I.J. Intelligent Systems and Applications, **2020, 6, 33-45**
Published Online December 2020 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2020.06.03

Deep Learning Sign Language Recognition System Based on Wi-Fi CSI

Marwa R. M. Bastwesy
Computers and Automatic Control Dept., Faculty of Engineering, Tanta University, Tanta, Egypt
E-mail: Marwa.Reda@f-eng.tanta.edu.eg

Nada M. ElShennawy
Computers and Automatic Control Dept., Faculty of Engineering, Tanta University, Tanta, Egypt
E-mail: Nada_elshennawy@f-eng.tanta.edu.eg

Mohamed T. Faheem Saidahmed
Computers and Automatic Control Dept., Faculty of Engineering, Tanta University, Tanta, Egypt
E-mail: Mohamed_ahmed1@f-eng.tanta.edu.eg

Received: 20 July 2020; Revised: 16 August 2020; Accepted: 23 August 2020; Published: 08 December 2020

Abstract: Many sensing gesture recognition systems based on Wi-Fi signals are introduced because of the commercial off-the-shelf Wi-Fi devices without any need for additional equipment. In this paper, a deep learning-based sign language recognition system is proposed. Wi-Fi CSI amplitude and phase information is used as input to the proposed model. The proposed model uses three types of deep learning: CNN, LSTM, and ABLSTM with a complete study of the impact of optimizers, the use of amplitude and phase of CSI, and preprocessing phase. Accuracy, F-score, Precision, and recall are used as performance metrics to evaluate the proposed model. The proposed model achieves 99.855%, 99.674%, 99.734%, and 93.84% average recognition accuracy for the lab, home, lab + home, and 5 different users in a lab environment, respectively. Experimental results show that the proposed model can effectively detect sign gestures in complex environments compared with some deep learning recognition models.

Index Terms: Wireless, Device-free sensing, Channel State Information, Sign Language Recognition, Deep Learning, WiFi Imaging.

1. Introduction

In recent literature, human activity recognition has attracted tremendous interest as it serves a wide range of applications such as gesture recognition [1], indoor localization, fall detection [2], activity recognition [3], and many other human-centric applications.

The activity and gesture recognition studies traditionally can be classified into two categories: computer vision-based approaches and wearable devices based approaches. The former includes video-based methods using cameras [4] and infrared depth sensors such as kinect [5]. The latter requires sensing devices like SignAloud [6]. However, most of these methods have some limitations: cameras based systems require the detected object to be within the coverage area, line-of-sight (LOS) and under sufficient lighting conditions. Also, it has a privacy problem. Sensor-based systems discomfort the user because the user should wear a special device all the time to detect the motion.

In recent years, another sensing-based gesture recognition method is introduced [7,8,9] based on Wi-Fi signals. It is concerned as a promising technology because of its advantages: low cost deployment and it does not require any special hardware. So, it is called device-free systems. Comparing to the previous approaches, wireless sensing technology can perceive an object in a passive manner over virtually any region at any time, regardless of ambient lighting conditions.

Wi-Fi signal has many features to support it in human activity and gesture recognition such as its better coverage, therefore it works in light-of-sight and non-light-of-sight (LOS/NLOS) scenarios, Wi-Fi signals can travel through and detect what is behind the wall for that reason it is called ”Walls have eyes” technology.

Recently, Wi-Fi based recognition systems are further popularized as it utilizes the commercial off-the-shelf Wi-Fi devices without the need for any additional special equipment.

By analyzing the Wi-Fi signals, a unique pattern is obtained which describes the environment based on the channel state information (CSI) attribute of Multiple-Input Multiple-Output (MIMO) in addition to Orthogonal Frequency Division Multiplexing (OFDM).

In SignFi [10] is a sign language recognition system based on Wi-Fi CSI. The authors presented a 9-layer CNN deep learning model for gesture recognition after pre-processing raw CSI measurements. SignFi achieves 98.91%, 98.01%, 94.81%, 86.66%, 98% recognition accuracy in the home, lab, home+lab, 5-users, and self test, respectively.

In this paper, a deep learning framework based on WiFi channel state information (CSI) based on the convolution neural network (CNN), Long Short Term Memory (LSTM), and Attention based Bidirectional Long Short-Term Memory (ABLSTM) are presented to recognize a sign word. The amplitude and phase information of a CSI waveform as features represented the correlation between the Wi-Fi CSI and the sign word are leveraged.

The main limitation of existing solutions is that all gesture recognition systems lack the ability to recognize gestures in a complex environment with different users.

The objective of this research is to improve the classifications of sign gestures in the complex and multi-human environment.

The main contributions of this paper are summarized as follows:

- Both CSI amplitude and phase are used as base signals to build a unique pattern for each sign gesture.
- We propose CNN deep learning framework for Wi-Fi CSI based gesture recognition.
- We study the impact of signal preprocessing, and the importance of CSI phase information in a multi-human environment.

The rest of this paper is organized as follows: background and the related works are reviewed in section II. In section III, the problem formulation and the proposed model are presented. The performance evaluation of the proposed model is introduced in section IV. Finally, the conclusions and future works are listed in section V.

2. Backgroung and Related Works

Wireless gesture recognition system based on Wi-Fi signals recently has great attention because of its features. A passive device-free wireless gesture recognition system based on Wi-Fi signals can be classified into three major categories: Radio Frequency (RF) based approaches, Received Signal Strength (RSS) based approaches, and Channel State Information (CSI) based approaches. Here, some of previous approaches are reviewed to describe the latest developments in device-free gesture recognition systems.

2.1. Radio Frequency (RF) based approaches

In this approach, a special purpose hardware called Software Defined Radio (SDR) that provides informative features of signals with devices like Universal Software Radio Peripheral (USRP) and Radio-frequency identification (RFID) readers are implemented for device-free recognition system. These systems have the ability to track human positions through walls and detect simple hand gestures.

In 2013, Pu et al. presented WiSee [9] recognized nine gestures and achieves 94% recognition accuracy by employing a USRP receiver to extract the Doppler shifts from the reflected Orthogonal Frequency Division Multiplexing (OFDM) Wi-Fi signals caused by body gestures.

