochs with a perfectly adjusted learning rate. In this study, a small learning rate is initiated. When validation accuracy did not improve in six consecutive epochs, the learning rate was updated by 0.75 times of its previous value. This model utilized Adam optimizer [52] to minimize error by setting parameters $\alpha = 0.001$ (learning rate), $\beta_1 = 0.9$ (decay rate for the first moment), $\beta_2 = 0.999$ (decay rate for the second moment) and $\epsilon = 1e-08$ (constant to sum of mini-batch variances). Finally, categorical cross-entropy was used to calculate the error for the optimizing algorithm.

A 10-fold cross-validation (CV) approach was used to train and evaluate the proposed CSI-IANet and compare its performance with other state-of-the-art techniques. Before training, the hyperparameters were defined as described in the previous section. The labeled, segmented CSI data are processed from the CSI signals and divided into ten folds. As shown in Fig. 7, nine randomly selected folds were used for training, and the remaining fold was used for testing. This procedure was repeated ten times, and the overall recognition performance was calculated by averaging the results of each repetition. A desktop computer with Intel Core i7 3.90 GHz CPU and NVIDIA Titan XP Pro GTX1080Ti 12 GB GPU, 1 TB HDD, and 32 GB RAM were utilized for the experiment.

The network was run in a TensorFlow environment. For the evaluation of the proposed model, three metrics (accuracy, F1-score, and Cohen’s Kappa) have been reported. To obtain the reliability of the results, all data were evaluated using 10-

| Section          | Layer Type                  | Patch size/Stride | Output size   |
|------------------|-----------------------------|-------------------|---------------|
| Feature Extraction | Convolution                | 3×3/1             | 256×178×32    |
|                   | Batch Normalization         |                   | 256×178×32    |
|                   | Linear Rectified Unit (ReLU)|                   | 256×178×32    |
|                   | Max pool                   | 3×3/1             | 84×59×32      |
|                   | Inception (a)              |                   | 42×30×96      |
|                   | SAM                        |                   | 42×30×96      |
|                   | Convolution                | 3×3/1             | 42×30×64      |
|                   | Batch Normalization         |                   | 42×30×64      |
|                   | Linear Rectified Unit (ReLU)|                   | 42×30×64      |
|                   | Max pool                   | 3×3/1             | 14×10×64      |
|                   | Inception (b)              |                   | 7×5×192       |
|                   | SAM                        |                   | 7×5×192       |
|                   | Convolution                | 3×3/1             | 7×5×128       |
|                   | Batch Normalization         |                   | 7×5×128       |
|                   | Linear Rectified Unit (ReLU)|                   | 7×5×128       |
|                   | Max pool                   | 3×2/1             | 2×2×128       |

| Recognition | Flatten                  | 1×512               |
|            | Dropout 20%              | 1×512               |
|            | Dense                    | 1×256               |
|            | Softmax                  | 1×128               |

layer, dense layer-1, dense layer-2 and softmax layer. The refined feature obtained from the previous layer is passed through the flattened layer of the recognition phase. Subsequently, the dropout layer could deactivate 20% of neurons to avoid overfitting. Dense layer-1 is composed of 256 neurons and utilized the ReLU activation function. In contrast, the dense layer-2 used 128 neurons with the ReLU activation function. Finally, the softmax layer classifies the CSI signals into 13 different groups. A summary of the different layers of the proposed model is presented in Table 1.

**D. METHODOLOGY**

The methodological steps involved in the proposed recognition method are described in the block diagram in Fig. 7. This was done in two phases. In the first phase, pre-processing of raw CSI signal and data split was performed, and in the second phase, model training and evaluation were performed.

