IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 June 2024, accepted 8 July 2024, date of publication 15 July 2024, date of current version 18 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3428550

RESEARCH ARTICLE

WiFi-TCN: Temporal Convolution for Human Interaction Recognition Based on WiFi Signal

CHIH-YANG LIN [1], (Senior Member, IEEE), CHIA-YU LIN [2], YU-TSO LIU [2], YI-WEI CHEN [2], AND TIMOTHY K. SHIH [2], (Senior Member, IEEE)

[1]Department of Mechanical Engineering, National Central University, Taoyuan 32001, Taiwan
[2]Department of Computer Science and Information Engineering, National Central University, Taoyuan 32001, Taiwan

Corresponding author: Timothy K. Shih (tshih@g.ncu.edu.tw)

This work was supported by the National Science and Technology Council under Grant 110-2221-E-008-104-MY3, Grant 111-2221-E-008-110-MY3, and Grant 112-2918-I-008-008.

ABSTRACT The quest for efficient and non-intrusive human activity recognition (HAR) in indoor environments has led to the burgeoning field of WiFi-based HAR. This method promises applications from healthcare monitoring to elderly care due to its cost-effectiveness and ease of deployment compared to traditional sensor-based systems. However, WiFi-based HAR faces significant challenges in maintaining performance across diverse environments and subjects, largely due to WiFi signal variability. Addressing this issue necessitates training models on extensive datasets. Recent studies have utilized conventional models, including Convolutional Neural Networks (CNNs), and sequence-to-sequence (Seq2Seq) models such as LSTM, GRU, or Transformer. Despite their precision, these models are computationally intensive and require more training data. To tackle these limitations, we propose a novel approach that leverages a Temporal Convolutional Network with Augmentations and Attention, referred to as TCN-AA. This model enhances computational efficiency and accuracy in the face of dataset variability by combining temporal convolutional layers with data augmentation strategies and an attention mechanism to focus on critical features. Our method is computationally efficient and significantly improves accuracy, even with a threefold increase in data size through augmentation techniques. Experiments on a public dataset indicate our approach outperforms state-of-the-art methods, achieving 99.42% accuracy. This work represents a significant step forward in the practical application of WiFi-based HAR, paving the way for broader adoption in critical areas like healthcare and security.

INDEX TERMS Enter attention, channel state information (CSI), data augmentation, human activity recognition, temporal convolution network (TCN), WiFi signal.

I. INTRODUCTION

The recognition of HAR in an indoor environment using WiFi applications has emerged as a significant and widely researched field [1], [2]. Literature review [3], [4], [5], [6], [7] indicates that camera-based and sensor-based HAR systems are widely utilized in addressing this issue. However, these approaches often come with concerns regarding non-intrusiveness and high privacy. Specifically, camera-based systems require not only the initial hardware but also substantial storage for video data, alongside concerns about privacy violations in sensitive environments. Sensor-based solutions, particularly those relying on wearable technology, necessitate a significant investment in multiple units for comprehensive coverage and face challenges regarding user compliance due to the discomfort and inconvenience of wearing devices for extended periods [8]. In contrast, WiFi-based HAR systems leverage existing network infrastructure, significantly reducing the cost and intrusiveness of deploying a dedicated monitoring system. The effectiveness of WiFi as an HAR solution is further accentuated by its non-intrusive nature and the ubiquity of WiFi infrastructure in modern indoor environments. This makes it a promising alternative solution for HAR [2], [3], [9], [10], capable of overcoming the limitations associated with traditional sensor and camera-based approaches. Despite its potential, WiFi-based HAR is

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

not without its challenges, including limited coverage and complexities in multi-subject scenarios.

In the realm of HAR, there are three distinct types of WiFi signals that are employed, namely hardware-based radio signals, Received Signal Strength Indicator (RSSI), and Channel State Information (CSI). The utilization of RSSI signals has garnered widespread attention in a multitude of sensing applications, including indoor positioning and tracking [11], [12]. However, due to the inherently coarse nature of RSSI signals, it poses a challenge to attain high levels of precision in fine-grained HAR. For instance, certain interactions as described in [9] that entail fine-grained activities such as ''kicking with the left leg'', ''kicking with the right leg'', ''pointing with the left hand'', and ''pointing with the right hand'' are particularly difficult to accurately determine due to the limitations posed by RSSI signals.

The CSI provides detailed information at the physical layer that is invaluable for precise activity recognition. The utilization of techniques such as Multi-Input Multiple-Output (MIMO) and Orthogonal Frequency Division Multiplexing (OFDM) [13] allows for the acquisition of additional phase and amplitude information of each sub-carrier from the WiFi signal that is spread between the transmitting and receiving antennas and operates at a specific carrier frequency. As a result of limb movement, variations in wireless signals can be observed. Early research has established the superiority of CSI compared to the traditional RSSI [14]. Despite the ability of CSI to capture the aggregate impact of multipath interference caused by the human body, variations in the signals may still persist, even when the same subjects are engaged in the same activity. To mitigate this issue, it is crucial to gather a substantial amount of data to prevent the model from unduly focusing on differences between subjects rather than the activity itself.

Addressing these gaps, our research introduces a novel Temporal Convolution Network with Augmentations and Attention (TCN-AA) approach. Our model innovatively combines the strengths of temporal convolution networks (TCN) with advanced data augmentation techniques and an attention mechanism, setting a new standard in WiFi-based HAR. By doing so, TCN-AA not only significantly improves the accuracy and efficiency of activity recognition over existing models but also demonstrates remarkable adaptability in handling the inherent signal variability among different subjects.

To this end, some researchers have utilized data augmentation to increase the diversity of the dataset [15], [16], though the outcomes remain unsatisfactory.

Our main contributions are as follows:

1) We propose a novel model known as Temporal Convolution Network with Augmentations and Attention (TCN-AA). Our approach diverges from conventional CNNs and Seq2Seq models such as Long-Short Term Memory (LSTM) [17] and Gated Recurrent Unit (GRU), instead relying on a TCN. By utilizing this architecture, we are able to reap the benefits of convolution networks, such as the capacity to process sequential inputs, as well as significantly reduced training duration.
2) We incorporate data augmentation techniques on WiFi signals to expand the available data, and implement an attention mechanism to enhance the model's focus on the activity and expedite convergence.
3) The empirical results of our proposed experiments demonstrate that TCN-AA outperforms existing state-of-the-art methods and achieves the highest accuracy.

The remainder of this paper is organized as follows. Section II presents a comprehensive overview of the previous studies on WiFi-based HAR. In Section III, We provide a detailed description of our environment setup and system design, including our proposed model, TCN-AA, and an attention mechanism to enhance the model's focus on the activity and expedite convergence. In Section IV, conduct a comprehensive evaluation of our model and method, including the TCN-AA model, attention mechanism of expedite convergence, and compare it to existing approaches. Finally, the conclusion and potential directions for future research are presented in Section V.

II. RELATED WORK

A comprehensive overview of recent studies pertaining to WiFi-based HAR will be presented. The available approaches for WiFi-based HAR can be mainly divided into two subgroups, based on their methodologies: CNNs, and Seq2Seq models.

A. CNN-BASED APPROACHES

The majority of early studies on WiFi-based HAR employ CNNs based models as a starting point. This is due to the characteristic pattern of activity present in the CSI signals, which make CNNs an appropriate choice for recognition. For instance, Wang et al. [18] utilized a modified U-Net of deep convolutional neural network and annotations on 2D images to perform body segmentation and pose estimation from CSI signals. Their model was evaluated on more than 105 frames across 16 indoor scenes and achieved scores of 0.91 for AP@50, 0.65 for mIoU, and an average mAP score of 0.38 over AP@50-AP@95. While their performance is not optimal, their study demonstrated the potential to reconstruct fine-grained 2D spatial information of human bodies from CSI signals.

Kabir et al. [19] developed the CSI-based Inception Attention Network (CSI-IANet) incorporating CNNs and spatial-attention and evaluated it using a dataset of WiFi-based human-to-human interactions (HHI), which is the same dataset used in this paper [9]. The HHI dataset includes 12 different human-to-human interactions performed by two subjects and will be described in detail in section III. The CSI-IANet achieved an average accuracy of 91.3%, making it the first CNN-based model to surpass 90% accuracy on this dataset. The authors applied a Butterworth low-pass filter to denoise the CSI signal, employed an inception module to provide the model with varying receptive fields, and utilized spatial attention to achieve such high accuracy.