In 2015, Wang et al. proposed the RF-IDraw system [11] which had been designed with commercial RFID readers and tags. RF-IDraw leverages the in-air hand gestures to interact with a device by a virtual touch screen. RF-IDraw achieves 97.5% character recognition accuracy and 92% word recognition accuracy. However, such systems suffer from burdensome cost installation as they require special purpose devices whose costs are high In contrast, CSI systems require only access point to transmit Wi-Fi signals which are now ubiquitously available everywhere and a receiver to receive signals.

2.2. Received Signal Strength (RSS) based approaches

Here, RSS values that can be extracted from Wi-Fi signals are used in recognition systems. RSS is a measure of signal amplitude.

In 2013, Sigg et al. introduced [12] RSS based recognition system. It achieved 71.6% recognition accuracy using Decision tree algorithm and 72.2% recognition accuracy kNN algorithm for walking, standing, lying down, and crawling activities.

In 2015, Abdelnasser et al. proposed WiGest [1] which builds on the commercial off-the-shelf (COTS) Wi-Fi devices and analyzes the variations in strength values of the received Wi-Fi signals compared to the transmitted signal, Received Signal Strength (RSS) values, due to hand movement. The WiGest system aims to track the user's hand gestures around his mobile without holding it and map them to perform a certain action. WiGest achieved 87.5% accuracy, using a single access point and 96% using three overhead access points. However, RSSI is an inefficient measure as it cannot give detailed information about each feature. Moreover, the performance of RSS-based systems degrades under complex environments owing to the fluctuations of RSSI values caused by multipath changes. Meanwhile, RSS-based gesture recognition systems lack the ability to detect gestures through walls and in a complex environment.

2.3. Channel State Information (CSI) based approaches

Channel State Information (CSI) is a more informative metric than RSSI. CSI contains unique fine-grained information for each gesture.

In 2015, a CSI based activity recognition system that leverages the correlation between the amplitude of CSI waveforms and different human activities is introduced [13]. CARM achieved more than 96% recognition accuracy between eight activities.

Also, in [14] a hand gesture recognition system based on CSI, WiG is proposed. WiG achieved a 92% and 88% recognition accuracy for line-of-sight(LOS) scenario and non-line of sight(NLOS) scenario, respectively.

In 2016, WiFinger is developed in [15] which is a finger gesture recognition system based on CSI values. WiFinger built a unique pattern in the amplitude of the CSI waveform for each finger gesture. WiFinger achieved more than 90% recognition accuracy for nine digits gesture from American Sign Language(ASL).

Also, In [8], WiGer is presented. It leverages the changes in the amplitude of CSI due to hand gestures. WiGer achieved more than 92% recognition accuracy in six scenarios for 7 gestures.

In 2018, WiCatch is presented [16]. It works based on the relationship between the variations of CSI and hand gestures to decrease the interference between signals and preserve the weak signals caused by hand gestures. WiCatch applies a novel interference elimination algorithm. WiCatch achieves more than 94% recognition accuracy between two hand gestures.

In summary, all previous approaches still suffer from an accurate detection gesture in complex environments. In contrast, this research is able to recognize the different hand, arm, head, and finger gestures with acceptable accuracy in different environments.

3. Proposed Model

3.1. System Overview

In this paper, a deep learning framework based on Wi-Fi channel state information (CSI) is proposed as shown in Fig.1. To recognize a sign word, three different deep learning methods are used: CNN, LSTM, and ABLSTM.

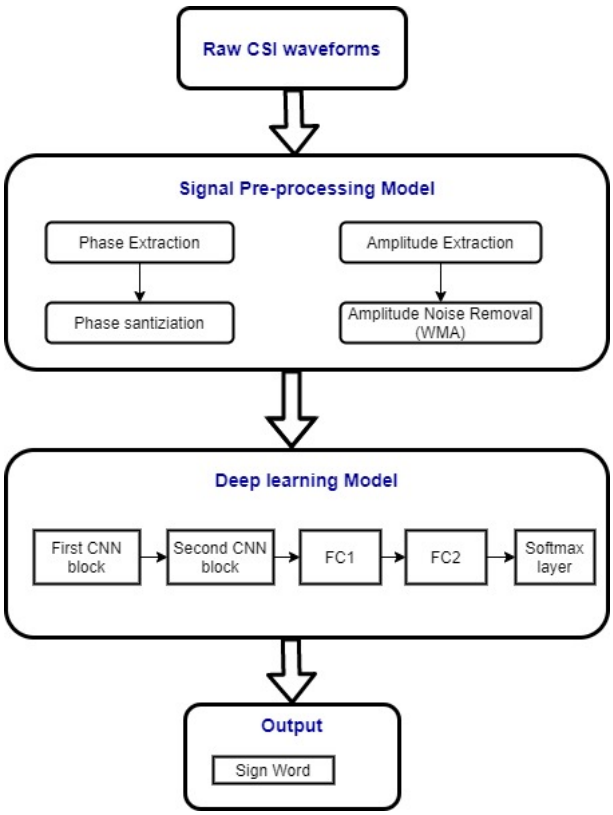


Fig.1. Flow chart of the proposed model

Our system consists of the following stages:

1. Input phase: Raw CSI data collected by SignFi[10]
2. Signal processing phase: weighted moving average on CSI amplitude and phase calibration algorithm on CSI phase to preserve the original waveform information
3. Classification phase: the combination of preprocessed amplitudes and phases is fed to the used CNN deep learning algorithm which consists of two convolution blocks and two fully connected layers followed by softmax layer which works as a classifier.

Our model is worked based on mainly three tiers. The first one is extracting Wi-Fi CSI amplitude and phase that represent the sign word. The second tier, the signal pre-processing phase. Finally, the classification phase is built.

3.2. The channel state information

Channel state information (CSI) is a fine grained informative and stable measurement. In wireless communication, CSI conveys the characteristics of the physical environment which describes how Wi-Fi signals propagate in the channel between the transmitter and the receiver with the effects of reflections and refractions.