Three steps must be followed to design a statistical model for classification: i. Model building, ii. Training and model validation, and iii. Model evaluation. The quality of model development and training depends on the amount of data with sufficient variety. Moreover, the proper selection of the hyperparameters (i.e., the number of epochs, learning rate, batch size, activation function, etc.) also provoked model quality. This study was performed using a publicly available CSI dataset. The training set was used to select the hyperparameters of the proposed model, and a validation set was used to evaluate its performance. The proposed CSI-IANet model was trained for up to 100 epochs with 64 batch sizes. An early stop callback for validation loss with 10 epochs of patience was used to end the training if no improvements were identified. The learning rate is a hyperparameter that governs how much the weights of the network need to be altered with respect to the loss gradient. The model can learn to best estimate the function given the available resources in a certain number of training epochs with a perfectly adjusted learning rate. In this study, a small learning rate is initiated. When validation accuracy did not improve in six consecutive epochs, the learning rate was updated by 0.75 times of its previous value. This model utilized Adam optimizer [52] to minimize error by setting parameters $\alpha = 0.001$ (learning rate), $\beta_1 = 0.9$ (decay rate for the first moment), $\beta_2 = 0.999$ (decay rate for the second moment) and $\epsilon = 1e-08$ (constant to sum of mini-batch variances). Finally, categorical cross-entropy was used to calculate the error for the optimizing algorithm.

A 10-fold cross-validation (CV) approach was used to train and evaluate the proposed CSI-IANet and compare its performance with other state-of-the-art techniques. Before training, the hyperparameters were defined as described in the previous section. The labeled, segmented CSI data are processed from the CSI signals and divided into ten folds. As shown in Fig. 7, nine randomly selected folds were used for training, and the remaining fold was used for testing. This procedure was repeated ten times, and the overall recognition performance was calculated by averaging the results of each repetition. A desktop computer with Intel Core i7 3.90 GHz CPU and NVIDIA Titan XP Pro GTX1080Ti 12 GB GPU, 1 TB HDD, and 32 GB RAM were utilized for the experiment.

The network was run in a TensorFlow environment. For the evaluation of the proposed model, three metrics (accuracy, F1-score, and Cohen’s Kappa) have been reported. To obtain the reliability of the results, all data were evaluated using 10-
**FIGURE 5.** Proposed Model Architecture.
fold cross-validation. One of the most prevalent evaluation metrics in classification issues is accuracy, which is defined as the total correctly identified predictions divided by the total of predictions produced given a dataset. Accuracy is adequate when the target class is well balanced, but it is not a wise choice when the target class is unbalanced. As the dataset was slightly unbalanced, hence, for the complete picture of the model evaluation, other metrics such as F1-score and Cohen’s Kappa ($k$-score) were considered. The values used for the calculation are listed in Table 2 and equation (7)-(10). Here, true positive ($TP$) is a result in which the model accurately identifies the positive class, true negative ($TN$) is a result in which the model accurately identifies the negative class, false positive ($FP$) is a result in which the model incorrectly identifies the positive class, false negative ($FN$) is a result in which the model incorrectly identifies the negative class.

**TABLE 2.** Confusion matrix.

| Predicted Class | True Class | Class=1 | Class=0 |
|-----------------|------------|---------|---------|
| Class=1         | True Positive (TP) | False Positive (FP) |
| Class=0         | False Negative (FN) | True Negative (TN) |

**FIGURE 6.** Spatial Attention Module.

**FIGURE 7.** Overall system architecture.
TABLE 3. The interpretation of the Cohen’s kappa score (k-score).

| k-Score Value | Interpretation         |
|---------------|------------------------|
| k-score ≤ 0   | Poor agreement         |
| 0 ≥ k-score ≤ 0.2 | Slight agreement    |
| 0.2 ≥ k-score ≤ 0.4 | Fair agreement       |
| 0.4 ≥ k-score ≤ 0.6 | Moderate agreement   |
| 0.6 ≥ k-score ≤ 0.8 | Substantial agreement |
| 0.8 ≥ k-score ≤ 1 | Almost perfect agreement |

The F1-score represents the harmonic mean of the two measures (recall and precision). The numerical value starts from 0 to 1, where 0 stands for worst value whereas 1 stands for best value. In case of imbalanced number of sample datasets in interested classes, the F1-score can utilize to evaluate the recognition performance efficiently [39], [40], [48]. On the other hand, the Cohen’s Kappa score (k-score) can measure the agreement between the projected classes and the real classes that match them, eliminating any coincidences. The Cohen’s Kappa score [41], [49] in particular, allows us to evaluate the recognition performance produced by random guessing based on the number of samples in each class. The significance of the Cohen’s Kappa score (k-score) is elaborated in Table 3.

