Enhancing Robotic Arm Activity Recognition with Vision Transformers and Wavelet-Transformed Channel State Information

Rojin Zandi*, Kian Behzad*, Elaheh Motamedi*, Hojjat Salehinejad†‡, and Milad Siami*
*Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA †Kern Center for the Science of Healthcare Delivery, Mayo Clinic, Rochester, MN, USA ‡Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, USA {zandi.r, behzad.k,motamedi.e}@northeastern.edu, salehinejad.hojjat@mayo.edu, m.siami@northeastern.edu

arXiv:2407.06154v1 [cs.RO] 8 Jul 2024

***Abstract*—Vision-based methods are commonly used in robotic arm activity recognition. These approaches typically rely on line-of-sight (LoS) and raise privacy concerns, particularly in smart home applications. Passive Wi-Fi sensing represents a new paradigm for recognizing human and robotic arm activities, utilizing channel state information (CSI) measurements to identify activities in indoor environments. In this paper, a novel machine learning approach based on discrete wavelet transform and vision transformers for robotic arm activity recognition from CSI measurements in indoor settings is proposed. This method outperforms convolutional neural network (CNN) and long short-term memory (LSTM) models in robotic arm activity recognition, particularly when LoS is obstructed by barriers, without relying on external or internal sensors or visual aids. Experiments are conducted using four different data collection scenarios and four different robotic arm activities. Performance results demonstrate that wavelet transform can significantly enhance the accuracy of visual transformer networks in robotic arms activity recognition.**

***Index Terms*—Channel state information, robotic arm activity recognition, transformer networks, wavelet transform.**

I. Introduction

Franka Emika arms are a series of collaborative robotic arms, designed to work alongside humans [1]. These arms are known for their lightweight and compact design as well as their advanced capabilities in safety and ease of use. Franka Emika arms are utilized across a diverse range of applications in logistics, warehousing, automotive industry, and healthcare due to their versatility, precision, and ability to perform repetitive and complex tasks.

Activity recognition in robotic arms is the process of identifying and categorizing specific activities or movements being performed by a robotic arm. Activity recognition is essential for enabling robotic arms to comprehend their surroundings, interact proficiently with objects, and autonomously execute tasks [2]. To this aim, robotic arms are generally equipped with various sensors such as accelerometers, gyroscopes, joint encoders, force sensors, and vision systems [3]. These sensors gather data about the arm's movements, positions, forces exerted, and the surrounding environment. Activity recognition can pose challenges owing to task complexity, environmental variations, and intricacies and noise within sensory data. Vision systems, such as light detection and ranging (LiDAR), typically demand an unobstructed line of sight (LoS), which could be unfeasible in specific surroundings. Privacy also raises concerns, especially in the context of household robots [4].

Wireless sensing and detection leverages wireless signals and arrays of sensors to gather diverse types of data. In wireless communications, channel state information (CSI) refers to the information about the current state of a communication channel, such as its quality, fading, interference, and other characteristics [5]. Changes in the environment, including activities performed by robots and humans, can significantly affect CSI. Collected CSI measurements from Wi-Fi signals can be used for human activity recognition (HAR) using various machine learning methods [6]–[12] and hand movement velocity estimation [13].

Recently, CSI has also been used for robotic arms activity recognition (RAR) [14], [15]. A Franka Emika arm was utilized to perform four different activities. During performing the activities, an access point (AP), sniffer, and transmitter collected CSI measurements. A convolutional neural network (CNN), trained on the collected measurements, was used for robot arm activity recognition. Continuing from the prior work on RAR through Wi-Fi sensing, which employed CNNs, this study advances RAR using vision transformers (ViT) [16] and compares them with a convolutional neural network with long short-term memory (CNN-LSTM) [17] and basic transformer models [18]. Our ViT-based approach demonstrates remarkable enhancements over the earlier CNN-based approach, showcasing improved recognition accuracy and generalization.

To further refine our methodology, we have incorporated the discrete wavelet transform (DWT) as a novel noise reduction technique [19]. Noises such as multipath interference, which creates fading effects due to signals reflecting off surfaces, and co-channel interference from other wireless devices using

This material is based upon work supported in part by grants ONR N00014-21-1-2431, NSF 2121121, the U.S. Department of Homeland Security under Grant Award Number 22STESE00001-01-00, and by the Army Research Laboratory under Cooperative Agreement Number W911NF-22-2-0001 (all at Northeastern University). The views and conclusions contained in this document are solely those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security, the Army Research Office, or the U.S. Government.