In the frequency domain, the Wi-Fi channel with multiple transmitting and multiple receiving antennas (MIMO) can be expressed as in equation (1) [17]:

$$Y = H * X + N \tag{1}$$

where

Y	Received Signal Vector
H	complex channel matrix that contains the CSI values
X	Transmitted Signal Vector
N	Noise Vector

The CSI is estimated for each Orthogonal Frequency Division Multiplexing (OFDM)subcarrier in IEEE 802.11n links [18]. The CSI for each subcarrier can be represented as:

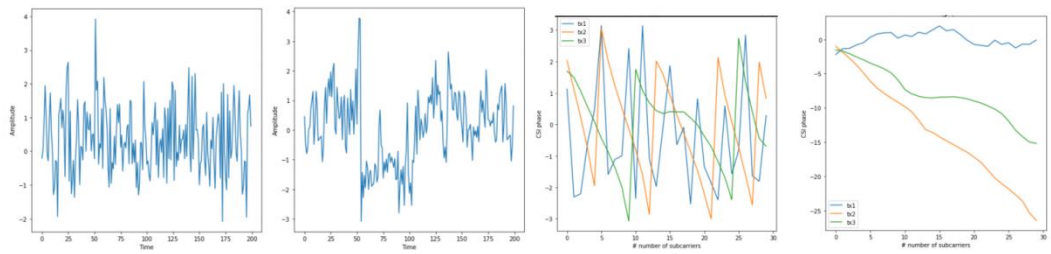
$$h = |h| e^{i\theta} \tag{2}$$

where |h| and Θ are the amplitude and phase respectively. The variations of the CSI amplitude and phase for each subcarrier in time represent unique CSI features.

3.3. Signal Preprocessing

The collected CSI stream from the Commercial-Off-The-Shelf (COTS) Wi-Fi devices is distorted [19-13] due to multi-path propagation in the environment, transmission power variations, the mismatch between the transmitter and receiver clocks, and frequencies and the errors of hardware and software. So, we need to reduce the CSI amplitude and phase noises by using a filter in the pre-processing signal phase. In our model, Weighted Moving Average (WMA) filter and phase sanitization are used to remove the randomness in CSI amplitude and phase, respectively.

1. Amplitude Noise Removal: WMA filter is applied to reduce the environmental variations in the CSI amplitude. Fig.(2a). and Fig.(2b). show the CSI amplitude for extracted subcarrier No.3 before and after the noise elimination by using WMA filter.



(a) Raw CSI amplitude of subcarrier 3 (b) CSI amplitude after Weighted moving average filter (c) Raw CSI Phase (d) Processed CSI Phase

Fig.2. Signal processing for our proposed framework

WMA uses equation (3) [20] to get the weighted average values.

$$A_t^{'} = \frac{1}{m+(m-1)+...+1}.\left[m.A_t + (m-1).A_{t-1} + ... + A_{t-m+1}\right] \quad (3)$$

where

A'_t	Wighted average amplitude at time t_i by its historical m values
m	Degree the current value is related to the historical records [2]

In our model, WMA filter has m=30.

2. Phase Sanitization: The CSI phase values are random due to the Carrier Frequency Offsets (CFO) and Sampling Frequency Offsets (SFO) between the transmitter and the receiver which cause random phase shifts. Therefore, we employ the phase sanitization method [21] to eliminate the random values and preserve the information reflecting sign details. The measured CSI phase Φ_i' from the i^{th} subcarrier can be expressed as in equation (4) [3]:

$$\phi_i^{'} = \phi_i - 2\pi \frac{k_i}{N}\delta + \beta + Z \quad (4)$$

where

Φ_t	True phase
k_i	Subcarrier index
N	Size of Fast Fourier Transform (FFT) which is 64 in IEEE 802.11n [22]
δ	Timing offset at the receiver due to SFO
β	Phase offset due to CFO
Z	Noise

As shown in Fig.(2c)., the raw CSI phases are folded within the range[-π,π]. From equation 4, due to the unknown values of δ and β , we can not obtain the real phase. Phase sanitization is also used for phase calibration, which used to remove the effects of SFO and CFO that represented in δ and β, respectively, by utilizing the linear fit after unwrapping the CSI phases to get the true phase values [20].

Fig.(2c). shows the raw CSI phase before applying the phase sanitization technique, where the processed CSI phase is shown in Fig.(2d).

3.4. Gesture recognition algorithm

Deep neural networks with convolution neural networks (CNNs) are employed to recognize the gestures in our feature extraction phase. CNN is used to extract a unique feature for corresponding sign word based on CSI amplitude and phase of this word.

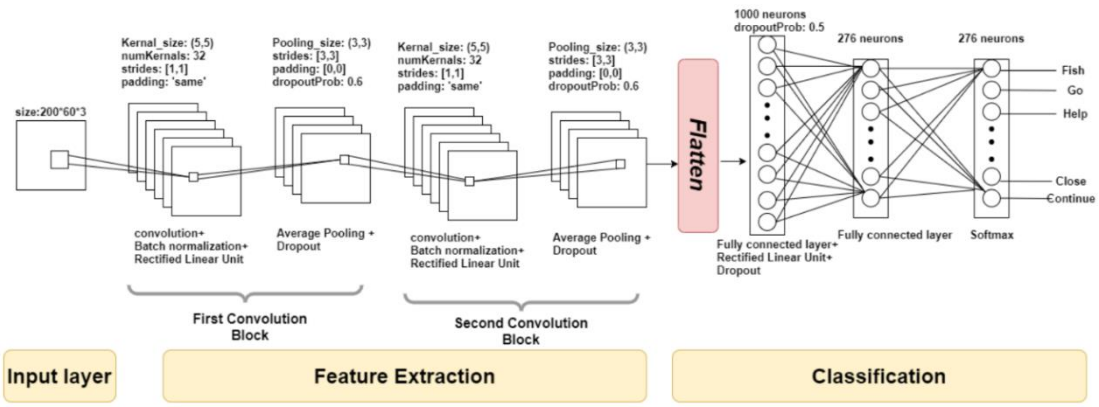


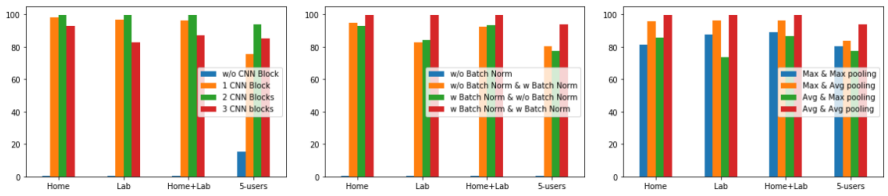
Fig.3. Overview of our proposed framework

The proposed CNN model is illustrated in Fig.3. It has three tiers: Input layer, Feature extraction, and finally, classification.