Subsequently, Alazrai et al. [1] proposed the HHI-AttentionNet, which employed a depth-wise CNN and a customized attention mechanism, achieving 95.47% accuracy on the same dataset [9]. The authors applied a Butterworth low-pass filter to remove significant peaks, such as outliers and high-frequency noise, and a Gaussian smoothing function to remove short peaks. Both CSI-IANet and HHI-AttentionNet utilized segmentation to split the data into smaller windows. However, this approach can compromise the integrity of the data as it requires fixing the length of each sample, and the computing power required increases as the length of each sample decreases. In contrast, this paper fixes the length of each sample by chopping some parts of the steady state and applying a one-dimensional Discrete Wavelet Transform (DWT) to downsample the data, which will be described in detail in Section III. Despite the high accuracy of 95.47% achieved by HHI-AttentionNet, there is still room for improvement. Among the studies utilizing the same dataset [9], the 95.47% accuracy is the highest achieved by a CNN-based model and it appears to be the limit for CNN-based models.

B. SEQ2SEQ MODEL APPROACHES

The limitations of CNN models have led researchers to investigate alternative approaches for processing time-sequence data. One such approach is the use of Seq2Seq models. In recent years, there has been an increasing trend in the use of Seq2Seq models for WiFi-based HAR.

An attention-based bi-directional long short-term memory (ABLSTM) network approach has been proposed in [20] for the implementation of a passive human activity recognition system. The system utilizes CSI obtained from WiFi signals to recognize six common daily activities including falling, walking, running, standing up, sitting down, and lying down. To achieve this, the authors first employ a bi-directional LSTM to encode the CSI signal, and then use attention mechanism on the LSTM output to make the final prediction. The proposed approach demonstrates remarkable performance with an average accuracy of 96.7% and 97.3% on two different environments, namely the Activity Room and the Meeting Room. However, it is worth noting that the high accuracy achieved in this study is based on data collected from a single individual performing coarse-grained activities, which may have made the recognition task easier for the model.

A different Seq2Seq approach is Two-Stream Convolution Augmented Human Activity Transformer (THAT) [21] that utilizes CSI signals to recognize seven daily activities [20], including an additional activity, "picking up." The authors propose a network architecture that combines the strengths of convolutional neural networks and multi-head self-attention transformers to achieve high recognition accuracy. The architecture utilizes two distinct input streams, a temporal stream and a channel stream, which are fed into separate CNN-based transformers. This design allows the model to generate distinct features through convolution in the temporal and channel domains, ultimately leading to improved performance. The results of the study show that the use of two input streams significantly improves recognition accuracy, with an average accuracy of 98.6% across four different scenes. Although this approach is faster than state-of-the-art Seq2Seq-based methods, it still requires a substantial amount of time for training and only uses data from a single subject.

Another example of a device-free deep learning model, H2HI-Net [22], presents an illustration of a multimodal approach to tackle the human-to-human interaction problem. The authors propose a combined model that encompasses a residual neural network and a bi-directional gated recurrent unit (Bi-GRU), which is evaluated using the HHI dataset [9]. Preprocessing of the CSI signal involves denoising through the application of the Butterworth filter and dimensionality reduction via principal component analysis. The residual neural network and Bi-GRU are then utilized to encode the spatial and temporal embeddings, respectively. The interactions are finally predicted through the concatenation of these two embeddings with a two-layer Dense Net, resulting in an average accuracy of 96.39%. While this performance is satisfactory, the authors still employ the Seq2Seq approach, which results in increased computational demands. Conversely, this paper utilizes a TCN with an attention mechanism and data augmentation, resulting in improved accuracy exceeding 96.39% while ensuring low computational requirements.

III. MATH

This section provides an overview of the dataset, the proposed methods, and the corresponding model. Initially, a description of the dataset is presented, which encompasses details on the interactions during data collection, the environment, and the environmental settings. Subsequently, the preprocessing and augmentation techniques employed in the proposed methods are provided. Finally, the section concludes with a discussion of the TCN model and the hyperparameters that resulted in optimal performance. The methodology of the proposed approach is presented in a flowchart in Fig. 1.

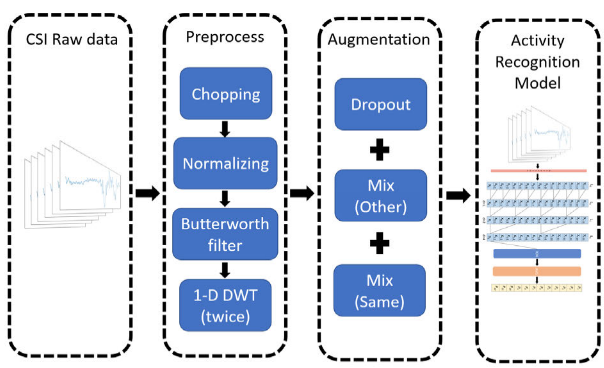

FIGURE 1. The flowchart of our methods.

A. CSI DATASET OF HUMAN-TO-HUMAN INTERACTION

In this paper, the proposed method is validated using the publicly available CSI dataset [9]. This dataset comprises 64 subjects consisting of 40 distinct pairs, who were required to perform a total of 12 interaction activities (i.e., Nc = 12), including approaching, departing, handshaking, high five, hugging, kicking with the left leg, kicking with the right leg, pointing with the left hand, pointing with the right hand, punching with the left hand, punching with the right hand, and pushing. The subjects performed 10 trials for each activity, with each trial lasting between 5 and 6 seconds, including a 2-second steady state. The packet length (N_p) varied between [1040, 2249].

Data collection was performed using a commercial off-the-shelf access point, such as the Sagemcom 2704 WiFi router, as the transmitter and a desktop computer equipped with an Intel 5300 NIC as the receiver. The MIMO WiFi streams consisted of 2 internal antennas at the transmitter (Nt = 2) and 3 external antennas at the NIC (Nr = 3). Data was collected using a bandwidth of 20 MHz and a frequency of 2.4 GHz, and the CSI was captured using a publicly available CSI tool [23] with 30 subcarriers (i.e., Ns = 30).

The Intel 5300 NIC chipset utilizes the 802.11n CSI to present channel matrices, each having a signed 8-bit resolution for the real and imaginary parts. In essence, each transmitter-receiver pair derived from the CSI signals can be represented as $H(T_i, R_j)$, where h_y^x denotes the complex number corresponding to the x^{th} packet and the y^{th} subcarrier, with x $\in [1, \ldots, N_p]$ and y $\in [1, \ldots, N_s]$. The notation (T_i, R_j) indicates that the signal is transmitted from the i^{th} antenna at the transmitter and received by the j^{th} antenna at the receiver. The CSI signals can be conceptualized as a vector as:

$$H\left(T_i, R_j\right) = \begin{bmatrix} h_1^1 & \cdots & h_N^1 \\ & \ddots & \\ \vdots & h_y^x & \vdots \\ & & \ddots \\ h_1^{N_p} & \cdots & h_{N_s}^{N_p} \end{bmatrix}$$

B. PREPROCESSING

The task of directly recognizing CSI signals from their values proves to be a challenge due to the inherent noise present in CSI signals. Fluctuations in CSI values are a common occurrence, even in environments without any human presence, due to the presence of internal and external sources of interference. The internal interferences include transmission rate adaptations and variations in power, while external interferences are comprised of background radio noise and interference caused by moving objects. As such, it is imperative that the CSI values undergo preprocessing prior to further analysis.