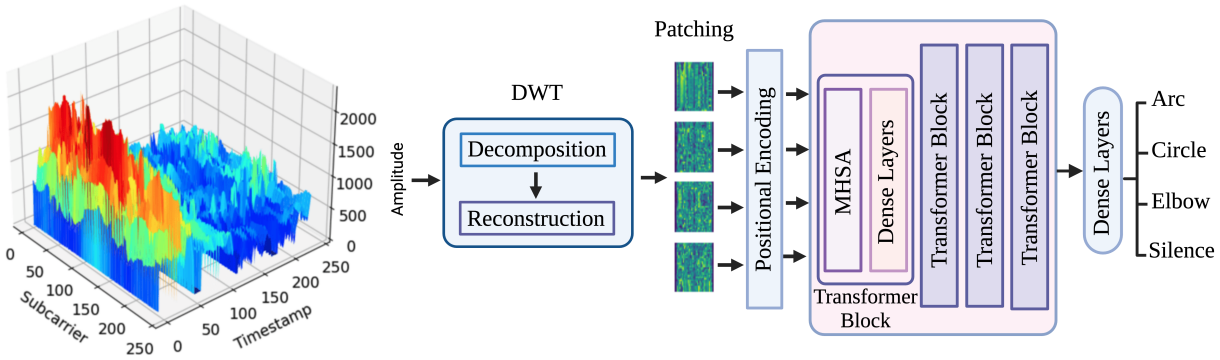


Figure 1. The proposed model for robotic arm activity recognition incorporates discrete wavelet transform (DWT) and multi-head self-attention (MHSA). The inputs to the model consist of the amplitudes derived from channel state information measurements.

the same frequency, can obscure the subtle changes in CSI data that are indicative of specific movements or activities. DWT, renowned for its efficacy in signal processing, operates by decomposing a signal into various frequency components, enabling the isolation and reduction of noise while preserving crucial signal features. This addition is particularly significant in Wi-Fi-based activity recognition, where signal clarity is paramount. By achieving superior results on the same dataset, our findings underscore ViT's potential for RAR tasks and contribute to the evolution of Wi-Fi-based activity recognition techniques.

II. Proposed Method

In the context of RAR, analyzing CSI data from both spatial and temporal dimensions is critical. Temporal analysis reveals the evolution of wireless signal characteristics over time, essential for tracking robotic movements and deciphering complex gestures [20], while spatial analysis provides insights into signal strength variations and multipath effects across locations. The CSI data is collected as a complex matrix where generally its amplitude is used for activity recognition [7], [14], [21]. Let $\{(\mathbf{X}_1, y_1), \ldots, (\mathbf{X}_N, y_N)\}$ represent a set of N CSI measurements where $\mathbf{X}_n \in \mathbb{R}^{S\times T}$ encapsulates the amplitude of the CSI and $y_n \in \{c_1, c_2, \ldots, c_M\}$ is the corresponding activity class for M number of possible activity classes. Figure 1 shows different steps of the proposed method where a CSI measurement $\mathbf{X}_n$ is the input and the target class is y_n. For convenience, a sample $\mathbf{X}_n$ is represented as $\mathbf{X}$ unless stated. The pseudocode of the model is presented in Algorithm 1 and each stage is detailed as follows.

A. Discrete Wavelet Transform for Noise Reduction

Let $\tilde{\mathbf{X}}_0 = \mathbf{X}$ represent an initial value, decomposed into a set of approximation coefficients using DWT as

$$\tilde{\mathbf{X}}_j[i,k] = \sum_{n=1}^{T} \tilde{\mathbf{X}}_{j-1}[i,n] \cdot \phi[n-2k], \tag{1}$$

and

$$\tilde{\mathbf{D}}_j[i,k] = \sum_{n=1}^{T} \tilde{\mathbf{X}}_{j-1}[i,n] \cdot \psi[n-2k], \tag{2}$$

where $\tilde{\mathbf{X}}_j[i,k]$ are the row-wise approximation coefficients and $\tilde{\mathbf{D}}_j[i,k]$ are the horizontal detail coefficients at level j using low-pass filter $\phi(\cdot)$ and high-pass filter $\psi(\cdot)$. A 1D DWT is then applied across each column of the row-transformed coefficients as

$$\tilde{\mathbf{X}}_j[k,l] = \sum_{m=1}^{S} \tilde{\mathbf{X}}_j[m,k] \cdot \phi[m-2l], \tag{3}$$