A. Input layer tier

In SignFi application, the size of each CSI matrix is size(csi) = (1,3,30) and there are 200 CSI samples for each sign gesture. So, the size of each CSI trace for each sign gesture is (3,30,200) [10]. In the input layer tier, the pre-

processed CSI amplitude and phase are concatenated and reshaped from (3,30,200) for each to single input (200,60,3).



(a) Number of CNN block (b) Batch Normalization Layer (c) Pooling Layer

Fig.4. Impact of batch normalization, pooling and number of CNN blocks on recognition accuracy using 5 flod cross validation.

B. Feature Extraction tier

This tier consists of two CNN blocks. Each of them has convolution layer, batch normalization layer, Rectified Linear Unit (ReLU) layer, average pooling, and dropout layer sequentially.

- Convolution Layer: It consists of 32 filters; each has 5 x 5 size and stride equals 1. Also, a border of zeros are added around input edges by making padding equals 1 to apply the convolution operation on all inputs. The output of this layer is a set of feature maps which are the input of the next layer.
- Batch Normalization Layer: The inputs are normalized to speed up the training process and improve the performance of the network by preventing the impact of overfitting.
- ReLU Layer: The Rectified Linear Unit (ReLU) layer is a nonlinear activation function. It sets any input value less than zero to zero. Its formula is as follows:

$$f(x) = \max(0, x) \tag{5}$$

- Average Pooling Layer: The role of pooling layer is reducing the number of connections to decrease the number of parameters to be learned in the network. In this framework, average pooling is applied with size 3 x 3 that returns the average of the input within this window and stride equals 3.
- Dropout Layer: The dropout layer is used to avoid overfitting in the training stage. It makes the overall network performance better. It randomly eliminates some of its inputs which have a certain probability.

A comprehensive study was done about: the batch normalization layer, pooling layer, and number of CNN block. It is presented in Fig.4.

Fig.4a. shows the impact of the used CNN block numbers. Our study is done with 4 scenarios (without CNN, 1 CNN, 2 CNN, and 3 CNN), as clearly seen in Fig.4a., two CNN block has the best accuracy value.

The study of batch normalization is shown in Fig.4b. For 2 CNN block, we study four cases, without batch normalization layer in both block, without batch normalization layer in the first block and with batch normalization layer in the second one, first and second block have with and without batch normalization layer, respectively, and finally, the both CNN has batch normalization layer. The results show that the use of batch normalization layer in both CNN block has the best accuracy value.

Fig.4c. shows the impact of pooling layer study by using the two most used methods (MAX and AVG pooling), the results reveal that the use of average pooling method for both CNN block achieves the highest accuracy value.

Our final conclusion is that the batch normalization layer has a major role in the feature extraction tier. But, the used method of pooling layer and the number of CNN block is changed based on the application of the framework.

C. Classification tier

- Fully-connected Layer: Two fully connected neural network layer are used. It plays an important role in the combination of all the learned features from the previous layers to share in the classification process. Two fully connected layers are used with size 1000 neurons and 276 neurons, respectively. The first one has ReLU activation function and a dropout rate equals 0.5. The second one is equal to the number of classes which equals 276.
- Softmax Classification Layer: It applies the softmax function to assign each CSI input to one of the 276 sign classes based on the predicted probabilities.

During the training process, Stochastic Gradient Descent with Momentum (SGDM) [23] optimization algorithm is used to update weights and biases during with 0.02 learning rate and 0.9 momentum.

4. Implementation and Evaluation

In this paper, gesture recognition based on Wi-Fi CSI is proposed. Deep learning is used in the proposed model. The proposed model is compared with the models introduced in [10,24,25,26] with respect to recognition accuracy. The experimental data was analyzed on Google Colab with pro version and 2 Tera. In Google Colab, we had a server with 26 GBytes Ram and P100 GPU. This model is implemented using Python 3.6 and Keras which is a Python deep learning library [27]. This section is divided into two parts: dataset description and result analysis.

4.1. Data Description

SignFi dataset [10] is used in training, testing, and validation process of the proposed model. It was collected by employing the access point(AP) with three external antennas as a transmitter and receiver equipped with Intel 5300 NIC with one internal antenna for gesture recognition. The dataset contains CSI traces obtained from two different environments, a laboratory, and a home scenario, with different room sizes, the distance between the access point (AP), receiver and transmitter's antenna orientations. The duration of each sign is between 0.5 seconds to 2.5 seconds. Table 1. summaries the used SignFi dataset for evaluating the proposed CNN model. The datasets can be downloaded from https://yongsen.github.io/SignFi/

Table 1. SignFi Dataset Summary

Dataset	Environment	Number of users	Number of sign	Number of repetitions for each sign	Number of instances
Home	Home	1	276	10	2760
Lab	Laboratory	1	276	20	5520
5-users	Laboratory	5	150	10 times per each user	7500

4.2. Results analysis

In this section, 5 fold cross validation is applied to ensure that an overfitting problem does not exist. We also split the data into 70% for training and 30% for testing to ensure the model can predict accurately new data. The results show that the classification performance degrades in the context of multi-human environment.