During the data collection process, the transmitter transmits data to the receiver at a specified frequency, which can result in fluctuations in the actual sampling rate due to internal and external sources of interference. At times, the overhead demands of the device may be excessive, leading to a significant drop in the sampling rate for a short duration. Data collected in this manner is often considered flawed due to a lower number of total packets in comparison to other sets of data. Our dataset encompasses packets of data ranging in size from 1040 to 2249. In order to discard data with fewer than 1500 packets, we have established a threshold of 1500 (i.e., N_p = 1500). For data with a packet count exceeding 1500, we have implemented a process of chopping the steady state in order to standardize the packet count at 1500. The final input data is represented in the form of $(N_t \times N_r, N_p, N_s)$. Despite the consistency in the activity being performed by the same subject-pair, variations in the CSI value can occur due to external interference. To mitigate this issue, normalization is imperative. The normalization process involves transforming the data to a range of [−1,1] for each Transmitter-Receiver pair. Subsequently, a Butterworth filter, a low-pass filter that effectively eliminates most high-frequency noise, is applied to each TR-pair. Finally, to achieve down sampling, a one-dimensional DWT is applied twice, reducing N_p from 1500 to 375. In the process of utilizing one-dimensional Discrete Wavelet Transform (1D-DWT), the original signal undergoes decomposition into two sets of coefficients: the approximation coefficients and the detailed coefficients. The approximation coefficients depict the low-frequency information, while the detailed coefficients correspond to high-frequency information. Given that the most significant interactions occur in the low-frequency domain, we choose to retain the approximation coefficients and discard the detailed coefficients.

The efficacy of our methodology is illustrated in Fig. 2, where we present the amplitude data of three separate interactions. The application of a Butterworth filter led to a substantial reduction in high-frequency noise. Furthermore, repeating the 1D-DWT process resulted in unchanged feature patterns and a decrease in both data length and training time. This is of significant value for researchers without access to advanced GPUs. Additionally, 1D-DWT plays a crucial role in removing high-frequency noise, as demonstrated in our experimental section.

C. AUGMENTATION

In contrast to the abundance of datasets available for image recognition, the availability of datasets pertaining to WiFi is relatively scarce. This disparity is demonstrated by the fact that datasets for image recognition, such as ImageNet and COCO [24], contain an order of magnitude greater number of samples compared to WiFi-based datasets. Thus, the implementation of data augmentation techniques for WiFi-based datasets is of utmost importance. Our research endeavors addressed this issue by employing and comparing the efficacy of three distinct augmentation techniques. The methods employed are presented as follows:

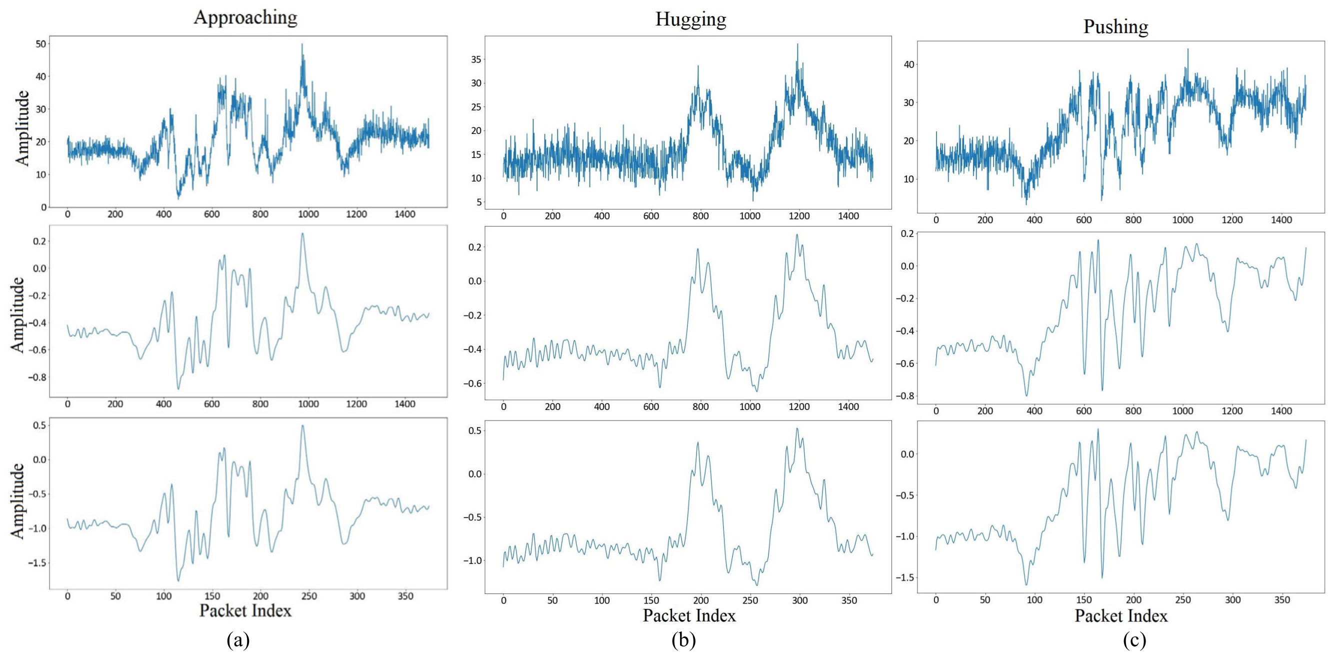

FIGURE 2. **The flowchart of our methods. (a) is the amplitude data of interaction "Approaching", (b) is the amplitude data of "Hugging", and (c) is the amplitude data of "Pushing". For the figures in first row, they represent the raw data. For the figures in middle row, they represent the data after a Butterworth lowpass filter and normalization. For the figures in third row, they represent the data after applying 1D DWT twice. As you can see, the length of packets in third row is 375.**

1) DROPOUT

In our proposed method, a portion of the CSI values was randomly set to zero with a probability λ, which was selected randomly from the range of (0, 0.07). This technique simulates the effect of signal loss and interference, a common occurrence in real-world wireless signal transmission. This method aims to make our model more resilient to such losses by training it to recognize activities even with incomplete data inputs.

2) MIX SAMPLES WITH DIFFERENT LABELS

The presence of external interferences, such as background radio noise and disturbances from moving objects, has been previously described. The intensity of background radio noise varies dynamically over time. To mitigate this issue, a technique of mixing samples with diverse labels has been proposed as a means of enhancing the robustness of the model. This approach is designed to simulate real-world scenarios where Wi-Fi signals are subject to a variety of distortions, thereby enhancing the robustness and generalization capability of our TCN-AA model. The following equation illustrates the mixing procedure:

$$D = A\cdot(1-\varepsilon_1) + B\cdot\varepsilon_2 + C\cdot\varepsilon_3. \tag{1}$$

where the new sample, D, is obtained as a result of mixing sample A, B, and C. The label of D is inherited from sample A, while the labels of samples B and C differ from that of sample A. In the course of our experiments, the value of ε_k was randomly selected from the range of (0, 0.05) to ensure the new sample D retains characteristics from all three, thereby introducing a controlled level of noise and complexity into the training data.

3) MIX SAMPLES WITH SAME LABELS

This method utilizes the same mixing equation as (1). The primary distinction is the mixing of samples that possess equivalent labels. This approach is motivated by the observation that moving object disturbance is a result of variations in subjects' heights and body shapes, even when they are performing the same activity. Given that moving object disturbance is expected to be more pronounced than background radio noise, it was deemed necessary to mix samples with the same label to enhance the robustness of the model.

D. MODEL

The CSI signal exhibits a strong correlation with time-sequence data, making it well-suited for the application of time-sequence deep learning models such as LSTM. The LSTM model operates through the use of three gates: the input gate, the forget gate, and the output gate. These gates enable the model to retain past memories; however, they also require substantial computational resources, resulting in longer training times. In contrast, the TCN [25], [26] presents a more suitable solution for our needs. The TCN utilizes one-dimensional causal convolution and dilated convolution in the temporal domain, making it a more efficient model compared to LSTM.