V. RESULT AND DISCUSSION

The proposed CSI-IANet was evaluated, and its performance was compared with other state-of-the-art techniques. The evaluation results show that the proposed model outperforms existing techniques. In this section, the details of the evaluation results are presented with a proper explanation. The proposed CSI-IANet model obtained an average recognition accuracy of 91.30% across the 13 HHI classes. A confusion matrix with a heatmap of the proposed CSI-IANet is shown in Fig. 8. Thirteen different HHIs are considered here. The average recognition accuracies for each of the 13 classes are displayed on the main diagonal of the confusion matrix. Misclassifications occur for two reasons: some interactions are quite similar, and the beginning and end of certain interactions are identical to steady-state interactions. There is some overlap for a couple of interactions because of the similarities between the interactions. From the confusion matrix, it is assumed that some misclassifications arise for interactions between punching with the left hand and punching with the right hand. Similarly, a mismatch also arises for the interaction of kicking with the left leg and kicking with the right leg. In addition, misclassification may occur because of the similarities between steady-state interactions with other HHIs (hand shaking, high fives, pointing with left hand, and pointing right hand) as the beginning and end of these interactions are identical. Fig. 9 shows the accuracy and F-1 measure in each of the interaction class.

For performance evaluation of the proposed CSI-IANet model, accuracy, F1-score, and Cohen’s Kappa (k-score) were utilized. The fold-wise results of different performance metrics (accuracy, F1-score, and Cohen’s Kappa (k-score)) are tabulated in Table 4. It shows that the fifth fold yields the highest results for accuracy, F1-score, and k-score, which are 91.98%, 0.92, and 0.90, respectively. Moreover, the second fold yielded the lowest values for accuracy, F1-score, and k-score were 90.46%, 0.90, and 0.88, respectively. However, there was no major fluctuation in the results for individual folds, and they provided almost similar results. Accuracy was calculated as a percentage.

The t-SNE algorithm was applied to visualize these features to understand how the proposed model represents the CSI data in the high-dimensional feature space. To do this, first, the feature vector is extracted from the previous classification layer of the proposed model. Next, t-SNE is applied to map the features onto a 2D space and then visualize the embedding representations of the dataset. Fig. 10 clearly shows 13 well-separated clusters of CSI data. The clear and wide margin among the 13 classes shows how well the CSI data are separated in the feature space. This indicates that the distributions of the features are quite different, demonstrating the good generalization capabilities of the proposed model.

We used 10-fold cross validation to test and train the proposed model. Table 4 shows that the 5th fold achieves the highest accuracy, F1-score and k-score among the 10-fold. Therefore, the training and test accuracy and loss curve for 5th fold of 10-fold cross validation are presented in Fig. 11 for better intuition. This shows that the accuracy and loss curve became steady after 60 epochs.

To evaluate the performance of the proposed model, it was compared with three state-of-the-art techniques. The pretrained CNNs, ResNet-50, Inception-V3, and DenseNet-121 were utilized for comparison. The number of neurons in the last layer was set to 13. In addition, the number of epochs was set to 50, and the Adam optimizer algorithm was used to tune the pretrained models. Moreover, the proposed model was also compared with the E2EDLF [17] to recognize HHIs. The performance comparison of the proposed CSI-IANet with other state-of-the-art techniques is presented in Table 5. The average recognition accuracies computed across all thirteen HHI classes for the ResNet-50, Inception-V3, DenseNet-121, and E2EDLF are 68.45%, 70.12%, 69.59%, and 86.30% respectively. The average recognition accuracies computed across all 13 HHI classes for ResNet-50, Inception-V3, DenseNet-121, and E2EDLF were 68.45%, 70.12%, 69.59%, and 86.30%, respectively. The average recognition F1-score computed across all HHI classes for ResNet-50, Inception-V3, DenseNet-121, and E2EDLF were 0.68, 0.70, 0.70, and 0.86, respectively. Furthermore, the average k-score com-
The confusion matrix of the proposed CSI-IANet model for HHI recognition.