A. Proposed model accuracy evaluation:

5 fold cross validation for evaluation with the home dataset, lab dataset and 8,280 instances from home and lab environment together, as well as 5 users dataset, are used for system validation. Also, a self test runs which is 5 fold cross-validation for each user.

First, all data are randomly divided into 5 folds. Second, one fold is selected for testing and the rest for training, resulting in 5 runs. Finally, the average accuracy of all the 5 runs is computed. Accuracy is calculated based on equation (6) [3]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Where

TP	True Positive rate
TN	True Negative rate
FP	False Positive rate
FN	False Negative rate

We perform an accuracy comparison between the proposed model against some deep learning CSI based recognition approaches which can extract features automatically to examine the effectiveness of the proposed model.

The proposed model is compared with the models which were introduced in [10,24,25,26]. In [10] is presented a model with a 9-layer CNN deep learning model for gesture recognition after pre-processing raw CSI measurements. SignFi achieves 98.91%, 98.01%, 94.81%, 86.66%, 98% recognition accuracy in the home, lab, home+lab, 5-users and self test, respectively. In [24]], the authors proposed a general deep neural network (DNN) model that does not require any data pre-processing and uses raw CSI measurements as model input. The general DNN model achieves 99.89%, 99.98%, 99.65% recognition accuracy in home, lab, and the self test, respectively. The general DNN model does not study the recognition accuracy for lab and home dataset together and 5-user dataset.

Our proposed method has an average accuracy of 99.674%, 99.855%, 99.73%, 93.84%, 99% of the home, lab, home+lab, 5-users, and finally self test, respectively. As shown in Fig.5. As clearly seen in Fig.5., the proposed model accuracy is nearly equal to the accuracy of [24].

B. Performance evaluation

In this section, the evaluation of the proposed framework with respect to another deep learning technologies is introduced. There are two types of deep learning: LSTM and ABLSTM.

- Long Short Term Memory (LSTM): LSTM is the recurrent neural network architecture that overcomes the vanishing and exploding of the gradient problem. LSTM consists of a memory cell and three gates, an input gate, an output gate and a forget gate to remain important information with dependencies [25].
- Attention based Bidirectional Long Short-Term Memory (ABLSTM): It is a combination of Bidirectional Long Short-Term Memory (BLSTM) model and attention model. BLSTM processes sequential data in forward and backward directions because it has forward and backward Long Short Term Memory (LSTM) layers. So, informative features are automatically extracted. The role of attention model is that it assigns different weights to different features. More important features are given larger weights that make the overall performance better [26].

LSTM and ABLSTM work on time series data. As Wi-Fi CSI data are typical time series with temporal dependency, the LSTM and ABLSTM have achieved a remarkable performance for WiFI CSI based human activity recognition in [25] [26].

For comparison, We employ the schemes in LSTM model [25] and ABLSTM [26] for SignFi datasets. We use raw CSI amplitude and phase as the input feature vector to LSTM [25] that has one layer with 200 LSTM units. The LSTM method achieves an average accuracy of 74.06%, 49.82%, 76.92%, 56.9% and 63.5%, respectively for home, lab, home+lab, 5-users, and self test. We fed the raw CSI phase and amplitude into the attention based bidirectional long short-term memory model introduced in [26]. The ABLSTM achieves 95.44%, 96.2%, 94.94% , 73.83% ,70% recognition accuracy in the home, lab, home+lab , 5-users, and self-test, respectively. According to the result, the average recognition accuracy of our system is 99.674%, 99.855%, 99.73%, 93.84%, 99% of the home, lab, home+lab, 5-users, and self-test, respectively. The results are shown in Fig.5. Table 2. Summarizes the recognition accuracy of all approaches. It can be observed that the performance of all approaches in lab environment is higher than the home environment. This is due to the complexity of home environment. The proposed model outperforms other approaches due to automatic feature extractions as they are suitable to accurately recognize different sign gestures in complex environments.

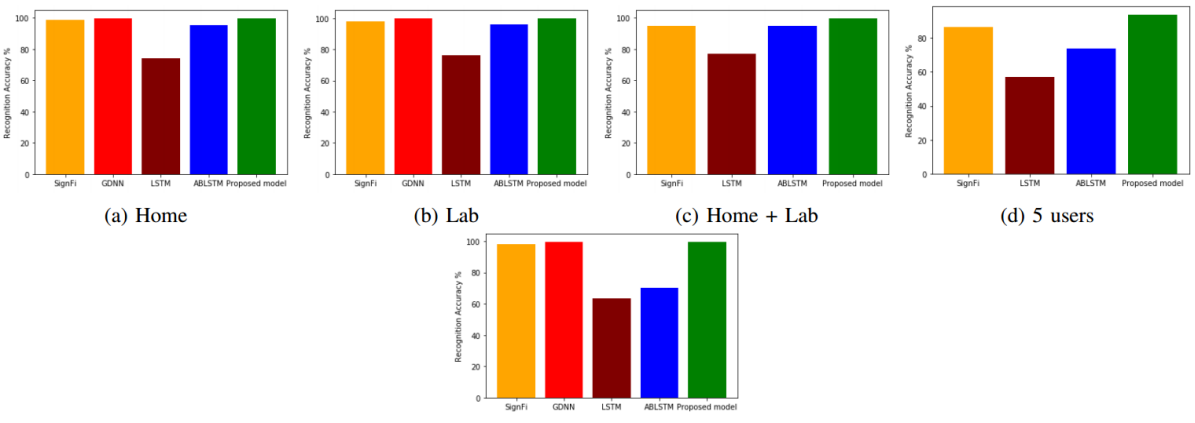


Fig.5. The Recognition Accuracies of All Approaches for Different Datasets

Table 2. The Recognition Accuracies of the Proposed Framework Versus the Other Approaches

Method	Home	Lab	Home+Lab	5-users	Self Test
SignFi [10]	98.91%	98.01%	94.81%	86.66%	98%
General DNN [24]	99.89%	99.98%	-	-	99.65%
LSTM [25]	74.06%	76.33%	76.92%	56.9%	63.5%
ABLSTM [26]	95.44%	96.2%	94.94%	73.83%	70%
Proposed Method	99.674%	99.855%	99.73%	93.84%	99%

The performance evaluation is presented by using time consumption in training and testing processes, F-score, precision, sensitivity, the impact of pre-processing data, and some other features.

