Adversarial Human Activity Recognition Using Wi-Fi CSI

†Harshit Ambalkar, †Xuyu Wang, ‡Shiwen Mao
†Department of Computer Science, California State University, Sacramento, CA 95819, USA
‡Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849-5201, USA
Email:hambalkar@csus.edu, xuyu.wang@csus.edu,smao@ieee.org

***Abstract*—Human activity recognition has been used for various applications in Internet of Things (e.g., health monitoring, security, and sport-related monitoring). Wi-Fi channel state information (CSI) is widely used for activity recognition, where CSI can capture human activities that influence wireless channel. In this paper, we study the impact of adversarial attacks on deep neural network (DNN) based human activity recognition with Wi-Fi CSI. First, we discuss the system framework, where activity recognition can be considered as a classification problem and a specific DNN model is introduced. Then, we discuss adversarial attack problem for DNN-based human activity recognition and formulate three white-box attacks. In the experiment with a public Wi-Fi CSI dataset, our results show that the performances of DNN-based human activity classification are greatly influenced by three white-box adversarial attacks.**

I. Introduction

Human activity recognition has gained great attentions in Internet of Things (IoT), which is widely employed in human health monitoring (e.g., fall detection for elders), human-computer interaction (HCI), security and surveillance, sport-related analysis. Currently, several techniques have been used for human activity recognition including sensor-based, radar-based, RFID-based, and Wi-Fi based methods [1]. For sensor-based methods, the wearable device (e.g., smartwatch) with motion sensors (e.g., an accelerator and a gyroscope) can recognize different activities such as step counter. Moreover, commodity radars such as Doppler-based or frequency-modulated continuous-wave (FMCW)-based can be exploited for contactless activity monitoring. Also, RFID is also used for activity monitoring and pose estimation with cheap RFID tags. Currently, Wi-Fi channel state information (CSI) becomes a mainstreaming research for activity monitoring using signal processing and machine learning methods. For example, CARM [2] is the first system work to use CSI to analyze activity recognition, where discrete wavelet transform (DWT) is used to extract features of CSI, and hidden Markov model (HMM) is exploited for activity classification, which requires manual feature extractions.

Compared with traditional machine learning, deep learning techniques have high learning capacity and could automatically extract the features, which also benefits human activity classification. In this paper, we mainly focus on Wi-Fi CSI based methods. For example, long short-term memory (LSTM) technique is first used for human activity recognition with Wi-Fi CSI amplitudes, which can obtain better performance compared with CARM system [3]. In addition, attention based Bi-directional LSTM (BiLSTM) model is proposed to improve the accuracy in different Wi-Fi CSI datasets [4]. In addition, unsupervised adversarial domain adaption is also used to address wireless environment change problems [5]. Generative adversarial networks (GAN) is used to augment the training data to improve activity classification accuracy [6].

Although deep learning could improve the performance of human activity recognition with Wi-Fi CSI, the deep neural networks (DNN) are easily misled by adversarial examples that are generated by adding a subtle perturbations [7]. Fast Gradient Sign Method (FGSM) attack method is the first to use an one-step attack method to generate adversarial examples [8]. To enhance the performance of adversarial examples, other iterative-based methods are also proposed such as Projected Gradient Descent (PGD) [9] and Momentum Iterative Method (MIM) [10]. Currently, adversarial attacks are being studied for different applications. For example, adversarial attacks have been only used for sensor-based [11] and radar-based human activity recognition [12]. Also, there are some work to study the impact of adversarial attacks on wireless communication systems (e.g., modulation recognition, end-to-end communication system, indoor localization) [13], [14].

Motivated by the previous works, we study the impact of adversarial attacks on DNN-based human activity recognition with Wi-Fi CSI. The main idea is to leverage adversarial examples to evaluate the models' classification performance by adding a small perturbation to Wi-Fi CSI. Specifically, we use Wi-Fi CSI amplitude information for human activity recognition, because CSI amplitude is stable compared with CSI phase information (where carrier frequency offset (CFO) brings random phase errors over different packets). In this paper, we first discuss the system model including the offline stage and the online stage. In the offline stage, we propose a modified BiLSTM model, which can capture the CSI sequence features, and automatically study the importance of features and time steps, thus improving the accuracy of human activity recognition. Then, we introduce three white-box attack methods (i.e., FGSM, PGD, and MIM) and evaluate the performances of DNN-based human activity recognition model using a public Wi-Fi CSI dataset.

The main contributions of this paper are summarized as follows.