The result obtained from 10-fold CV of the proposed CSI-IANet model.

| Metrics                  | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Fold 6 | Fold 7 | Fold 8 | Fold 9 | Fold 10 | Average   |
|--------------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|----------|-----------|
| Accuracy (%)             | 90.56  | 90.46  | 91.78  | 90.31  | 91.98  | 91.09  | 91.25  | 91.76  | 91.91  | 91.85    | 91.30±0.62 |
| F1-Score                 | 0.91   | 0.90   | 0.92   | 0.90   | 0.92   | 0.91   | 0.92   | 0.92   | 0.92   | 0.92     | 0.91±0.0062 |
| Cohen’s Kappa (K-Score)  | 0.89   | 0.88   | 0.90   | 0.88   | 0.89   | 0.89   | 0.89   | 0.90   | 0.90   | 0.90     | 0.89±0.0077 |

The performance analysis of the proposed CSI-IANet model demonstrates that it outperforms the existing best model E2EDLF by 5% in terms of accuracy, F1-score, and k-score. This improvement might be due to the new architecture of the proposed model and the optimal hyper-parameter selection. Thus, our proposed model can be used for the recognition of HHIs.

The runtime of the proposed CSI-IANet for training and recognition was calculated and compared with those of ResNet-50, Inception-V3, DenseNet-121, and E2EDLF techniques. Table 6 tabulates the runtime comparison between the proposed CSI-IANet with others, in terms of training and recognition time in average ± standard deviation values. All the time values were measured over ten repetitions of the 10-fold cross validation procedure.
TABLE 5. Performance evaluation comparison of CSI-IANet model.

| Classifier       | Accuracy (%) | F1-score     | k-score  | Number of trainable Parameter |
|------------------|--------------|--------------|----------|------------------------------|
| ResNet-50        | 68.45±0.47   | 0.682±0.0017 | 0.662±0.0014 | 26,667                       |
| Inception-V3     | 70.12±0.56   | 0.701±0.0031 | 0.692±0.0036 | 26,667                       |
| DenseNet-121     | 69.59±0.37   | 0.694±0.0023 | 0.674±0.0043 | 13,355                       |
| E2EDLF           | 86.30        | 0.86         | 0.85     | 935,053                      |
| Proposed CSI-IANet | 91.30±0.62   | 0.913±0.0062 | 0.894±0.0077 | 546,321                      |

TABLE 6. Runtime comparison.

| Model            | Training time (sec) | Recognition time (sec) |
|------------------|---------------------|------------------------|
| ResNet-50        | 12715.67±1436       | 0.0021±0.00015         |
| Inception-V3     | 8798.16±655         | 0.0019±0.000021        |
| DenseNet-121     | 9867.86±149         | 0.0020±0.000023        |
| E2EDLF           | 934.27±3.56         | 0.0002±0.000018        |
| Proposed CSI-IANet | 5497.35±1.7        | 0.0003±0.000025        |

In addition to obtaining refined features, the spatial attention model has also been utilized. The proposed classifier was composed of three sections. The data processing section applies a Butterworth low-pass filter to denoise the CSI signal and perform segmentation. The raw data are used to preserve more features other than conversion into another representation. Then, the feature extraction layer utilizes the inception module with spatial attention to obtain the refined feature that is fed to the recognition layer. The recognition layer utilized a flatten, dropout, dense, and softmax layer to classify it into 13 different activities. The proposed CSI-IANet shows better performance in terms of accuracy and number of interactions that are being recognized. In the future, we can adopt channel attention with the spatial attention module to obtain more refined features.