Time consumption of the proposed frameworks and the main SignFi system is listed in Table 3. SignFi takes 8.28ms to finish training and 0.62ms for testing. While LSTM model takes 7.2ms and 3ms for training and testing,

respectively. ABLSTM takes 27.2ms for training which can be considered as a long training time among all approaches and 10ms for testing. The training time of the proposed CNN model is 1ms which is the shortest training time among all approaches while the testing time is 0.66ms which is nearly equal SignFi training time. With the above results, it can be concluded that the proposed model can be used for Wi-Fi CSI based sign gesture recognition in real-time since it achieves acceptable accuracy and its test time is small.

Table 3. The Training and Testing Time of All the Approaches

Time	SignFi [10]	General DNN [24]	LSTM [25]	ABLSTM [26]	Proposed Method
Training (ms)	8.28	-	7.2	27.2	1
Testing (ms)	0.62	-	3	10	0.66

The proposed frameworks are evaluated based on the most performance metrics used Accuracy, F1-score, Precision, and Recall. These metrics expressed as shown in equation (7), (8), and (9) [3]

$$F1-score = \frac{2TP}{2TP + FP + FN} \tag{7}$$

$$\Pr ecision = \frac{TP}{TP + FP} \tag{8}$$

$$\mathrm{Re}call(Sensitivity) = \frac{TP}{TP + FN} \tag{9}$$

Table 4., Table 5. and Table 6. show the results for f1-score, accuracy, precision, and recall for LSTM, ABLSTM, and CNN frameworks, respectively in all standard environment. Fig.6. presents the recognition accuracy of the proposed CNN model for each environment by using 70% of data for training and 30% of data for testing.

Table 4. Performance of Lstm Model

Method	**Accuracy%**	**F1-score%**	**Precision%**	**Recall%**
Home	89.61	89.3	91.1	89.6
Lab	94.26	94.2	94.6	94.3
Home+Lab	89.05	88.8	90.2	89.00
5-users	65.02	65.3	68.7	65.00

Table 5. Performance of Ablstm Model

Method	**Accuracy%**	**F1-score%**	**Precision%**	**Recall%**
Home	97.22	97.2	97.6	97.2
Lab	97.16	97.00	97.2	97.2
Home+Lab	94.48	94.3	95.2	94.5
5-users	71.11	71.6	74.6	71.1

Table 6. Performance of the Proposed Cnn Model

Method	**Accuracy%**	**F1-score%**	**Precision%**	**Recall%**
Home	99.52	99.5	99.5	99.5
Lab	99.8	99.8	99.8	99.8
Home+Lab	99.5	99.5	99.5	99.5
5-users	91.8	91.8	91.8	91.8

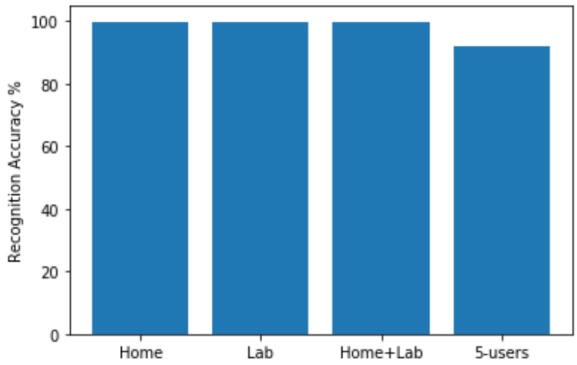


Fig.6. Leave-one-subject-out for proposed model

The impact of the data pre-processing phase in CNN framework is also studied. The accuracy of the framework with and without the pre-processing phase is shown in Fig.7. By using raw CSI amplitude and phase information as features. The recognition accuracy drops from 99.674% to 94.78% in the home dataset. In the lab dataset, the recognition accuracy drops from about 99.855% to 94.891%. Moreover, the recognition accuracy achieves 92.96% and 78.71% in home + lab, and 5-users datasets. So, the results confirm that amplitude noise removal and phase sanitization play a vital role in improving the performance of our proposed model.

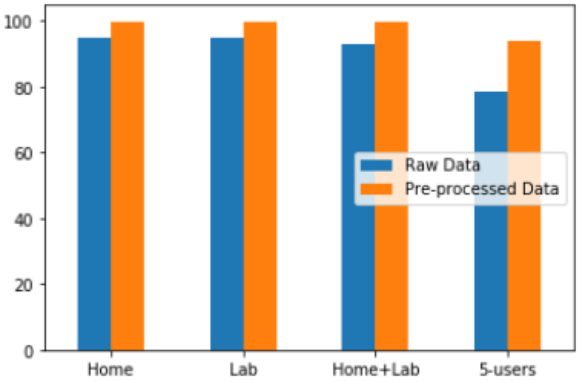


Fig.7. Impact of preprocessing

The impact of CSI phase and amplitude on the performance of the proposed CNN model is also investigated. Fig.8. presents the recognition accuracies of the proposed CNN model with both CSI phase and amplitude values, with CSI phase values, and with CSI amplitude values in different environments. The highest recognition accuracy is obtained when we use CSI amplitude information only or CSI phase information only when the dataset is performed by a single user as in home, lab, and both home and lab datasets. Moreover, the performance of the model drops if we use the CSI phase or amplitude as features to 76.067% and 85.733%, respectively in the dataset collected from more than one user as in the 5-user dataset. So, to improve the recognition accuracy, we recommended the use of both CSI phase and amplitude information to extract a unique pattern of features for each sign word independent on the user that performs it.