- To the best of our knowledge, this is the first work to

study the impact of adversarial attacks on DNN-based human activity recognition using Wi-Fi CSI.
- We discuss the system model, including system architecture, and attention-based BiLSTM model for human classification. Then, we also introduce three white-box attacks methods.
- Using a public Wi-Fi CSI dataset, our experimental results show that the three white-box attack methods greatly mislead the performance of the used DNN model for human activity recognition.

In the following, the preliminaries are introduced in Section II. We present the system model in Section III and adversarial attack models in Section IV. Our experimental study in Section V. Section VI summaries this paper.

II. Preliminaries

A. Channel State Information Preliminaries

Many wireless communication systems (e.g., Wi-Fi, LTE, and 5G) leverage orthogonal frequency-division multiplexing (OFDM) techniques in physical layer to obtain high data rate and address frequency selective channel fading. Generally, OFDM systems divide a large bandwidth into several orthogonal small bands (i.e., subcarriers). Also, cyclic prefix is used as a guard interval to address intersymbol interference (ISI). Fast Fourier transform (FFT) and inverse FFT (IFFT) are also exploited in the receiver and the transmitter to implement the OFDM systems, respectively.

Recently, several 802.11n/ac measurement tools (e.g., Intel WiFi Link 5300 NIC [15]) become public, which can easily extract CSI data from off-the-shelf Wi-Fi devices. Our model uses the Intel 5300 NIC to collect CSI data including 30 out of the 56 subcarriers at the WiFi receiver for a 20MHz or 40MHz channel. Generally, Wi-Fi CSI can capture the multipath effect in indoor environments, which includes static paths and dynamic paths. For wireless sensing, the static paths (i.e., static vector) is constant, while the dynamic paths (i.e., dynamic vector) is variable that is reflected by the moving object in different activities, or different locations. Thus, the complex CSI value is also defined by

$$H(f,t) = (H_s(f) + \sum_{i \in D} a_i(f,t) e^{-j\frac{2\pi d_i(t)}{\lambda}}) e^{-j2\pi\Delta f t} \quad (1)$$

where $H(f,t)$ is the CSI for the carrier frequency f at time t, $H_s(f)$ is the static vector, D is the set of dynamic paths, $a_i(f,t)$ and $d_i(t)$ are the channel attenuation and the ith path length at time t, $e^{-j2\pi\Delta f t}$ is the phase shift because of the carrier frequency difference.

Human activities (e.g., walking, standing up) will influence CSI dynamic paths, thus leading to different complex CSI values over time. The complex CSI value can be defined $H_i = |H_i| \exp(j\angle H_i)$, where $|H_i|$ and $\angle H_i$ are the amplitude response and phase response of subcarrier i, respectively. Generally, CSI amplitude and phase difference between two antennas are stable, which can be used for wireless sensing applications (e.g., indoor localization, vital sign monitoring, and human activity recognition) [16], [17].

B. Adversarial Machine Learning

When a large dataset (e.g., image, audio, text, or wireless data) is available, deep neural networks (DNN) have become a powerful tool to solve the complex real-world problems (e.g., classification, regression, and data compression and generation). However, adversarial machine learning (i.e., a type of machine learning methods) can fool deep neural networks by adding a small perturbation into the input data (i.e. adversarial examples) [7]. Specifically, a well-trained DNN model is easily attacked by adversarial examples, which will lead to incorrect classification.

Currently, adversarial examples can be generated by white-box attacks and black-box attacks. When performing white-box attacks, the attacker can access the entire information including network framework, training weights and gradients, and the dataset. Thus, the white-box attack can always obtain a stronger attack by carefully crafting the adversarial examples. For black-box attacks, the adversary does not know DNN framework and weights, which can query the output of the DNN model with the available input data. Also, the black-box attack can use a local substitute DNN with a synthetic dataset to generate the adversarial examples, which can not only misclassify the substitute DNN model but also the target DNN model [18].

III. System Model

Deep Learning has been used for human activity recognition problems to improve high classification accuracy, compared with the traditional machine learning. However, DNN models are vulnerable to the adversarial examples, which are only slightly different from the original data. In the section, we will discuss the system framework under adversarial attacks and introduce DNN-based human activity classification problem.