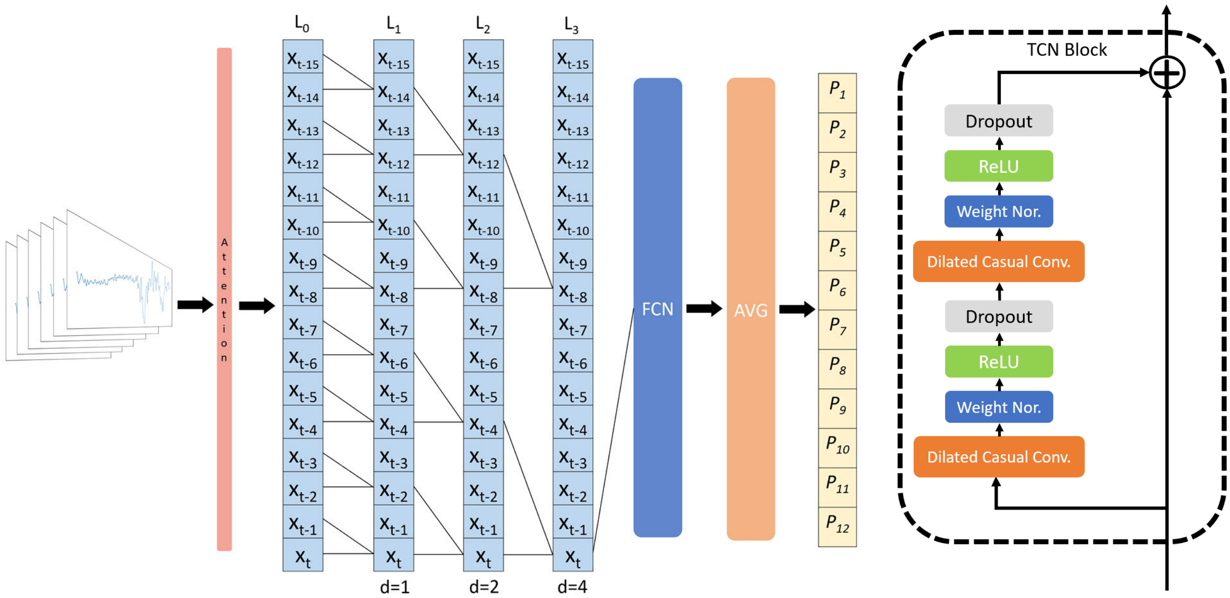

FIGURE 3. The architectural overall of our system in this paper. (a) An overview of the TCN model which give an example of kernel size = 2. The L_0 represents the input after attention. The dilation size of L_1, L_2, and L_3 is 1,2, and 4, respectively. The output, P_1 to P_{12}, is the probabilities of 12 classes. (b) A detail of each TCN layer.

In addressing the challenges of WiFi signal-based HAR, we specifically chose a TCN model enhanced with a tailored attention mechanism. We are the first to apply TCNs in the domain of WiFi CSI-based HAR. This decision was driven by the intrinsic temporal properties of CSI signals, which are rich in sequential data that traditional convolutional networks might not fully exploit. Our attention mechanism allows for dynamic weighting of temporal features, enabling the model to focus on the most significant segments of the data sequence for HAR.

The selection of the TCN model itself was justified by its three core strengths in handling time-sequenced data, particularly for CSI signals. First, the convolution operation within TCN is adept at feature extraction from time-series data, a capability well-documented in prior research. This is especially pertinent for CSI signals, where temporal patterns are crucial for accurate activity recognition. Second, the efficiency of TCN in terms of computational resources significantly reduces training times, a critical advantage over LSTM models which, despite their efficacy in capturing long-term dependencies, demand considerable computational power and time due to their complex gating mechanisms. Finally, the incorporation of causal and dilated convolutions in TCN enables effective learning from historical data without the need for the extensive data that LSTM models typically require.

Our work deviates from the traditional focus on RNNs and LSTMs by leveraging TCNs, which offer several significant advantages. TCNs capture long-range dependencies with their extended receptive fields, surpassing the capabilities of traditional RNN architectures. This is achieved without the common issues of vanishing or exploding gradients, which are often encountered in RNNs. Additionally, TCNs enable parallel processing of input sequences, resulting in faster training times and more stable convergence.

By integrating our attention mechanism into the TCN model, as illustrated in Fig. 3, our approach is distinctively positioned to harness the temporal dynamics of CSI signals, offering a nuanced understanding of human activities in WiFi-based environments. This targeted design not only elevates the model's performance in recognizing nuanced human activities but also underscores our contribution to advancing WiFi-based HAR methodologies. Notably, our research is pioneering in applying TCNs to WiFi signal classification, marking a significant contribution to this field.

While TCNs are highly effective in capturing temporal dependencies and handling sequential data within a specific context, they often struggle to generalize across different environments or scenarios. This limitation arises because TCNs, like many deep learning models, can be sensitive to variations in the input data distribution that were not present during training.

In practical terms, this means that our TCN-based model, although proficient in recognizing human activities from WiFi CSI signals within a controlled setting, may not perform as reliably when applied to different environments with varying signal characteristics. Such scenarios could include changes in the physical layout, interference from other electronic devices, or different patterns of human movement that were not represented in the training data.

1) CAUSAL CONVOLUTIONS

It is a commonly accepted practice in Seq2Seq models to maintain the same length of output as input, as well as to preserve past information. The TCN accomplishes this goal through the use of causal convolutions. To implement the TCN model, a full one-dimensional convolutional network architecture is employed, with a kernel size of k and a padding size of " k −1". The padding is accomplished using zeros, ensuring that each layer's output maintains the same length as the input, as demonstrated in Fig. 4. This is due to the fundamental principle of TCN, which prevents future information from leaking into the past. As a result, the output at time t is always convolved only with information from time t and earlier.

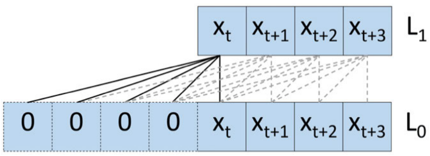

FIGURE 4. An example of 1D causal convolution while the kernel size is 5, the padding size is 4, and padding with 0.

2) DILATED CONVOLUTIONS

The utilization of Seq2Seq models presents a challenge in terms of memory capacity. While an increase in the model's ability to retain historical information results in improved performance, it also leads to a larger memory footprint. To address this issue, two possible solutions have been proposed. The first approach entails the utilization of large kernel sizes in causal convolutions; however, this option appears to be impractical. The alternative solution, as introduced by Yu and Koltun [27], involves the use of dilated convolutions, which have since become a widely utilized technique. The dilated mechanism enables the TCN to possess an exponentially expanding receptive field, as demonstrated in Fig. 5.

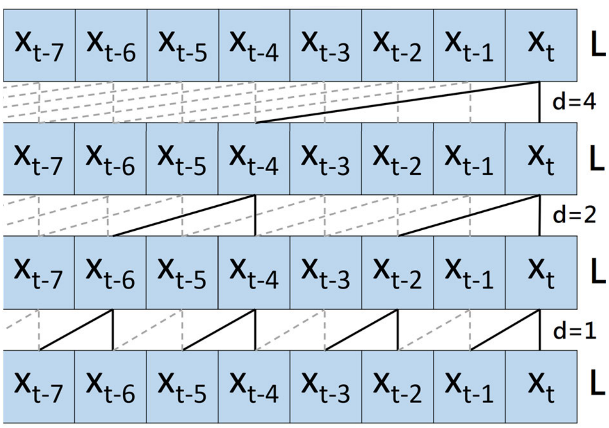

FIGURE 5. An example of dilated convolution while the kernel size is 2 and the delated size increases from 1 to 4 exponentially.

In our experiments, the dilated size was initially set at 1 and grew exponentially with a base of 2. For instance, in our model, there are three layers of TCN, each with a dilated size of 1, 2, and 4 respectively. Furthermore, we increased the kernel size (k = 15) in order to maintain a larger receptive field. Although training with a larger kernel size may take longer, it remains a more cost-effective solution compared to the use of LSTM networks. The comparison between the performance of LSTM and TCN will be discussed in Section IV.