VI. CONCLUSION

This study developed a CSI-based inception attention network (CSI-IANet) for human–human interaction recognition. Instead of using deep learning, we utilized an inception module that widens the network to save computational power.

REFERENCES

[1] J. K. R. Aggarwal, S. Michael, “Human activity analysis: a review,” ACM Computing Surveys, vol. 43, no. 3, pp. 1-43, 2011.
FIGURE 10. Two-dimensional t-SNE visualization of the learned representations of the proposed CSI-IANet model visualized 10% data of entire dataset.

[2] O. D. Lara and M. A. Labrador, “A survey on human activity recognition using wearable sensors,” IEEE Commun. Surveys Tuts., pp. 1192-1209, 2013.

[3] S. M. Flores, B. Sareen, and A. Vagga, “Performance of RFID tags in near and far field in personal wireless communications,” in Proc. IEEE Intl. Conf. on ICPCWC, 2005.

[4] X. Li, Y. He, and X. J. R. S. Jing, “A survey of deep learning-based human activity recognition in radar,” Remote sensing, vol. 11, no. 9, pp. 1068, 2019.

[5] F. Adib and D. Katabi, “See through walls with Wi-Fi,” in Proc. of the ACM SIGCOMM 2013 conference on SIGCOMM, pp. 75-86, 2013.

[6] M. Yousefi, M. Mah, and A. Agravala, “Challenges: device-free passive localization for wireless environments,” in Proc. of the 13th annual ACM intcl. conf. on Mobile computing and networking, pp. 222-229, 2007.

[7] J. Wilson and N. Patwardi, “See-through walls: motion tracking using variance-based radio tomography networks,” IEEE Trans. Mobile Comput., vol. 10, no. 5, pp. 612-621, 2010.

[8] J. Wilson and N. Patwardi, “Radio tomographic imaging with wireless networks,” IEEE Trans. Mobile Comput., vol.9, no. 5, pp. 621-632, 2010.

[9] 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, 2015.

[10] M. Raja, V. Ghaderi, and S. Sigg, “WiBot! In-vehicle behaviour and gesture recognition using wireless network edge,” in IEEE 38th Intl. Conf. on Distributed Computing Systems (ICDCS), pp. 376-387, 2018.

[11] K. Ohara, T. Maekawa, and Y. Matsushita, “Detecting State changes of indoor everyday Objects using Wi-Fi channel state information,” in Proc. ACM interactive, mobile, wearable Ubiquitous Technol., vol. 1, no. 3, pp. 1-28, 2017.

[12] M. H. Kabir, M. R. Hoque, K. Thapa, and S.-H. Yang, “Two-layer hidden Markov model for human activity recognition in home environments,” Intl. Journal of Distributed Sensor Networks, pp. 1-12, January 2016.

[13] S. Yousef, H. Narui, S. Dayal, S. Ermon, and S. Valaee, “A survey on behavior recognition using Wi-Fi channel state information,” IEEE Commun. Mag., vol. 55, no. 10, pp. 98-104, 2017.

[14] C. Feng, S. Arshad, S. Zhou, D. Cao, and Y. Liu, “Wi-Multi: A three-phase System for Multiple Human Activity Recognition with Commercial Wi-Fi Devices,” IEEE Internet Things J., vol. 6, no. 4, pp. 7293-7304, 2019.

[15] Z. Chen, L. Zhang, C. Jiang, Z. Cao, and W. Cui, “Wi-Fi CSI based passive human activity recognition using attention based BLSTM,” IEEE Trans. Mob. Comput., vol. 18, no. 11, pp. 2714–2724, 2019.

[16] L. Wang, T. Gu, X. Tao, H. Chen, J. Lu, “Recognizing multi-user activities using wearable sensors in a smart home,” Pervasive and mobile computing, vol. 7, no. 3, pp. 287-298, 2011.

[17] R. Alazrai, M. Hababeh, B. A. Alsafiy, M. Z. Ali, and M. I. Daoud, “An end-to-end deep learning framework for recognizing human-to-human interactions using Wi-Fi signals,” IEEE Access, vol. 8, pp. 197695–197710, 2020.