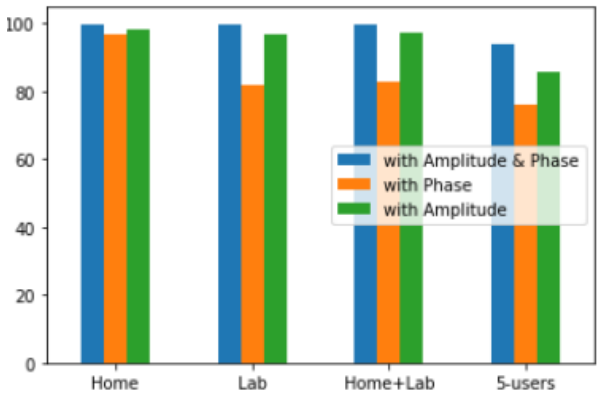


Fig.8. Impact different features

The Impact of using different optimizers with the proposed CNN model on the performance metrics is also studied. The used optimizers are Adam [28] with 0.001 learning rate, AdaGrad [29] with 0.01 learning rate, AdaDelta [29] with 1.0 learning rate, AdaMax [29] with 0.002 learning rate, and Stochastic Gradient Descent with Momentum (SGDM) [23] with 0.02 learning rate and 0.9 momentum. Table 7., Table 8., Table 9., Table 10. and Table 11. show the effect of adam, AdAdaGrad, AdaDelta, AdaMax and SGDM. SGDM optimizer outperforms other optimizers. It achieves the highest results of Lab, Home+Lab, and 5 Users datasets. But, it is nearly equal to the results of AdaMax optimizer of Home environment.

Table 7. Performance of the Proposed Cnnmodel using Adam Optimizer

Method	Accuracy%	F1-score%	Precision%	Recall%
Home	98.91	99.5	99.5	99.5
Lab	97.94	97.9	98.3	97.9
Home+Lab	98.31	98.3	98.6	98.3
5-users	81.42	81.8	84.1	81.4

Table 8. Performance of the Proposed Cnn Model using Adagrad Optimize

Method	Accuracy%	F1-score%	Precision%	Recall%
Home	99.27	99.3	99.5	99.3
Lab	97.52	97.5	98.00	97.5
Home+Lab	98.19	98.2	98.5	98.5
5-users	79.42	79.8	82.7	79.4

Table 9. Performance of the Proposed Cnn Model using Adadelta Optimizer

Method	Accuracy%	F1-score%	Precision%	Recall%
Home	97.95	97.8	98.6	97.9
Lab	98.85	98.8	99.00	98.9
Home+Lab	98.00	98.00	98.3	98.00
5-users	81.2	83.1	81.00	81.02

Table 10. Performance of the Proposed Cnn Model Using Adamax Optimizer

Method	Accuracy%	F1-score%	Precision%	Recall%
Home	99.64	99.6	99.7	99.6
Lab	97.77	97.7	98.1	97.8
Home+Lab	97.46	97.4	97.8	97.5
5-users	81.11	81.3	83.3	81.1

5. Conclusions and Future Works

In this paper, a framework for a Wi-Fi CSI based gesture recognition system based on a Convolution Neural Network (CNN) as a classification algorithm is developed. Weighted moving average filter and phase sanitization are used for amplitude noise removal and removing phase offsets from raw CSI measurement, respectively to use them as base signals for the CNN classification algorithm. CNN is able to learn automatic features that represent the relationship between Wi-Fi CSI variance and sign gestures. The effectiveness of our proposed model is verified by comparing the recognition accuracy of our proposed model with other deep learning approaches. The proposed framework exceeds the other models. It achieves99.674%, 99.855% and 99.73% recognition accuracy at home, lab, and home+lab environment for 276 sign gestures collected by a single user, respectively. Moreover, the proposed model achieves 93.84% recognition accuracy for 7,500 samples performed by 5 users in a lab environment which means that our model is robust as it achieves acceptable accuracy in the context of multi-human environments The use of transfer learning techniques such as ResNet is considered as a future work.

References

[1] H. Abdelnasser, M. Youssef, and K. A. Harras,“Wigest A ubiquitous wifi-based gesture recognition system,” in 2015 IEEE Conference on Computer Communications (INFOCOM). IEEE, 2015, pp. 1472–1480.

[2] Y. Wang, K. Wu, and L. M. Ni, “Wifall: Device-free fall detection by wireless networks,” IEEE Transactions on Mobile Computing, vol. 16, no. 2, pp. 581–594, 2016.

[3] T. Z. Chowdhury, “Using wi-fi channel state information (csi) for human activity recognition and fall detection,” Ph.D. dissertation, University of British Columbia, 2018.