A. System Architecture

Fig. 1 represents the system architecture of the human activity recognition, which includes an offline training stage and an online test stage. We consider Intel 5300 NIC to collect WiFi CSI data, where CSI amplitude is used for activity recognition because its stability. Generally, we can obtain 90 CSI subcarriers from three antennas as the input of DNN model. In the offline stage, the training dataset is exploited to train the DNN model to classify K activities. Note that in this paper, we use the public dataset including seven different activities. In the online stage, we use three three white-box attacks to generate adversarial examples, which will be introduced to the new Wi-Fi CSI amplitude data. In addition, the trained DNN model is used to validate the performance of human activity recognition in the online stage.

B. Problem Formulation

In human activity recognition problem, we denote x as the input sequence data (i.e., CSI amplitudes from 500 packets and 90 subcarriers) in a sliding window, and define y the output label (e.g., seven activities in this paper). Further, we exploit f to represent the DNN model function, $\mathcal{L}$ to denote the loss

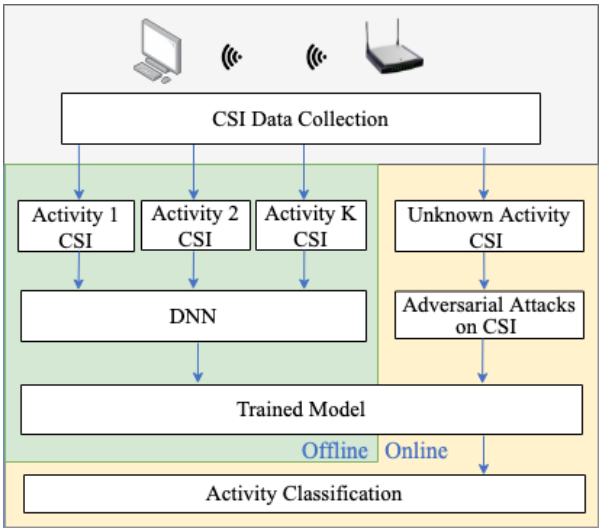

Fig. 1. System architecture.

function of the DNN model (i.e. categorical cross-entropy for human activity classification problem), and θ to represent the weight parameters of the DNN model. For human activity classification problem with Wi-Fi CSI, the objective of the problem is to minimize the loss function to seek the optimal weight parameters of the DNN model, which is formulated by

$$\arg\min_{\theta} \mathcal{L}(f(x,\theta),y). \tag{2}$$

By minimizing the loss function in th training stage, the optimal weight parameters θ^* are obtained, which will be employed for human activity recognition in the online stage using new Wi-Fi CSI amplitudes.

To validate the effect of adversarial attacks on DNN based human activity recognition, the specific DNN model is used in the proposed system. The table I summarizes the used DNN model. First, the DNN model uses different layers including input layer, bidirectional layer, attention layer, and three dense layers. The last dense layer employs Solftmax function to classify seven different activities. We also employ Adam as the optimizer for the DNN model, using a batch size of 128 and the number of epochs of 50.

Specifically, before three dense layers, we use BiLSTM and attention blocks in the DNN model. Compared with LSTM that only processes CSI amplitudes in one direction (past information), BiLSTM can consider the past and future information, which includes a forward layer and a backward layer. In addition, the use of attention model can focus on the interest CSI signal parts, and obscure the rest for human activity classification [4]. Specifically, the attention model can automatically study the importance of features and time steps, where larger weights are assigned to more important features and time steps. In summary, BiLSTM can learn the sequential features as the input of the attention model (i.e., the self-attention), which can effectively improve the accuracy of human activity recognition.

TABLE I
DNN MODEL

Model	Type of Layer	Output Shape	Loss Function
DNN	Input Layer	500, 90	Categorical Cross-entropy
	Bidirectional Layer	500, 20	
	Attention Layer	20	
	Dense Layer	64	
	Dense Layer	32	
	Dense Layer	16	
	Dense Layer (Softmax)	7	

IV. Adversarial Attack Models

In this section, We will discuss the adversarial attack problem for DNN-based human activity recognition, and then introduce three white-box adversarial attack methods (e.g., FGSM, PGD, and MIM) to validate the robustness performance of the proposed system.

A. Problem Formulation

The widely used DNN model can be misled by adversarial examples by adding a small perturbation to the new Wi-Fi CSI amplitude. Generally, the objective of the adversary is to destroy the performance of the DNN model by maximizing the loss function, which is formulated by

$$\arg\max_{x_{adv}} \mathcal{L}(f(x_{adv},\theta^*),y), \tag{3}$$

where x_{adv} is the adversarial example. The adversarial example x_{adv} can be obtained by $x_{adv} = x + \eta$, where η is the perturbation. Traditionally, when the trained DNN model f with parameter θ^* can be accessed, we use a box-constrained optimization problem (e.g., L-BFGS attack needs to use a binary search to find the optimal parameter value) [7] to generate an adversarial example x_{adv}. However, it will have high time complexity, which becomes impractical in real-world applications. Therefore, we consider the one-step attack method (i.e., FGSM) and two iterative attack methods (PGD and MIM) in this paper, which are discussed as the follows.