and

$$\hat{\mathbf{D}}_j[k,l] = \sum_{m=1}^{S} \tilde{\mathbf{X}}_j[m,k] \cdot \psi[m-2l], \tag{4}$$

where $\tilde{\mathbf{X}}_j[k,l]$ are the final approximation coefficients and $\hat{\mathbf{D}}_j[k,l]$ are the vertical detail coefficients at level j. Additionally, the diagonal detail coefficients $\overline{\mathbf{D}}_j[k,l]$ can be computed by applying $\psi(\cdot)$ across both dimensions as

$$\overline{\mathbf{D}}_j[k,l] = \sum_{m=1}^{S}\sum_{n=1}^{T} \tilde{\mathbf{X}}_{j-1}[m,n] \cdot \psi[m-2k] \cdot \psi[n-2l], \tag{5}$$

where the 2D decomposition results in one approximation matrix $\tilde{\mathbf{X}}_j$ and three detail matrices $\tilde{\mathbf{D}}_j$, $\hat{\mathbf{D}}_j$, and $\overline{\mathbf{D}}_j$ at each level j, capturing horizontal, vertical, and diagonal details respectively. After each level of decomposition, the approximation coefficients are updated iteratively for J levels. Then, a denoised CSI matrix is computed by applying the inverse 2D DWT as

$$\begin{aligned}\hat{\mathbf{X}}_j[m,k] = &\sum_{l=1}^{\beta} \tilde{\mathbf{X}}_j[k,l] \cdot \phi^{-1}[2m-l] \\ &+ \hat{\mathbf{D}}_j[k,l] \cdot \psi^{-1}[2m-l],\end{aligned} \tag{6}$$

and

$$\overline{\mathbf{D}}_j[m,k] = \sum_{l=1}^{\beta} \overline{\mathbf{D}}_j[k,l] \cdot \psi^{-1}[2m-l], \tag{7}$$

where $\beta = \frac{T}{2^j}$. The inverse DWT on rows using the newly obtained column coefficients is computed as

$$\begin{aligned}\hat{\mathbf{X}}_{j-1}[m,n] = &\hat{\mathbf{X}}_{j-1}[m,n] \\ &+ \sum_{k}^{\gamma} \overline{\mathbf{D}}_j[m,k] \cdot \psi^{-1}[2n-k],\end{aligned} \tag{8}$$

where $\gamma = \frac{S}{2^j}$. The inverse filters $\phi^{-1}(\cdot)$ and $\psi^{-1}(\cdot)$ are the reconstruction low-pass and high-pass filters, respectively. A

denoised matrix is obtained ffter the final reconstruction level, which is the reconstructed version of the original CSI matrix with reduced noise.

Algorithm 1 Denoising CSI Amplitude Measurements

Input: CSI amplitude $\mathbf{X} \in \mathbb{R}^{S \times T}$, number of levels J
Output: Denoised amplitude matrix $\tilde{\mathbf{X}} \in \mathbb{R}^{S \times T}$
procedure DWTDENOISING($\mathbf{X}$, J)
    Initialize $\tilde{\mathbf{X}}_0 = \mathbf{X}$
    for $j = 1 \rightarrow J$ **do**
        for $n = 1 \rightarrow T$ and $m = 1 \rightarrow S$ **do**
            Compute approx. coefficients using (1) & (3)
            Compute detail coefficients using (2), (4) & (5)
        end for
        Update $\tilde{\mathbf{X}}_{j-1} \leftarrow \tilde{\mathbf{X}}_j$
    end for
    for $j = J \rightarrow 1$ **do**
        for $l = 1 \rightarrow \beta$ and $k = 1 \rightarrow \gamma$ **do**
            Up-sample and convolve using (8)
        end for
    end for
    Set $\tilde{\mathbf{X}} \leftarrow \tilde{\mathbf{X}}_0$
    return $\tilde{\mathbf{X}}$
end procedure