3) TEMPORAL ATTENTION MECHANISM

The widespread success of transformers in recent years is largely attributed to their utilization of the attention mechanism. The implementation of an appropriate attention mechanism can result in both an improvement in accuracy and an acceleration of convergence speed during the training process. We adopts a mechanism of attention that is primarily inspired by the works of Hao et al. [26] and Vaswani et al. [28]. The input (H) is first linearly transformed into three distinct vectors, named query (Q), key (K), and value (V), respectively, as expressed by the following formulas:

$$\mathrm{Q} = \mathrm{W_Q} \cdot \mathrm{H}. \tag{2}$$

$$\mathrm{K} = \mathrm{W_K} \cdot \mathrm{H}. \tag{3}$$

$$\mathrm{V} = \mathrm{W_V} \cdot \mathrm{H}. \tag{4}$$

$$\text{Attention }(\mathrm{Q, K, V}) = \text{softmax}(\frac{QK^T}{\sqrt{d_K}})\mathrm{V}. \tag{5}$$

$$H' = H \cdot \text{Attention}(Q, K, V). \tag{6}$$

The weight of the linear transformation, represented by $\mathrm{W_Q}$, $\mathrm{W_K}$, and $\mathrm{W_V}$, is learned during the training phase. The outcome of the softmax operation performed on the product of the query (Q) and key (K) serves as a weighted factor which is divided by the square root of $\mathrm{d_K}$, the dimension of key (K), and then multiplied with the value (V). This attention mechanism allows the model to dynamically concentrate on the most pertinent features of the input during the prediction process. Upon computing the attention score, it is multiplied with the input (H) to produce a new input (H'). Subsequently, some modifications are made to (5) in accordance with the principles of causal convolutions. As previously mentioned, the fundamental principle of causal convolutions is that future information must not penetrate into the past. Thus, there is no requirement for calculating the correlation between the past and the future. As a result, (5) is revised as follows:

$$\text{Attention }(\mathrm{Q, K, V}) = \text{softmax }(\mathrm{L}(\frac{QK^T}{\sqrt{d_K}}))\mathrm{V}. \tag{7}$$

where the function L represents the lower triangular function. Upon performing the multiplication of the query and keys, and division by the square root of $\mathrm{d_K}$, the values above the main diagonal in the resulting matrix are set to zero, as demonstrated in Fig. 6. In consideration of time and performance constraints, we have limited the incorporation of attention to the first layer, as depicted in Fig. 3. The impact of adding attention at various layers will be assessed in Section IV for comparison purposes.

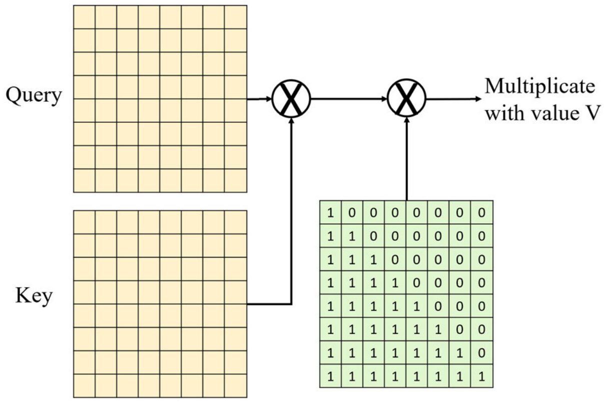

FIGURE 6. An example of the lower triangular function.

4) FCN AND AVERAGE POOLING

Our proposed model consists of three hierarchical layers of TCN blocks. The number of filters utilized in each layer has been set to a uniform value of 50 (i.e., $N_f^{L1} = N_f^{L2} = N_f^{L3} = 50$). As a result, the output of TCN is a three-dimensional vector which is in the shape of ($N_t \times N_r$, N_p, N_f^{L3}). The first dimension represents the TR-pair, the second dimension represents the temporal dimension, and the third dimension represents the frequency domain. Only the last term of the second dimension is subsequently fed into a fully connected network (FCN) [29], which takes the shape of ($N_t \times N_r$, 1, N_f^{L3}). The output of the FCN is then represented as ($N_t \times N_r$, 1, N_c). Considering the existence of $N_t \times N_r$ TR-pairs, each with a distinct signal, we interpret the output of the FCN as $N_t \times N_r$ individual probability distributions, each pertaining to one of the 12 classes, based on the respective TR-pair. In this scenario, we apply an average pooling on the $N_t \times N_r$ probability distributions to yield the final output of the model, which is of the shape (1, 1, N_c).

E. VARIABLE AND HYPERPARAMETERS

Here we provide a summary of the symbols and variables utilized, along with the experimentally determined hyperparameters that were chosen to achieve optimal performance.

We used a total of 200 epochs to train our model, following a 10-fold cross-validation approach to ensure robustness and generalizability. Our hyperparameters were selected based on empirical rules and extensive experimentation to achieve optimal performance.

Our hyperparameters are provided and these settings were chosen to balance the trade-off between model complexity and generalization. Specifically, the number of TCN layers and filters was set to capture sufficient temporal features while avoiding overfitting. The kernel size of 15 was selected to cover a wide range of temporal dependencies. We included the attention mechanism at the first layer to enhance the model's focus on relevant features early in the network.

The dropout rate of 0.5 was used to prevent overfitting, and a batch size of 32 was chosen to ensure stable gradient updates. The AdamW optimizer, which combines Adam optimization with weight decay, was selected for its effectiveness in training deep networks. We applied an exponential decay rate of 0.988 per epoch to the learning rate to gradually reduce it, helping the model converge more effectively.

TABLE 1. Variables and meanings in TCN-AA.

Variable	Meaning
N_c	The number of classes (i.e., 12).
N_p	The number of packets.
N_t	The number of antennas at the transmitter (i.e., 2).
N_r	The number of antennas at the receiver (i.e., 3).
N_s	The number of subcarriers (i.e., 30).
h_y^x	A complex number corresponding to the x^{th} packet and the y^{th} subcarrier for a particular TR-pair.
$H(T_i, R_j)$	A CSI value transmitted by the i^{th} transmitter and received by the j^{th} receiver.
λ	The dropout rate in data augmentation.
ε_k	The mixing rate in data augmentation where $k \in [1, 2, 3]$ in our method.
N_f^{Lm}	The number of filters in the m^{th} layer of TCN where $m \in [1, 2, 3]$ in our model.

TABLE 2. Hyperparameters of the TCN-AA model.

Hyperparameter	Value
Number of TCN layer	3
The number of filters at each layer	[50, 50, 50]
Kernel size (k)	15
Attention Mechanism	Add at 1st layer only
Dropout rate	0.5
Batch size	32
Optimizer	AdamW (Adam + weight decay)
The rate of exponential decay of learning rate	0.988/per epoch
Epoch	200
K-fold	10 folds

IV. EXPERIMENTS

In this section, a comprehensive experimental analysis will be performed to evaluate various aspects of our proposed method. The first comparison will be against the performance of an LSTM model that has been trained by ourselves. Subsequently, a comparison will be made with the results of other models that have been implemented on the HHI dataset. As stated in Section III, the dataset consists of 12 unique interactions, each performed by 40 subject pairs for 10 trials, yielding a total of 400 samples per interaction (class). Finally, the performance of the model will be further scrutinized through a series of ablation studies that focus on factors such as augmentation, kernel size, dropout, and attention.

A. COMPARE TO LSTM

Initially, we trained the HHI dataset using a bidirectional LSTM network. The architecture of the LSTM network is depicted in Fig. 7, consisting of two layers with 180 neurons per layer.

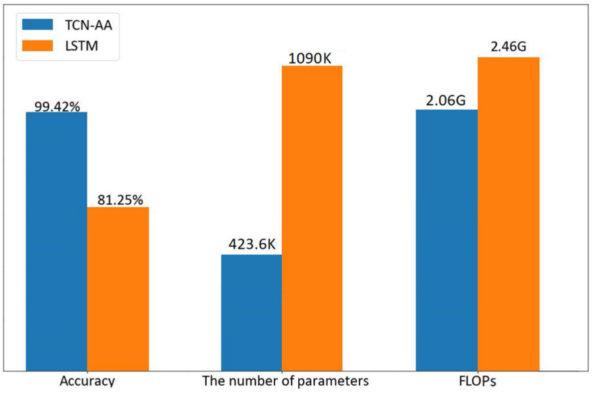

FIGURE 7. The comparison between our TCN-AA model and LSTM.

The LSTM was followed by a FCN [29] that outputted the likelihood of the 12 classes based on the features extracted by the LSTM. As demonstrated in Fig. 8, the accuracy of the LSTM model was approximately 81%, and its training time was three times longer compared to our proposed TCN model.