B. Fast Gradient Sign Method

The FGSM attack method is an effective one-step attack to reduce the time complexity, compared with L-BFGS attack. Based on the given input, the FGSM attack method can generate the perturbation η by calculating the gradient of the loss function [8], which is defined by

$$\eta = \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}(f(x,\theta^*),y)), \tag{4}$$

where ϵ is a hyper-parameter to adjust the magnitude of the perturbation. Given the loss function $\mathcal{L}$, we can obtain the perturbation η by calculating the first derivative of $\mathcal{L}(f(x,\theta^*),y)$ using the backpropagation algorithm. The generalization of FGSM is called the Fast Gradient Method (FGM) [19], where the perturbation o FGM is formulated by

$$\eta = \epsilon \cdot \frac{\nabla_x \mathcal{L}(f(x,\theta^*),y)}{\|\nabla_x \mathcal{L}(f(x,\theta^*),y)\|_2}. \tag{5}$$

Based on FGM method (5), we can conveniently generate the perturbation.

C. Projected Gradient Descent Attack

Based on the one-step method (e.g. FGM), PGD attack was proposed using an iterative version of FGM to enhance the attack performance [9]. The PGD attack method could improve the robustness of the DNN model against first-order attacks methods (e.g., FGM). Based on PGD method (i.e. an iterative method), we could generate the adversarial examples by

$$x_0^{adv} = x, \tag{6}$$

$$x_{N+1}^{adv} = \text{Clip}_{x,\epsilon}\left\{ x_N^{adv} + \alpha \cdot \frac{\nabla_x \mathcal{L}(f(x_N^{adv}, \theta^*), y)}{\left\| \nabla_x \mathcal{L}(f(x_N^{adv}, \theta^*), y) \right\|_2} \right\}, \tag{7}$$

where α a hyper-parameter in each iteration, which can be set to ϵ/N, if the ϵ parameter is provided. The generated small perturbation is around the original input x in the L^p ball. In addition, $\text{Clip}_{x,\epsilon}$ could project the perturbation back into the L^p ball. The PGD method is a stronger adversarial attack method, compared with one-step FGM/FGSM methods.

D. Momentum Iterative Method

Because PGD generates the adversarial examples greedily along the direction of the gradient in each iteration, the local maxima could be obtained easily, thus leading to the poor transferability. To address this problem, the momentum iterative method is used, which can leverage the gradient of the previous iterations to help update the perturbation. Based on the MIM method, the gradient is obtained by

$$g_{(N+1)} = \mu \cdot g_N + \frac{\nabla_x \mathcal{L}(f(x_N^{adv}, \theta^*), y)}{\left\| \nabla_x \mathcal{L}(f(x_N^{adv}, \theta^*), y) \right\|_2}, \tag{8}$$

where g_N includes the gradients from previous $N-1$ iterations with a decay factor μ. Then, we can generate the adversarial examples using the following equation,

$$x_{(N+1)}^{adv} = x_N^{adv} + \alpha \cdot \text{sign}(g_{(N+1)}), \tag{9}$$

where α could be set to ϵ/N when ϵ is given.

V. Experimentation and Results

A. Experiment Configuration

We use a public Wi-Fi CSI dataset for human activity recognition, which are collected in indoor environments [3]. A commodity Wi-Fi router and a laptop are used as the transmitter and the receiver, both of which have the Intel 5300 NIC. Also, the transmitter and the receiver are located three meters apart in line-of-sight (LOS) environment. Then, the receiver has a sampling frequency of 1 kHz, where three antennas are exploited to obtain 90 CSI values over a packet. To implement CSI data segmentation, a sliding window (i.e., 2 seconds) is employed. In addition, six persons collected CSI data with seven common daily activities including "bed", "stand-up", "fall", "pick-up", "run", "sit-down" and "walk", where each person conducted an activity with a period of 20 seconds. Then, the first column in the CSI dataset offers the timestamp; the second column to the 91st column provide CSI amplitude values (90 subcarriers over three antennas) and the remaining columns offer the phase information. In this paper, we only use the CSI amplitude data for human activity recognition. In addition, the data is divided into two subsets: 90% for training and validation, 10% for test.