B. Feature Extraction with Transformers

Each input sample $\tilde{\mathbf{X}}$ is divided to Z small patches $\hat{\mathbf{X}} \in \mathbb{R}^{P \times P}$, where $Z = \lceil \frac{ST}{P^2} \rceil$ and $\lceil \cdot \rceil$ is the ceiling function. Then, these patches are flattened as $\mathbf{x} \in \mathbb{R}^{P^2}$ and embedded to a vector using the positional encoder. Each sequence of these embedded vectors must pass through transformer blocks which contain a multi-head self-attention (MHSA), represented as $\mathbf{M}$, and a multi-layer perceptron (MLP) network. Flattening and embedding the Z patches, results in input sequence $\mathbf{\Lambda} \in \mathbb{R}^{Z \times D}$ where D is the latent vector size. The MHSA is an extension of self-attention $\mathbf{S}$ which runs L self-attentions. First, one must compute the query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ as

$$[\mathbf{Q}, \mathbf{K}, \mathbf{V}] = \mathbf{\Lambda} \mathbf{U}_{\mathbf{Q},\mathbf{K},\mathbf{V}}, \tag{9}$$

where $\mathbf{U}_{\mathbf{Q},\mathbf{K},\mathbf{V}} \in \mathbb{R}^{D \times 3D}$ is the set of parameters learned during training the network and $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V} \in \mathbb{R}^{Z \times D}$. Then, the attention weights are computed as

$$\mathbf{A} = \sigma\left(\frac{\mathbf{Q}\mathbf{K}^\top}{D^{\frac{1}{2}}}\right), \tag{10}$$

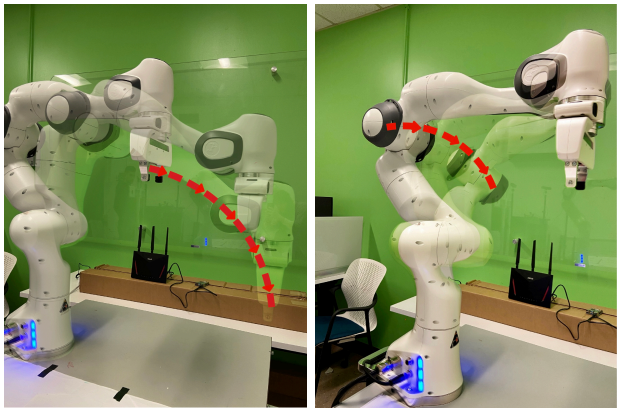
where $\sigma(\cdot)$ is the Softmax function and $\mathbf{A} \in \mathbb{R}^{Z \times Z}$. The self-attention $\mathbf{S}$ is computed as

$$\mathbf{S} = \mathbf{A}\mathbf{V}. \tag{11}$$

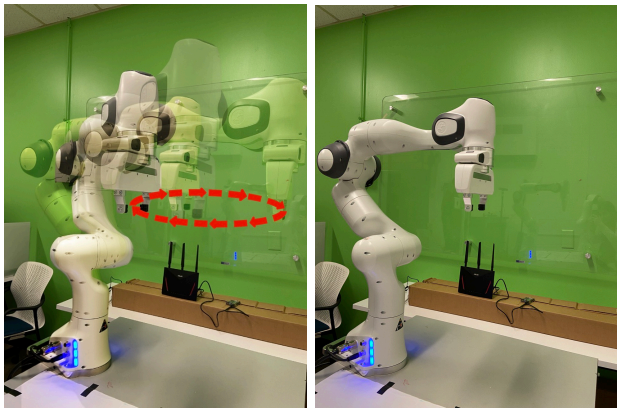
A collection of L self-attention headers are concatenated and projected as

$$\mathbf{M} = [\mathbf{S}_1 \oplus \mathbf{S}_2 \oplus \cdots \oplus \mathbf{S}_L]\mathbf{U}_{MHSA}, \tag{12}$$

where $\mathbf{M} \in \mathbb{R}^{Z \times D}$, $\mathbf{U}_{MHSA} \in \mathbb{R}^{(L \times D) \times D}$ and $\oplus$ is the concatenation operator. It is computed by applying standard multiplication on concatenated self-attentions and $\mathbf{U}_{MHSA}$.



(a) Arc (b) Elbow



(c) Circle (d) Silence

Figure 2. A Franka Emika arm performing four different activities: (a) Arc, (b) Elbow, (c) Circle in the XY-plane, and (d) Silence, [14].

Following the normalization and residual connection, the data passes through an MLP network. It processes the output of the MHSA layer-by-layer to extract more complex features and relationships, and then another round of normalization and residual connection, is done on the features.