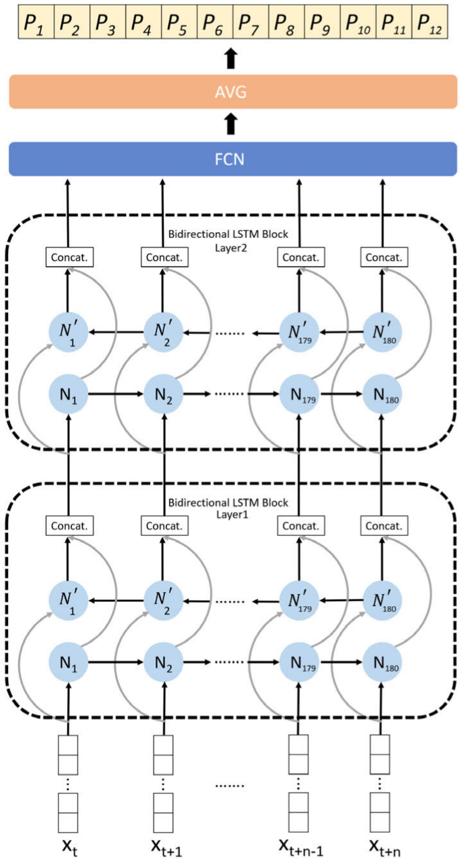

FIGURE 8. The architecture of LSTM model we used to compare.

B. COMPARE TO STATE-OF-THE-ART-METHODS

In this study, we will compare the performance of our TCN model with that of other methods that were applied to the same HHI dataset. These methods can be broadly categorized into two groups: traditional methods and deep learning-based methods. A customized support vector machine (SVM) proposed by [30] utilized principal component analysis (PCA) as preprocessing and the minimum redundancy maximum relevance (MRMR) algorithm to select a subset of the extracted features. Then, they applied SVM for classification, achieving an accuracy of 86.21%.

With respect to deep learning-based methods, Kabir et al. [19] introduced CSI-IANet, which utilized a Butterworth low-pass filter to denoise the CSI signal, employed three layers of CNN which is an inception module providing the model with varying receptive fields, and spatial-attention to first achieved an accuracy over 90% (91.3%). Kabir and Shin [31] presented DCNN, which employed only three layers of CNNs and achieved an accuracy of 88.66%. Shafiqul et al. [9] proposed HHI-AttentionNet, which utilized depth-wise CNNs and a customized attention mechanism, achieving an accuracy of 95.47%. Hao et al. [32] proposed a new Gabor residual block for CNN, combined with temporal attention and frequency attention, resulting in an accuracy of 86%. Alazrai et al. [33] proposed E2EDLF, which employed two layers of CNNs and achieved an accuracy of 86.3%.

Two additional methods based on Seq2Seq models were described. Khan et al. [34] presented Attention- BiGRUs, which used bidirectional gated recurrent units (Bi-GRUs) and multi-head attention, achieving an accuracy of 87%. Abdel-Basset et al. [22] proposed H2HI-Net, which utilized three layers of 1D CNNs with a residual network and BiGRUs, achieving an accuracy of 96.39%. As shown in Table 3, our TCN model achieved the best performance with an accuracy of 99.42%, outperforming H2HI-Net by 3%. In 2023, Zhou et al. [35] introduced CHA-Sens, a comprehensive residual convolution framework specifically designed for CSI-based human activity sensing. The CHA-Sens model, utilizing adaptive kernel size and stride, regional parameter-free attention mechanism, and shortcut of identity mapping, outperformed previous models on the HHIs dataset with an accuracy of 99.1%.

Fig. 9 illustrates the confusion matrix of our TCN-AA model's performance on the HHI dataset. The matrix showcases the model's predictive accuracy for each of the 12 interaction activities. High values along the matrix's diagonal represent a strong true positive rate for corresponding classes, indicating precise activity recognition.

The superiority of deep-learning-based methods over SVM methods is evident in our proposed approach. The

TABLE 3. The comparison over the models.

Model	Accuracy(%)
TCN-AA(Ours)	**99.42%**
SVM	86.21%
CSI-IANet	91.30%
DCNN	88.66%
HHI-AttentionNet	96.47%
GraSens	86.00%
E2EDLF	86.30%
Attention-BIGRU	87.00%
H2HI-Net	96.39%
CHA-Sens	99.1%

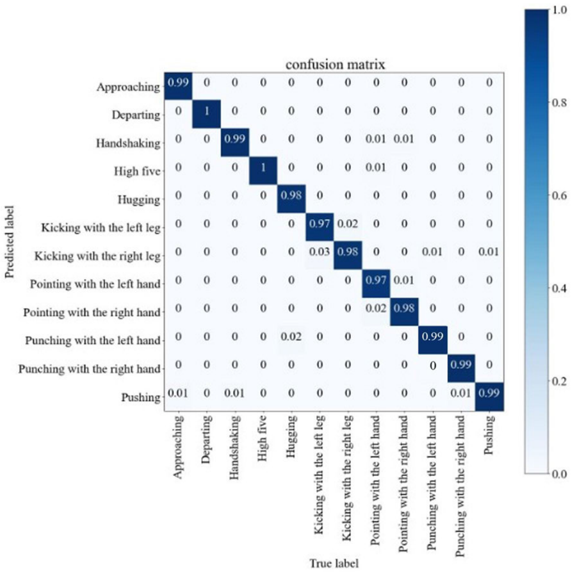

FIGURE 9. Confusion matrix of the TCN-AA model for the recognition of 12 human interaction activities using WiFi signals.

proposed augmentation methods and the use of the TCN model contribute to this advantage. The application of a causal convolution and a dilated convolution enables more efficient feature extraction in the temporal domain. Even without the proposed augmentation methods, our model demonstrates a high level of accuracy, registering 89.67%, which surpasses the performance of a significant number of other models. In comparison to Seq2Seq models, our model offers a higher level of efficiency while maintaining a smaller number of parameters, leading to reduced power consumption.

C. ABLATION STUDY

1) AUGMENTATION

As outlined in Section III, the proposed augmentation methods have resulted in a doubling of the amount of data. The raw data was utilized to evaluate performance, with the hyperparameters remaining the same, except for the kernel size, dropout rate, and attention. In this case, no attention networks were applied to the augmentation methods, and their performance was the sole focus of evaluation. The model was configured with a kernel size of 2 and a dropout rate of 0.2.

A comparison of three augmentation methods is presented, namely, dropout, mixing with the same label (i.e., Mix (Same)), and mixing with different labels (i.e., Mix (Other)). In the experiment, there were 400 samples in raw data for each class, and after each augmentation method, an additional 400 samples were generated for each class. As depicted in Table 4, the highest accuracy achieved using raw data alone was only 57%. However, the application of the different augmentation methods resulted in a significant increase in accuracy, with ranges between 88.54% and 90.18%. This demonstrates that the use of the augmentation methods can result in an improvement of around 30% in accuracy over the use of raw data alone.

TABLE 4. The comparison while doubling the amount of the data and fixed amount of the data. ***A*** represents the raw data. ***B***, ***C***, and ***D*** represent the method of dropout, mix(other), and mix(same), respectively. $(A+B)/2$ represents selecting a half of the amount of the data randomly.

Augmentation Method	Accuracy (%) Training	Validation	Loss Training	Validation
A	76.96%	56.81%	0.93	1.31
A+B	88.54%	85.21%	0.67	0.75
A+C	**90.18%**	**89.69%**	0.65	**0.69**
A+D	90.09%	88.74%	**0.64**	0.73
(A+B)/2	82.20%	69.15%	0.85	1.18
(A+C)/2	**82.47%**	69.57%	0.84	**1.16**
(A+D)/2	82.44%	**70.00%**	**0.83**	1.12

To evaluate the quality of the augmented data, a controlled set of experiments was conducted. Each augmentation method, namely dropout and mix with the same label, contributed an additional 400 samples per class, resulting in a doubled dataset of 800 samples per class. From this, a subset of 400 samples was randomly chosen for comparison with the original dataset. As indicated in Table 4, the model trained on augmented data demonstrated an average accuracy improvement of 13% over those trained exclusively on raw data. Fig. 10 provides a comparative visualization of the CSI data before and after the application of the augmentation methods, illustrating the enhanced signal diversity and potential for improved model robustness.

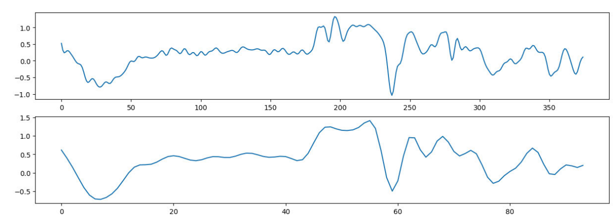

FIGURE 10. Comparative visualization of the 'Kicking with the right leg' class before (top) and after (bottom) data augmentation using dropout and mix with the same label methods.