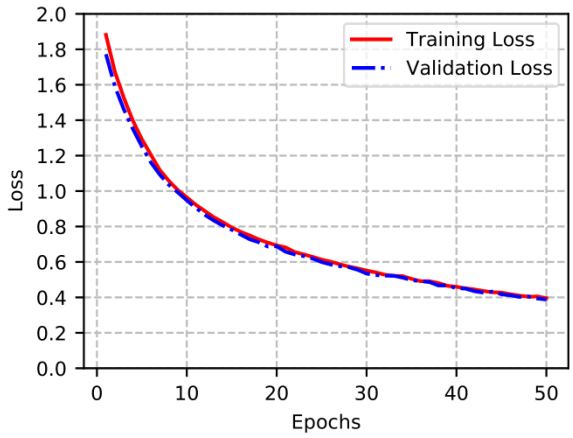

Fig. 2. Training loss vs validation loss.

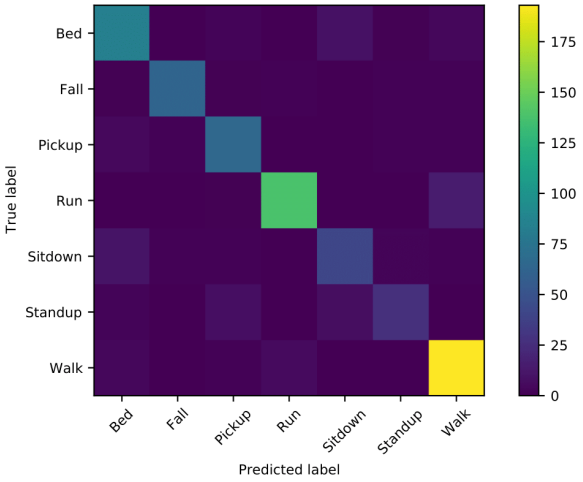

Fig. 3. Confusion matrix of human activity classification with clear test data.

We independently implemented three types of adversarial attacks (i.e., FGSM, PGD, and MIM) for human activity recognition. Three adversarial attacks have been performed in the testing stage over different epsilon values. In all the experiments, we use Python, Tensorflow, Keras, and Cleverhans libraries for training and testing the used model. In addition, Google Colab Pro is exploited as a cloud service to train the DNN model.

In the following section, we will discuss the performance of DNN-based human activity recognition, and validate their performances under three white-box adversarial attacks.

B. Results and Discussions

Fig. 2 shows the loss over different epochs for training and validation of human activity classification model. We use 50 epochs to train the DNN model in the offline stage. We can see

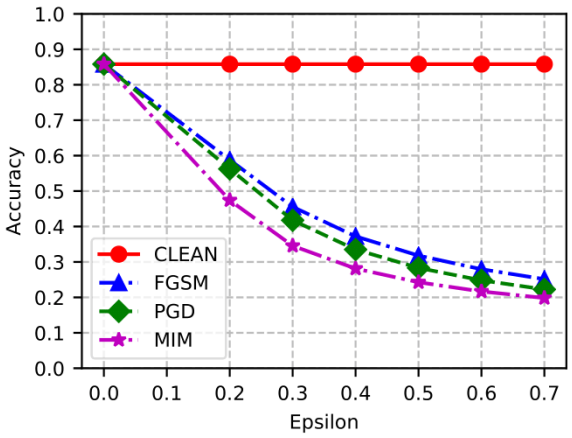

Fig. 4. Classification accuracy under different white-box attacks.

that the loss functions over training and validation decreases with the increase of epochs. Then, the loss function curves for training and validation will converge after the number of epochs is about 40. Fig. 3 shows the confusion matrix for the human classification model with clear test data, which can be used to analyze the performance of the classification model. We can see that the classification can predict the walking activity much more accurately because of a large movement from walking that will greatly influence CSI amplitude values. In addition, for other activities, the classification model can also obtain satisfied results.