C. Classification

The output features of the transformer blocks must be fed to the final MLP model to classify the CSI input samples to M classes. The loss function $\mathcal{L}$ of the model is computed for a batch of N samples as

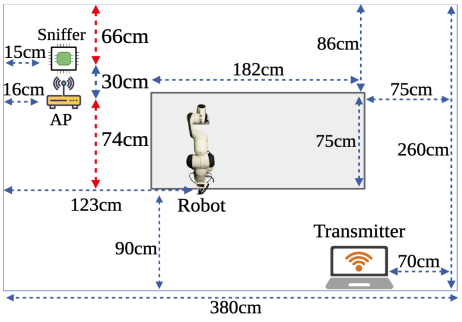
$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M} y_{n,m} \cdot \log\left(\sigma\left(\hat{y}_{n,m}\right)\right), \tag{13}$$

where $y_{n,m}$ is the true label for sample $\mathbf{X}_n$ and class m, and $\hat{y}_{n,m}$ is the is the corresponding logit predicted by the model.

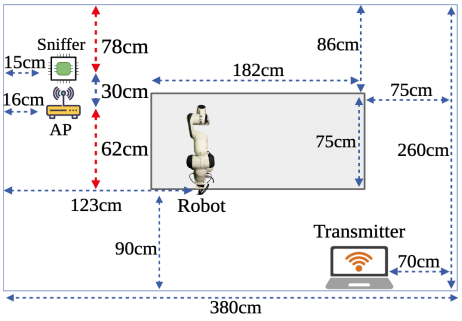
III. Experiments

A. Data

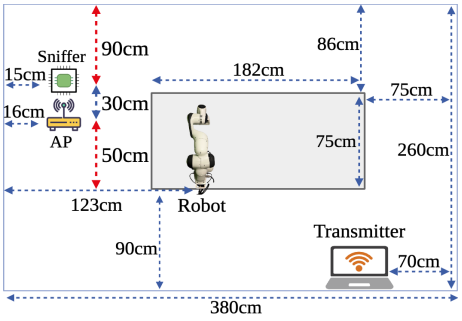
The dataset was acquired from a Franka Emika robotic arm [14] performing three activities which are (a) Arc, (b) Elbow, (c) Circular motion in the X-Y plane, and (d) No movement (Silence), as presented in Figure 2, based on the floor plans in Figure 3. Each activity was repeated for periods exceeding 5 minutes, capturing data at a frequency of 30Hz,



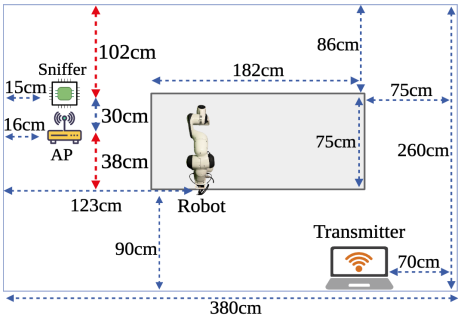
(a) Scenario 1



(b) Scenario 2



(c) Scenario 3



(d) Scenario 4

Figure 3. Different data collection scenarios. In each scenario, the red dashed arrows show the change in location of access point and sniffer, which have moved 12-cm horizontally.

resulting in the collection of $10,000$ packets by the router. The selected bandwidth was 80 MHz, utilizing $S = 256$ subcarriers. For every sample, a sequence of $T = 300$ consecutive CSI packets was employed. A total of 100 samples were collected for each activity class. As mentioned in section II, we focus on the amplitude of the CSI data and perform the removal of unused and pilot subcarriers. Consequently, each

Table I
AVERAGE AND STANDARD DEVIATION OF CLASSIFICATION METRICS OF EACH MODEL, AFTER 10-FOLD CROSS-VALIDATION, TESTED ON EACH SCENARIO IN PERCENTAGE.