Furthermore, we found that the impact of the amount of "Approaching" and "Departing" classes (class 1 and 2) on

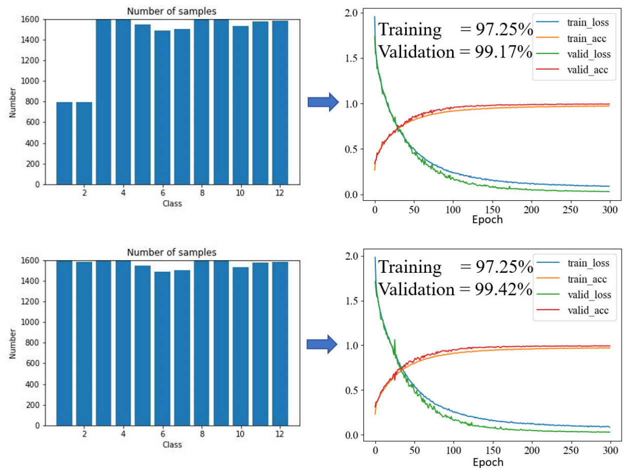

FIGURE 11. The comparison of different number of class 1 and 2.

accuracy are insignificant, as shown in Fig. 11. While the number of samples in class 1 and 2 are 800, the highest accuracy is 99.17%. The number of samples in class 1 and 2 are 1600, the highest accuracy is 99.42%.

2) KERNEL SIZE

In this experiment, an evaluation of the effect of different kernel sizes on performance was conducted. The evaluation was performed using raw data, and the attention network was not utilized. The results, as depicted in Table 5, indicate that accuracy increases and loss decreases as the kernel size increases. A plateau in both accuracy and loss appears to be reached when the kernel size reaches 15. Based on these observations, kernel size 15 was selected for use in subsequent experiments.

TABLE 5. The comparison of different kernel size.

Kernel Size	Accuracy (%)		Loss	
	Training	Validation	Training	Validation
2	62.55%	54.68%	1.17	1.28
3	73.87%	65.32%	0.95	1.08
7	92.75%	85.96%	0.40	0.53
11	96.88%	90.21%	0.26	0.42
15	98.16%	**90.64%**	**0.21**	**0.39**
19	98.28%	90.00%	**0.21**	0.40
23	98.42%	90.64%	0.22	0.46
27	97.85%	89.57%	0.23	0.45
31	**98.44%**	90.43%	0.22	0.46

3) DROPOUT

In this study, the impact of different dropout rates on performance during training in each TCN block was analyzed. The evaluation was performed on raw data, and the attention mechanism was not utilized. The results, as shown in Table 6, indicate that the lowest loss was achieved when the dropout rate was set to 0.6, while the highest accuracy was attained with a dropout rate of 0.5. Based on these findings, the

TABLE 6. The comparison of different dropout rate.

Dropout Rate	Accuracy (%)		Loss	
	Training	Validation	Training	Validation
0	**100%**	81.27%	**0.08**	1.39
0.1	**100%**	86.70%	0.10	0.71
0.2	99.72%	88.35%	0.14	0.53
0.3	99.40%	87.34%	0.17	0.49
0.4	98.84%	88.19%	0.19	0.44
0.5	97.87%	**89.89%**	0.23	0.42
0.6	96.40%	88.45%	0.27	**0.40**
0.7	92.37%	87.76%	0.37	0.42
0.8	84.51%	81.86%	0.52	0.53

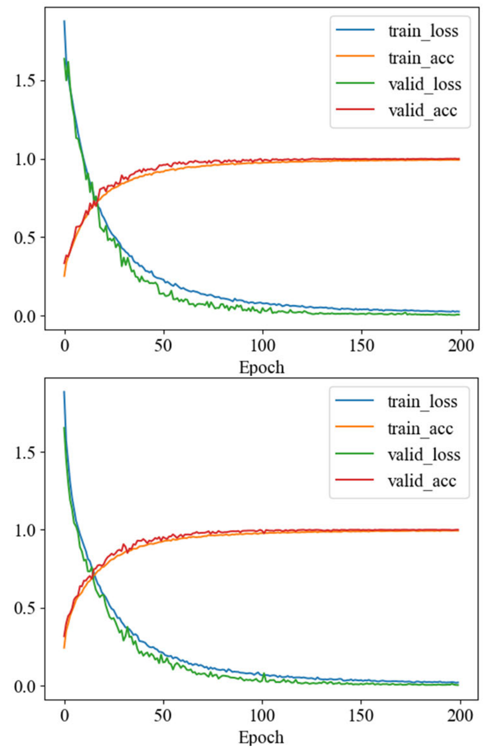

FIGURE 12. The loss and accuracy curve of one-layer attention. The curve above is applying attention before TCN and the curve below is applying after TCN.

dropout rate was set to 0.5 during the training phase. It should be noted that the validation accuracy, as depicted in Table 6, is expected to be higher than the training accuracy due to the operation of dropout. During the training phase, only 50% of the features are utilized for classification, while in the validation phase, 100% of the features are used.

4) ATTENTION

In the examination of attention, the performance was evaluated at various layers while attention was applied. Three separate experiments were conducted to assess the effect of attention on the model. Experiment 1 involved applying attention before TCN, experiment 2 involved applying attention to the output of TCN and feeding it to FCN, and experiment 3 involved applying attention at the beginning of each TCN layer. As shown in Fig. 12, the performance was almost

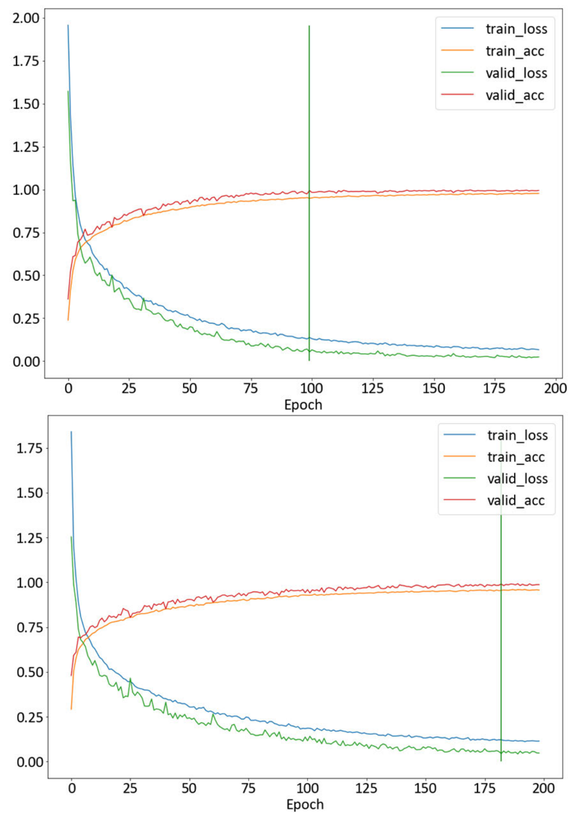

FIGURE 13. The loss and accuracy curve with attention and without attention.

identical when attention was applied in experiments 1 and 2. However, experiment 3 was discarded due to its high computational demands, as high accuracy could be achieved with a single layer of attention. Ultimately, it was decided to apply attention based on experiment 1.

Furthermore, the performance with and without attention was compared. According to Fig. 13, the models with attention reached 99% accuracy around the 100th epoch, and this convergence was faster than the models without attention, which reached 99% accuracy around the 180th epoch. This result demonstrates the significance of attention in achieving fast convergence during WiFi signal processing.

V. CONCLUSION AND FUTURE WORK

In this paper, a novel approach for recognizing human-human interactions using WiFi signals is presented. The proposed method incorporates an augmentation method, an attention mechanism, and a TCN-AA to efficiently extract features from time-sequence data while maintaining a low number of parameters. The experimental results demonstrate the superiority of the TCN-AA model over existing state-of-the-art methods, as it achieved a remarkable accuracy of 99.42% on the public dataset, outperforming the current best by 3%.

As future work, the applicability and adaptability of the proposed method in various environments, with different subjects, and for more complex behaviors will be investigated.

REFERENCES

[1] I. M. Shafiqul, M. K. A. Jannat, J.-W. Kim, S.-W. Lee, and S.-H. Yang, "HHI-AttentionNet: An enhanced human-human interaction recognition method based on a lightweight deep learning model with attention network from CSI," *Sensors*, vol. 22, no. 16, p. 6018, Aug. 2022.