[4] S. Oprisescu, C. Rasche, and B. Su, "Automatic static hand gesture recognition using tof cameras," in 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO). IEEE, 2012, pp.2748–2751.
[5] Z. Zhang, "Microsoft kinect sensor and its effect," IEEE multimedia,vol. 19, no. 2, pp. 4–10, 201
[6] U. of Washington. (2016) Signaloud demo. [Online]. Available: http://www.washington.edu/news/2016/04/12/uw-undergraduate-teamwins-10000-lemelson-mit-student-prize-for-gloves-that-translate-signlanguage/
[7] M. Scholz, S. Sigg, H. R. Schmidtke, and M. Beigl, "Challenges for device-free radio-based activity recognition," in Proceedings of the 3rd workshop on Context Systems, Design, Evaluation and Optimisation (CoSDEO 2011), in Conjunction with MobiQuitous, vol. 2011, 2011.
[8] M. Al-qaness and F. Li, "Wiger: Wifi-based gesture recognition system," ISPRS International Journal of Geo-Information, vol. 5, no. 6, p. 92, 2016.
[9] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, "Whole-home gesture recognition using wireless signals," in Proceedings of the 19th annual international conference on Mobile computing & networking. ACM, 2013, pp. 27–38.
[10] Y. Ma, G. Zhou, S. Wang, H. Zhao, and W. Jung, "Signfi: Sign language recognition using wifi," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 2, no. 1, p. 23, 2018.
[11] J. Wang, D. Vasisht, and D. Katabi, "Rf-idraw: virtual touch screen in the air using rf signals," ACM SIGCOMM Computer Communication Review, vol. 44, no. 4, pp. 235–246, 2015.
[12] S. Sigg, S. Shi, F. Buesching, Y. Ji, and L. Wolf, "Leveraging rf-channel fluctuation for activity recognition: Active and passive systems, continuous and rssi-based signal features," in Proceedings of International Conference on Advances in Mobile Computing & Multimedia, 2013, pp. 43–52.
[13] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Understanding and modeling of wifi signal based human activity recognition," in Proceedings of the 21st annual international conference on mobile computing and networking. ACM, 2015, pp. 65–76.
[14] W. He, K. Wu, Y. Zou, and Z. Ming, "Wig: Wifi-based gesture recognition system," in 2015 24th International Conference on Computer Communication and Networks (ICCCN). IEEE, 2015, pp. 1–7.
[15] H. Li, W. Yang, J. Wang, Y. Xu, and L. Huang, "Wifinger: talk to your smart devices with finger-grained gesture," in Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, 2016, pp. 250–261.
[16] Z. Tian, J. Wang, X. Yang, and M. Zhou, "Wicatch: a wi-fi based hand gesture recognition system," IEEE Access, vol. 6, pp. 16 911–16 923, 2018.
[17] C. Wang, S. Chen, Y. Yang, F. Hu, F. Liu, and J. Wu, "Literature review on wireless sensing-wi-fi signal-based recognition of human activities," Tsinghua Science and Technology, vol. 23, no. 2, pp. 203–222, 2018.
[18] C. Feng, S. Arshad, and Y. Liu, "Mais: Multiple activity identification system using channel state information of wifi signals," in International Conference on Wireless Algorithms, Systems, and Applications. Springer, 2017, pp. 419–432.
[19] Y. Xie, Z. Li, and M. Li, "Precise power delay profiling with commodity wi-fi," IEEE Transactions on Mobile Computing, vol. 18, no. 6, pp. 1342–1355, 2018.
[20] S. Liu, Y. Zhao, F. Xue, B. Chen, and X. Chen, "Deepcount: Crowd counting with wifi via deep learning," arXiv preprint arXiv:1903.05316, 2019.
[21] C. Wu, Z. Yang, Z. Zhou, K. Qian, Y. Liu, and M. Liu, "Phaseu: Realtime los identification with wifi," in 2015 IEEE conference on computer communications (INFOCOM). IEEE, 2015, pp. 2038–2046.
[22] C. Gao, Y. Li, and X. Zhang, "Livetag: Sensing human object interaction through passive chipless wifi tags," in 15th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 18), 2018, pp. 533–546.
[23] L. Bottou, "Stochastic gradient descent tricks," in Neural networks: Tricks of the trade. Springer, 2012, pp. 421–436.
[24] B. Wei, K. Li, C. Luo, W. Xu, and J. Zhang, "No need of data preprocessing: A general framework for radio-based device-free context awareness," arXiv preprint arXiv:1908.03398, 2019.
[25] S. Yousefi, H. Narui, S. Dayal, S. Ermon, and S. Valaee, "A survey on behavior recognition using wifi channel state information," IEEE Communications Magazine, vol. 55, no. 10, pp. 98–104, 2017.
[26] Z. Chen, L. Zhang, C. Jiang, Z. Cao, and W. Cui, "Wifi csi based passive human activity recognition using attention based blstm," IEEE Transactions on Mobile Computing, 2018.
[27] A. Gulli and S. Pal, Deep learning with Keras. Packt Publishing Ltd, 2017.
[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
[29] M. D. Zeiler, "Adadelta: an adaptive learning rate method," arXiv preprint arXiv:1212.5701, 2012.

Authors' Profiles

Marwa R. Bastwesy was born in 1994, Egypt.

She received the B.S. degree from Tanta university, Egypt, in June 2016. She is a demonstrator at the Department of Computers and Automatic Control Engineering, Faculty of Engineering, Tanta University, Egypt.

Her research interests are in artificial intelligence, object detection and recognition, deep learning, device-free sensing, wireless networks.

Nada M. Elshennawy was born in 1978, Egypt.

She received the PhD in WiMAX wireless networks from Tanta University, Egypt. She is an assistant professor at the Department of Computers and Automatic Control Engineering, Faculty of Engineering, Tanta University, Egypt.

Her research interests are in machine learning, computer vision and human behavior recognition, wireless networks, wireless sensor networks, neural networks.

Mohamed T F Saidahmed received the B.S. degree from Menofyia university (MU), Shebin El-Kom (SK), in June 1969, M.Sc. degree from Helwan university(UV) in May 1976, both in Egypt, and the D.Sc. Degree in System Science, Control and Networks from The George Washington University (GW), the School of Engineering and Applied Science (SEAS), Washington D.C., May 1983, USA. From1969 - 1977, he worked Demonstrator and Teaching Assistant in the Electrical Engineering (EE) dept., Faculty of Eng.(FE),MU, SK, Egypt. Toward Working for the D.Sc. Degree in USA, he worked Teaching Assistant from Fall 1979 to Spring 1983 at SEAS,GW,USA. From 1977 – Sep.1997, he has position of Assistant, Associate, and full Professor in the EE dept., FE, MU, SK, Egypt.

Since 1st October 1997, he has been full Professor in the Department of Computer Engineering and Automatic Control (CEAT) in FE,Tanta University (TU), Egypt, till now. From 1998 to 31st of July 2006, he was the Vice Dean and Dean of the FE , TU, Egypt. Since 2008, he spent his sabbatical leave in the EE Dept,Taif University,Taif,KSA. His research interests include singular systems, deep learning, AI, computer engineering, nonlinear and estimation of robust delayed singular control system.

How to cite this paper: Marwa R. M. Bastwesy, Nada M. ElShennawy, Mohamed T. Faheem Saidahmed, " Deep Learning Sign Language Recognition System Based on Wi-Fi CSI", International Journal of Intelligent Systems and Applications(IJISA), Vol.12, No.6, pp.33-45, 2020. DOI: 10.5815/ijisa.2020.06.03