Scenario	Model	Accuracy	Precision	Recall	F1-Score
1	CNN	90.9±3.2	91.6±3.9	90.9±3.2	90.4±3.9
	CNN-LSTM	79.2±5.3	79.4±6.0	79.3±5.3	79.0±5.6
	Transformer	69.8±5.4	74.5±5.2	69.8±5.4	70.0±5.3
	ViT	95.0±1.2	95.8±1.2	95.0±1.2	95.0±1.1
	ViT-DWT	**96.7±1.2**	**96.4±0.7**	**96.7±1.2**	**96.7±1.2**
2	CNN	85.7±4.1	85.6±4.7	85.7±4.1	85.6±4.6
	CNN-LSTM	81.8±4.7	86.5±4.2	81.8±4.7	81.3±5.1
	Transformer	63.5±4.6	73.6±4.3	63.5±4.6	63.3±4.0
	ViT	93.7±1.7	94.4±1.5	93.7±1.7	93.7±1.9
	ViT-DWT	**96.0±2.2**	**96.2±1.3**	**96.0±2.2**	**96.0±2.1**
3	CNN	84.1±5.4	87.0±4.7	84.1±5.4	84.4±5.0
	CNN-LSTM	82.4±2.7	82.8±1.9	82.4±2.7	82.5±1.5
	Transformer	70.4±3.4	79.1±2.8	70.4±3.4	70.0±3.0
	ViT	96.9±1.2	97.2±1.2	96.9±1.2	96.9±1.1
	ViT-DWT	**97.4±1.4**	**97.7±2.3**	**97.4±1.4**	**97.3±2.5**
4	CNN	88.1±4.3	88.0±4.3	88.1±4.3	87.8±3.9
	CNN-LSTM	83.6±4.0	85.3±3.7	83.6±4.0	83.8±4.1
	Transformer	65.4±5.3	75.5±4.9	65.4±5.3	62.3±4.4
	ViT	95.6±2.0	96.0±2.1	95.6±2.0	95.6±1.8
	ViT-DWT	**97.4±1.4**	**97.6±1.2**	**97.4±1.4**	**97.3±1.4**

sample undergoes a reshaping process, resulting in a matrix $\mathbf{X} \in \mathbb{R}^{234\times300}$, and then DWT with $J = 10$ is performed on $\tilde{\mathbf{X}}$ and finally the denoised signal is ready to be fed to the feature extraction model.

B. Training Setup

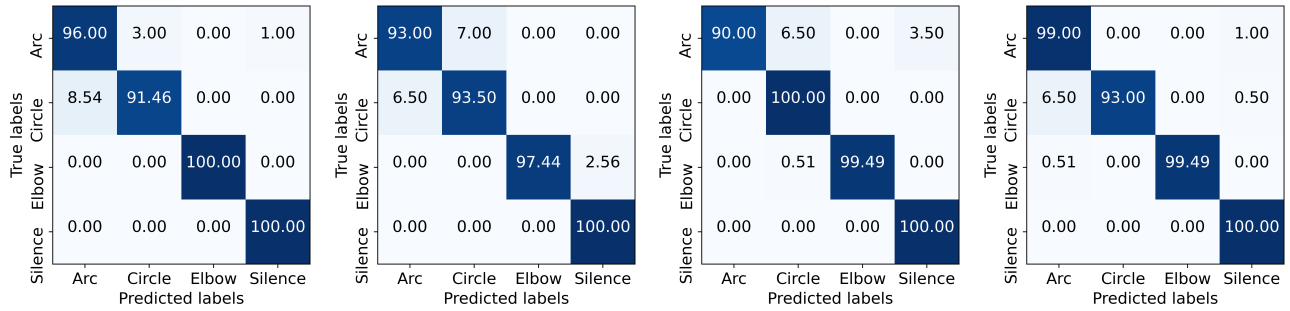
The inherent complexity and sequential nature of CSI measurements come with challenges for conventional pattern recognition methods, which led us to investigate alternative models beyond CNN. In the case of the CNN-LSTM model, we incorporated two LSTM layers into the existing CNN structure. In contrast, for training the transformer model, we flattened each CSI matrix and concatenated them into a vector.

The ViT model, tailored for CSI classification, employs a patch size of 20 to capture spatial hierarchies effectively. It is notable that the patches must be square and non-overlapping, so we zero-padded each CSI sample to have input samples with dimension 300×300, and after patching each sample results in 225 patches. It undergoes 60 epochs of training with a batch size of 64 and early stopping with 10 epochs patience, a learning rate of 1×10^{-2}, dropout rate of 0.3, and weight decay of 2×10^{-3}. A grid search was conducted using the validation dataset to tune the hyperparameters. A sensitivity analysis is provided in Subsection III-D.

C. Results Analysis

To investigate the impact of altering the sniffer's location, we adopted a leave-one-scenario-out (LOSO) approach, as previously demonstrated in [14]. This strategy involves training our models on three distinct scenarios and testing them on the fourth. In each scenario, both the sniffer and AP were relocated by 12 cm.

The average and standard deviation of the classification metrics for the CNN, CNN-LSTM, transformer, ViT, and



(a) Scenario 1 (b) Scenario 2 (c) Scenario 3 (d) Scenario 4

Figure 4. Average confusion matrix values of ViT-DWT, in percentage, after 10-fold cross-validation for scenarios in Figure 3.