[2] M. K. A. Jannat, M. S. Islam, S.-H. Yang, and H. Liu, "Efficient Wi-Fi-based human activity recognition using adaptive antenna elimination," *IEEE Access*, vol. 11, pp. 105440–105454, 2023, doi: 10.1109/ACCESS.2023.3320069.

[3] M. H. Uddin, J. M. K. Ara, M. H. Rahman, and S. Yang, "A study of real-time physical activity recognition from motion sensors via smartphone using deep neural network," in *Proc. 5th Int. Conf. Electr. Inf. Commun. Technol. (EICT)*, Dec. 2021, pp. 1–6.

[4] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognit. Lett.*, vol. 119, pp. 3–11, Mar. 2019.

[5] S. Münzner, P. Schmidt, A. Reiss, M. Hanselmann, R. Stiefelhagen, and R. Dürichen, "CNN-based sensor fusion techniques for multimodal human activity recognition," in *Proc. ACM Int. Symp. Wearable Comput.*, Sep. 2017, pp. 158–165.

[6] H. Liu, T. Xue, and T. Schultz, "On a real real-time wearable human activity recognition system," in *Proc. 16th Int. Joint Conf. Biomed. Eng. Syst. Technol.*, 2023, pp. 711–720.

[7] E. Essa and I. R. Abdelmaksoud, "Temporal-channel convolution with self-attention network for human activity recognition using wearable sensors," *Knowl.-Based Syst.*, vol. 278, Oct. 2023, Art. no. 110867.

[8] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, "A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions," *Knowl.-Based Syst.*, vol. 223, Jul. 2021, Art. no. 106970.

[9] R. Alazrai, A. Awad, B. Alsaify, M. Hababeh, and M. I. Daoud, "A dataset for Wi-Fi-based human-to-human interaction recognition," *Data Brief*, vol. 31, Aug. 2020, Art. no. 105668.

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

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

[12] J. Wilson and N. Patwari, "See-through walls: Motion tracking using variance-based radio tomography networks," *IEEE Trans. Mobile Comput.*, vol. 10, no. 5, pp. 612–621, May 2011.

[13] H. Li, X. He, X. Chen, Y. Fang, and Q. Fang, "Wi-motion: A robust human activity recognition using WiFi signals," *IEEE Access*, vol. 7, pp. 153287–153299, 2019.

[14] S. K. Yadav, S. Sai, A. Gundewar, H. Rathore, K. Tiwari, H. M. Pandey, and M. Mathur, "CSITime: Privacy-preserving human activity recognition using WiFi channel state information," *Neural Netw.*, vol. 146, pp. 11–21, Feb. 2022.

[15] J. Zhang, F. Wu, B. Wei, Q. Zhang, H. Huang, S. W. Shah, and J. Cheng, "Data augmentation and dense-LSTM for human activity recognition using WiFi signal," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4628–4641, Mar. 2021.

[16] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[18] F. Wang, S. Zhou, S. Panev, J. Han, and D. Huang, "Person-in-WiFi: Fine-grained person perception using WiFi," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jun. 2019, pp. 5452–5461.

[19] M. H. Kabir, M. H. Rahman, and W. Shin, "CSI-IANet: An inception attention network for human-human interaction recognition based on CSI signal," *IEEE Access*, vol. 9, pp. 166624–166638, 2021.

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

[21] B. Li, W. Cui, W. Wang, L. Zhang, Z. Chen, and M. Wu, "Two-stream convolution augmented transformer for human activity recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 1, pp. 286–293.

[22] M. Abdel-Basset, H. Hawash, N. Moustafa, and N. Mohammad, "H2HI-Net: A dual-branch network for recognizing human-to-human interactions from channel-state information," *IEEE Internet Things J.*, vol. 9, no. 12, pp. 10010–10021, Jun. 2022.

[23] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Tool release: Gathering 802.11n traces with channel state information," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 1, p. 53, Jan. 2011.
[24] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. 13th Eur. Conf. Comput. Vis.*, Zurich, Switzerland. Cham, Switzerland: Springer, Sep. 2014, pp. 740–755.
[25] S. Bai, J. Zico Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, *arXiv:1803.01271*.
[26] H. Hao, Y. Wang, S. Xue, Y. Xia, J. Zhao, and F. Shen, "Temporal convolutional attention-based network for sequence modeling," 2020, *arXiv:2002.12530*.
[27] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.
[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.
[29] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[30] R. Alazrai, A. Awad, M. Z. Ali, and M. I. Daoud, "An approach for recognizing two-human interactions using channel state information," in *Proc. 9th Int. Conf. Electr. Electron. Eng. (ICEEE)*, Mar. 2022, pp. 301–305.
[31] M. H. Kabir and W. Shin, "Deep convolution neural network based CSI signal classifier to recognize human-human interactions," in *Proc. IEEE Int. Conf. Consum. Electronics-Asia (ICCE-Asia)*, Nov. 2021, pp. 1–3.
[32] Y. Hao, Z. Shi, X. Mu, and Y. Liu, "GraSens: A Gabor residual anti-aliasing sensing framework for action recognition using WiFi," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 288–295.
[33] R. Alazrai, M. Hababeh, B. A. Alsaify, 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.
[34] M. M. U. Khan, A. B. Shams, and M. M. S. Raihan, "A prospective approach for human-to-human interaction recognition from Wi-Fi channel data using attention bidirectional gated recurrent neural network with GUI application implementation," 2022, *arXiv:2202.08146*.
[35] F. Zhou, W. Zhang, G. Zhu, H. Li, and Q. Shi, "CHA-Sens: An end-to-end comprehensive residual convolution framework for CSI-based human activity sensing," in *Proc. 26th Int. Conf. Comput. Supported Cooperat. Work Design (CSCWD)*, May 2023, pp. 113–118.

CHIH-YANG LIN (Senior Member, IEEE) is currently with the Department of Mechanical Engineering, National Central University, Taoyuan, Taiwan. Previously, he was the Dean of International Academy; the Chief of Global Affairs Office; and a member of the Department of Electrical Engineering, Yuan-Ze University, Taoyuan. He has been recognized as an IET Fellow and has contributed to over 200 papers that have been featured in a wide range of international conferences and journals. His research interests primarily focused on computer vision, machine learning, deep learning, image processing, big data analysis, and the design of surveillance systems.

CHIA-YU LIN received the B.Sc. degree in computer science and information engineering from the National Central University, Taiwan, where he is currently pursuing the Master of Science degree in computer science and information engineering. His research interests include human activity recognition, Wi-Fi-based recognition, and few-shot learning.

YU-TSO LIU received the M.S. degree in electrical engineering from the National Dong Hwa University, Taiwan. He is currently pursuing the Ph.D. degree in computer science and information engineering from the National Central University, Taiwan. He is also a Senior Engineer with TSEC Corporation, where he specializes in enterprise infrastructure network projects management. His research interests include wireless communication, the application of edge artificial intelligence, deep learning, and the Internet of Things.

YI-WEI CHEN received the B.Sc. degree in computer science and information engineering from the National Changhua University of Education, Taiwan. He is currently pursuing the Master of Science degree in computer science and information engineering from the National Central University, Taiwan. His research interests include human activity recognition and Wi-Fi-based recognition.

TIMOTHY K. SHIH (Senior Member, IEEE) is a Distinguished Professor with the National Central University (NCU), Taiwan. In NCU, he was the Vice Dean of College of EECS and the Founding Director of the Innovative AI Research Center. He was the Dean of the College of Computer Science, Asia University, Taiwan; and the Chairperson of the CSIE Department, Tamkang University, Taiwan. He is a fellow of the Institution of Engineering and Technology (IET). He was also the Founding Chairperson Emeritus of the IET Taipei Local Network. In addition, he is a Senior Member of ACM. He was the Conference Co-Chair of the 2004 IEEE Internet. He was the Founder and the Co-Editor-in-Chief of the *International Journal of Distance Education Technologies*, USA. He was an Associate Editor of IEEE Computing. He was an Associate Editor of IEEE Transactions on Learning Technologies, the *ACM Transactions on Internet Technology*, and IEEE Transactions on Multimedia. For more details, please visit http://tshih.minelab.tw.