Fig. 4 shows the accuracy of activity classification over different epsilon values under three white-box attacks (i.e., FGSM attack, PGD attack, and MIM attack). The adversarial examples are obtained by adding a small perturbation under different epsilon values, which can determine the strength of the noise in the original CSI amplitude data. We consider the range of epsilons in [0.00001, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7]. In Fig. 4, we can see that the accuracy for the clear test data (denoted by CLEAN) is about 0.85. However, under the adversarial examples, the used DNN model cannot obtain good results for all attacks. For example, at the epsilon with 0.7, the accuracy for three attacks will drop under 0.3. Moreover, we can notice that MIM and PGD attacks methods with the iterative method are better than the one-step FGSM attack. Therefore, we consider that all three white-box adversarial attacks can mislead the human activity classification model (i.e. the used DNN model).

VI. Conclusion

In this paper, we proposed adversarial machine learning for human activity recognition using Wi-Fi CSI. We discussed the system framework, where activity recognition can be considered as a classification problem and the attention-based BiLSTM model is introduced. Then, we discussed adversarial attack problem for DNN-based human activity recognition and formulated three white-box attacks (i.e., FGSM, PGD, and MIM). In the experimental part with a public CSI dataset, our results showed that the performances of DNN-based human activity classification are greatly influenced by three white-box adversarial attacks.

Acknowledgments

This work is supported in part by the NSF under Grants ECCS-1923163 CNS-2105416.

References

[1] J. Liu, H. Liu, Y. Chen, Y. Wang, and C. Wang, "Wireless sensing for human activity: A survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1629–1645, 2019.

[2] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Understanding and modeling of WiFi signal based human activity recognition," in *Proc. ACM MobiCom'15*, Paris, France, Sept. 2015, pp. 65–76.

[3] S. Yousefi, H. Narui, S. Dayal, S. Ermon, and S. Valaee, "A survey on behavior recognition using WiFi channel state information," *IEEE Commun. Mag.*, vol. 55, no. 10, pp. 98–104, Oct. 2017.

[4] 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. 2018.

[5] W. Jiang *et al.*, "Towards environment independent device free human activity recognition," in *Proc. ACM MobiCom'18*, New Delhi, India, Oct. 2018, pp. 289–304.

[6] D. Wang, J. Yang, W. Cui, L. Xie, and S. Sun, "Multimodal CSI-based human activity recognition using GANs," *IEEE Internet of Things Journal*, May 2021, in press.

[7] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, Dec. 2013.

[8] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, Dec. 2014.

[9] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, June 2017.

[10] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proc. IEEE CVPR'18*, Salt Lake City, UT, June 2018, pp. 9185–9193.

[11] R. K. Sah and H. Ghasemzadeh, "Adar: Adversarial activity recognition in wearables," in *Proc. IEEE/ACM ICCAD'19*, Westminster, CO, Nov. 2019, pp. 1–8.

[12] Z. Yang, Y. Zhao, and W. Yan, "Adversarial vulnerability in doppler-based human activity recognition," in *Proc. Int. Joint Conf. Neural Networks (IJCNN'20)*, Glasgow, UK, July 2020, pp. 1–7.

[13] Y. Lin, H. Zhao, Y. Tu, S. Mao, and Z. Dou, "Threats of adversarial attacks in DNN-based modulation recognition," in *Proc. IEEE INFOCOM'19*, Toronto, Canada, July 2020, pp. 1–10.

[14] M. Patil, X. Wang, X. Wang, and S. Mao, "Adversarial attacks on deep learning-based floor classification and indoor localization," in *Proc. 3rd ACM Workshop on Wireless Security and Machine Learning*, Virtual Conference, June 2021, pp. 7–12.

[15] D. Halperin, W. J. Hu, A. Sheth, and D. Wetherall, "Predictable 802.11 packet delivery from wireless channel measurements," in *Proc. ACM SIGCOMM'10*, New Delhi, India, Sept. 2010, pp. 159–170.

[16] X. Wang, L. Gao, S. Mao, and S. Pandey, "BiLoc: Bi-modal deep learning for indoor localization with commodity 5GHz WiFi," *IEEE Access J.*, vol. 5, pp. 4209–4220, Mar. 2017.

[17] X. Wang, C. Yang, and S. Mao, "TensorBeat: Tensor decomposition for monitoring multi-person breathing beats with commodity WiFi," *ACM Transactions on Intelligent Systems and Technology*, vol. 9, no. 1, pp. 8:1–8:27, Sept. 2017.

[18] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proc. 2017 ACM Asia Conf. Computer Commun. Security*, Abu Dhabi, UAE, Apr. 2017, pp. 506–519.

[19] T. Miyato, A. M. Dai, and I. Goodfellow, "Adversarial training methods for semi-supervised text classification," *arXiv preprint arXiv:1605.07725*, May 2016.