Table II
MEAN AND STANDARD DEVIATION OF ALL LEAVE-ONE-SCENARIO-OUT PERFORMANCE EVALUATIONS OF VIT AND VIT-DWT, FOR EACH ROBOTIC ARM ACTIVITY CLASS, IN PERCENTAGE.

Model	Activity	Accuracy	Precision	Recall	F1-Score
ViT	Arc	90.0±9.2	94.3±6.5	90.0±9.2	91.6±4.5
	Circle	95.0±5.9	92.5±7.9	95.0±5.9	93.3±3.7
	Elbow	97.4±4.4	100.0±0.0	97.4±4.4	98.6±2.3
	Silence	98.7±2.2	96.7±5.6	98.7±2.2	97.6±2.8
ViT-DWT	Arc	94.5±5.3	95.1±5.3	94.5±5.3	94.6±3.6
	Circle	95.9±5.0	94.5±4.9	95.9±5.0	95.1±3.8
	Elbow	99.1±2.9	100.0±0.0	99.1±2.9	96.0±4.3
	Silence	100.0±0.0	97.7±3.2	100.0±0.0	98.8±2.1

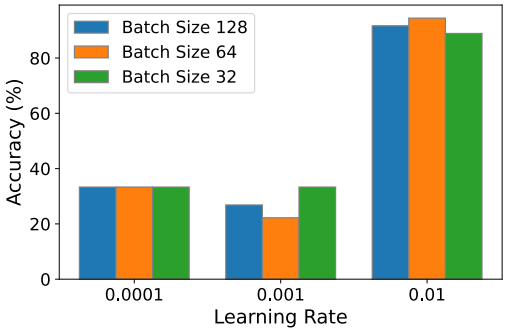


Figure 5. Performance of the ViT-DWT model after grid search for different learning rate and batch size values.

ViT-DWT models are reported in Table I after 10-fold cross-validation. Notably, the ViT model exhibits superior performance compared to the CNN model developed in [14]. The ViT model processes an input CSI measurement as a sequence of patches, where the attention mechanism can capture complex temporal and spatial relationships between patches with a higher performance than the CNN and regular transformer networks. This aligns with the efficacy of attention-based methods in Wi-Fi sensing, as demonstrated in prior work on HAR [22]. The ViT's superior performance over both CNN and regular transformers underscores the potential of attention mechanisms in this domain.

To have a better understanding of performance of the proposed ViT-DWT model, the average confusion matrix values of each scenario and corresponding classification metrics per activity class are presented in Figure 4 and Table II. This table presents the average and standard deviation of each class after 10-fold cross-validation and over all the LOSO experiments for the both ViT and ViT-DWT models. Notably, the ViT-DWT model achieved an average accuracy of over 96.7%, with the highest precision recorded for the Elbow activity class. This high precision indicates the model's ability to accurately classify activities without many false positives. Moreover, the model exhibited strong recall rates, emphasizing its capacity to correctly identify instances of each activity class. The F1-Score, which balances precision and recall, further substantiates the model's overall effectiveness.

D. Sensitivity Analysis

Sensitivity of the proposed model to different learning rate values $\{10^{-4}, 10^{-3}, 10^{-2}\}$ and batch sizes $\{32, 64, 128\}$ is presented in Figure 5 with respect to the accuracy of the model. The results show that a learning rate of 10^{-2} and batch size of 64 achieve the best performance. With respect to the activation function, by adopting rectified linear unit (ReLU), we enabled the model to capture nonlinear relationships while mitigating vanishing gradient issues. These choices collectively enabled our ViT model to achieve exceptional accuracy in classifying different activities. Key to this model's success is the integration of two attention heads per transformer layer. With four transformer layers and $D = 256$, this architecture strikes a balance between depth and breadth of feature extraction.

IV. CONCLUSION

Wi-Fi sensing has conclusively demonstrated its efficacy in HAR and RAR, representing a significant advancement in the field. Our investigation transcended traditional CNN approaches by incorporating advanced architectures like CNN-LSTM, transformer networks, and ViT, with ViT emerging as the superior model due to its unparalleled performance. This superiority was further enhanced by denoising CSI measurements with DWT. Our findings underscore the pivotal role of ongoing innovation in modeling techniques to drive the evolution and refinement of RAR systems. The objective is

to achieve greater precision and context awareness in understanding robotic activities, thereby pushing the boundaries of Wi-Fi sensing applications.