[18] R. Alazrai, A. Awad, A. Baha’A, M. Hababeh, and M. Daoud, “A dataset for Wi-Fi-based human-to-human interaction recognition,” Data Br., vol. 31, p. 105668, 2020.

[19] Y. He, Y. Chen, Y. Hu, and B. Zeng, “Wi-Fi vision: sensing, recognition, and detection with commodity MIMO-OFDM Wi-Fi,” J. Internet of things, vol. 7, no. 9, pp. 8296-8317, 2020.

[20] F. Naya, H. Noma, R. Ohmura, and K. Kogure, “Bluetooth-based indoor proximity sensing for nursing context awareness,” in 9 th IEEE Intl. Symposium on wearable computers (ISWC’05), pp. 212-213, 2005.

[21] D. Dardari, A. Conti, U. Fierrez, A. Giorgetti, and M. Win, “Ranging with ultra wideband signals in multipath environments,” in Proc. IEEE Proc. IRE* (through 1962), vol. 97, no. 2, pp. 404-426, 2009.

[22] H. Jiang, C. Cai, X. Ma, Y. Yang, and J. Liu, “Smart home based on Wi-Fi sensing: A survey,” in IEEE Access, vol. 6, pp. 13317-13325, 2018.

[23] R. Moore, R. Howard, P. Kuksa, and R. P. Martin, “A geometric approach to device-free motion localization using signal strength,” Technical Report DCST-TR-674, pp. 1-11, 2010.

[24] A. E. Kosba, A. Saeed, and M. Youssef, “Robust WLAN device-free passive motion detection,” in Proc. of IEEE Conf. on Wireless Communications and Networking (WCNC), pp. 3284-3289, 2012.

[25] J. Yang, Y. Ge, H. Xiong, Y. Chen, and H. Liu, “Performing joint learning for passive intrusion detection in pervasive wireless environments,” in Proc. of IEEE Conf. on computer communications (INFOCOM), pp. 1-9, 2010.

[26] A. Booranawong, N. Jindapetch, and H. Saiyo, “A system for detection and tracking of human movements using RSSI signals,” IEEE Sensor J., vol. 18, no. 6, pp. 2531-2544, 2018.

[27] M. S. Sigg, S. Shi, Y. J. Ji, and M. Beigl, “RF-sensing of activities from non-cooperative subjects in device-free recognition systems using ambient and local signals,” IEEE Trans. Mobile Comput., vol. 13, no. 4, pp. 907-920, 2014.

[28] H. Abdelnasser, M. Youssef, and K. A. Harras, “Wigest: A ubiquitous Wi-Fi-based gesture recognition system,” in Proc. of IEEE Conf. on computer communications (INFOCOM), pp. 1472-1480, 2015.

[29] Y. Gu, L. Quan, and F. Ren, “Wi-Fi-assisted human activity recognition,” in Proc. of IEEE Asia Pacific Conf. on Wireless and Mobile, pp. 60-65, 2014.

[30] N. Domodaran, E. Haruni, M. Kokkharova, J. Schäfe, “Device free human activity and fall recognition using Wi-Fi channel state information(CSI),” CCF Trans. Pervasive Comput. Interact, vol. 2, no. 1, pp. 1-17, 2020.

[31] J. Zhao, L. Liu, Z. Wei, C. Zhang, W. Wang, and Y. J. S. Fan, “ R-DEHM: CSI-based robust duration estimation of human motion with WiFi,” Sensors, vol. 19, no. 6, pp. 1421, 2019.

[32] H. Li, Y. He, X. Chen, Y. Fang, and Q. J. I. A. Fang, “Wi-motion: A robust human activity recognition using Wi-Fi signals,” IEEE Access, vol. 7, pp. 153287-153299, 2019.

[33] J. L. Y. Wang, Y. Chen, M. Gruteser, J. Yang, and H. Liu, “E-eyes: Device-free location-oriented activity identification using fine-grained Wi-Fi signatures,” in Proc. Annu. Int. Conf. Mob. Comput. Networking, MOBICOM, pp. 617-628, 2014.