REFERENCES

[1] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta, "R3m: A universal visual representation for robot manipulation," *arXiv preprint arXiv:2203.12601*, 2022.

[2] Xing Li, Junpei Zhong, and MM Kamruzzaman, "Complicated robot activity recognition by quality-aware deep reinforcement learning," *Future Generation Computer Systems*, vol. 117, pp. 480–485, 2021.

[3] Mary B Alatise and Gerhard P Hancke, "A review on challenges of autonomous mobile robot and sensor fusion methods," *IEEE Access*, vol. 8, pp. 39830–39846, 2020.

[4] Juan Li, Xiang He, and Jia Li, "2d LiDAR and camera fusion in 3d modeling of indoor environment," in *2015 National Aerospace and Electronics Conference (NAECON)*. IEEE, 2015, pp. 379–383.

[5] Yongsen Ma, Gang Zhou, and Shuangquan Wang, "Wifi sensing with channel state information: A survey," *ACM Computing Surveys (CSUR)*, vol. 52, no. 3, pp. 1–36, 2019.

[6] Yi Zhang, Yue Zheng, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang, "Widar3. 0: Zero-effort cross-domain gesture recognition with wi-fi," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[7] Hojjat Salehinejad and Shahrokh Valaee, "LiteHAR: Lightweight Human Activity Recognition from WiFi Signals with Random Convolution Kernels," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4068–4072.

[8] Hojjat Salehinejad, Navid Hasanzadeh, Radomir Djogo, and Shahrokh Valaee, "Joint human orientation-activity recognition using wifi signals for human-machine interaction," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[9] Qing Jiang, Ailin He, and Xiaolong Yang, "Device-free human activity recognition based on random subspace classifier ensemble," in *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2018, pp. 590–591.

[10] Siamak Yousefi, Hirokazu Narui, Sankalp Dayal, Stefano Ermon, and Shahrokh Valaee, "A survey on behavior recognition using wifi channel state information," *IEEE Communications Magazine*, vol. 55, no. 10, pp. 98–104, 2017.

[11] Hojjat Salehinejad, Navid Hasanzadeh, Radomir Djogo, and Shahrokh Valaee, "Contrastive representation of channel state information for human body orientation recognition in interaction with machines," in *2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2023, pp. 1–6.

[12] Radomir Djogo, Hojjat Salehinejad, Navid Hasanzadeh, and Shahrokh Valaee, "Fresnel zone-based voting with capsule networks for human activity recognition from channel state information," *IEEE Internet of Things Journal*, 2024.

[13] Navid Hasanzadeh, Radomir Djogo, Hojjat Salehinejad, and Shahrokh Valaee, "Hand movement velocity estimation from wifi channel state information," in *2023 IEEE 9th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, 2023, pp. 296–300.

[14] Rojin Zandi, Hojjat Salehinejad, Kian Behzad, Elaheh Motamedi, and Milad Siami, "Robot motion prediction by channel state information," in *2023 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2023.

[15] Rojin Zandi, Kian Behzad, Elaheh Motamedi, Hojjat Salehinejad, and Milad Siami, "Robofisense: Attention-based robotic arm activity recognition with wifi sensing," *arXiv preprint arXiv:2312.15345v3*, 2023.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[17] Ronald Mutegeki and Dong Seog Han, "A CNN-LSTM approach to human activity recognition," in *2020 international conference on artificial intelligence in information and communication (ICAIIC)*. IEEE, 2020, pp. 362–366.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[19] Luisa Pasti, Beata Walczak, DL Massart, and P Reschiglian, "Optimization of signal denoising in discrete wavelet transform," *Chemometrics and intelligent laboratory systems*, vol. 48, no. 1, pp. 21–34, 1999.

[20] Biyun Sheng, Fu Xiao, Letian Sha, and Lijuan Sun, "Deep spatial–temporal model based cross-scene action recognition using commodity wifi," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 3592–3601, 2020.

[21] Bo–Yi Chang and Jang–Ping Sheu, "Indoor localization with csi fingerprint utilizing depthwise separable convolution neural network," in *2022 IEEE 33rd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2022, pp. 1276–1281.

[22] Aichun Zhu, Zhonghua Tang, Zixuan Wang, Yue Zhou, Shichao Chen, Fangqiang Hu, and Yifeng Li, "Wi-atcn: Attentional temporal convolutional network for human action prediction using wifi channel state information," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 4, pp. 804–816, 2022.