[34] S. Arshad, C. Feng, C. Y. Liu, Y. Hu, R. Yu, S. Zhou, H. Li, “Wi-chase: A WiFi based human activity recognition system for sensorless environments,” in IEEE 18th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2017.

[35] X. Wang, C. Yang, and S. Mao “PhaseBeat: Exploiting CSI phase data for vital sign monitoring with commodity Wi-Fi devices,” IEEE 37th Intl. Conf. on Distributed Computing Systems (ICDCS), pp. 1230-1239, 2017.

[36] M. De Sanctis, E. Cianca, S. Di Domenico, D. Provenziani, G. Bianchi, and M. Ruggieri, “Wibecam: Device free human activity recognition through Wi-Fi beacon-enabled camera,” in Proc. of the 2 nd Workshop on Physical analytics, pp. 7-12, 2015.

[37] Z. Shi, J. A. Zhang, R. Xu, and Q. Cheng, “Deep learning networks for human activity recognition with CSI correlation feature extraction,” in Proc. of IEEE Intl. Conf. on Communications (ICC), pp. 1-6, 2019.

[38] H. Yan, Y. Zhang, Y. Wang, and K. Xu, “WiAct: A passive Wi-Fi based human activity recognition system,” IEEE Sensors Journal, vol. 20, no. 1, pp. 296-305, 2019.
Accuracy
Number of epochs
Loss
Training loss
Testing loss
Training accuracy
Testing accuracy

FIGURE 11. Accuracy and loss curve for the 5th fold training and testing.

[39] Z. Shi, J. A. Zhang, R. Xu, and G. Fang, “Human activity recognition using deep learning networks with enhanced channel state information,” in IEEE Globecom. Workshops (GC Work shop), pp. 1-6, 2018.
[40] “ML | Inception Network”, [online]. Available: https://www.geeksforgeeks.org/ml-inception-network-v1/, Accessed on: Oct 2021.
[41] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, “Residual attention network for image classification,” in Proc. of the IEEE Conf. on computer vision and pattern recognition, pp. 3156-3164, 2017.
[42] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in Proc. of the European Conf. on computer vision (ECCV), pp. 3-19, 2018.
[43] J. Park, S. Woo, J.-Y. Lee, and I. Kweon, “BAM: Bottleneck attention module,” arXiv preprint arXiv:1807.06514 (2018), 2018.
[44] J. Ba, V. Mnih, and K. Kavukcuoglu, “Multiple object recognition with visual attention,” arXiv preprint arXiv:1412.7755 (2014), 2014.
[45] Y. Mnih, N. Heess, and A. Graves, “Recurrent models of visual attention,” Advances in neural information processing systems, pp. 2204-2212, 2014.
[46] Z. Yang, Z. Zhou, and Y. Liu, “From RSSI to CSI: Indoor localization via channel response,” ACM Computing Surveys (CSUR), vol. 46, no. 2, pp. 1-32, 2013.
[47] Y. Wang, K. Wu, and L. Ni, “Wifall: Device-free fall detection by wireless networks,” IEEE Trans. Mobile Comput., vol. 16, no. 2, pp. 581-594, 2016.
[48] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, “Tool release: Gathering 802.11n traces with channel state information,” vol. 41, no. 1, pp. 53-53, 2011.
[49] Y.-M. Fang, H.-L. Feng, J. Li, and G.-H. Li, “Stress wave signal denoising using ensemble empirical mode decomposition and an instantaneous half period model” Sensors, vol. 11, no. 8, pp. 7554-7567, 2011.
[50] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, “Large kernel matters: improve semantic segmentation by global convolutional network,” in Proc. of the IEEE conf. on computer vision and pattern recognition, pp. 4353-4361, 2017.
[51] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proc. of the IEEE conf. on computer vision and pattern recognition, pp. 2881-2890, 2017.
[52] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

***