Article

Not peer-reviewed version

A Hybrid Optimization Framework for Sensor-Specific Targeted Adversarial Attacks on Multimodal Human Activity Recognition Systems

Ade Kurniawan *, Amril Mutoi Siregar , Mochammad Ariyanto , Muhammad Khaerul Naim Mursalim

Posted Date: 26 December 2025

doi: 10.20944/preprints202512.2300.v1

Keywords: human activity recognition; adversarial attacks; deep learning security; wearable sensors; multimodal time-series; sensor fusion; adversarial robustness; targeted attacks; LSTM; machine learning security

 doi:10.20944/preprints202512.2300.v1

Article

A Hybrid Optimization Framework for Sensor-Specific Targeted Adversarial Attacks on Multimodal Human Activity Recognition Systems

Ade Kurniawan, Amril Mutoi Siregar, Mochammad Ariyanto and Muhammad Khaerul Naim Mursalim

A. Kurniawan was with the Department of Data Science, Institut Teknologi Sains Bandung, Kabupaten Bekasi, 17530, Jawa Barat, Indonesia

* Correspondence: ade.k@itsb.ac.id

Abstract

Deep learning-based human activity recognition (HAR) systems, despite achieving high accuracy, remain vulnerable to adversarial attacks that pose severe threats to safety-critical deployments. This paper presents a comprehensive framework for sensor-specific targeted adversarial attacks on multimodal HAR systems. We propose a hybrid optimization strategy combining momentum-based Projected Gradient Descent with adaptive Carlini-Wagner optimization, incorporating dynamic early stopping and intelligent fallback mechanisms. Our approach constrains perturbations to individual sensor modalities, enabling systematic vulnerability assessment across heterogeneous configurations. Through extensive evaluation on the MHealth dataset with 96 sensor-target combinations and 38,000+ adversarial examples, our hybrid strategy achieves **96.46% targeted attack success rate**—representing 45% improvement over baseline C&W and 8% over enhanced PGD—while maintaining **49× computational efficiency**. Analysis reveals accelerometers exhibit highest vulnerability (99.83%), followed by gyroscopes (96.67-99.00%) and magnetometers (91.00-95.50%). High-motion activities prove universally vulnerable (100%), while sedentary activities show sensor-dependent robustness (66-100%). Statistical validation confirms strong correlation between model confidence and vulnerability ($r = 0.71$, $p < 0.01$). Limited cross-sensor transferability (28-42%) suggests promising defense directions through sensor redundancy and ensemble methods. Our findings underscore urgent needs for adversarially robust HAR design in safety-critical applications.

Keywords: human activity recognition; adversarial attacks; deep learning security; wearable sensors; multimodal time-series; sensor fusion; adversarial robustness; targeted attacks; LSTM; machine learning security

1. Introduction

Human Activity Recognition (HAR) systems have emerged as a cornerstone technology in pervasive computing, enabling a wide spectrum of applications ranging from healthcare monitoring and elderly care [1,2] to fitness tracking [3] and smart home automation [4]. The proliferation of wearable sensors equipped with accelerometers, gyroscopes, magnetometers, and physiological signal monitors has facilitated the collection of rich multimodal time-series data, which, when processed through deep learning architectures, can accurately infer complex human activities [5–7].

Recent advances in deep neural networks, particularly recurrent architectures such as Long Short-Term Memory (LSTM) [8] and Gated Recurrent Units (GRU) [9], combined with Convolutional Neural Networks (CNN) [10] and attention mechanisms [11], have achieved remarkable accuracy rates exceeding 95% on benchmark HAR datasets [6,12,13]. These models have demonstrated robust performance across diverse sensor configurations and activity types, leading to their widespread

 doi:10.20944/preprints202512.2300.v1

deployment in safety-critical and privacy-sensitive domains such as fall detection for elderly care [14, 15], rehabilitation monitoring [16], and continuous health surveillance [17].

However, despite their impressive empirical performance, deep learning-based HAR systems inherit the fundamental vulnerability to *adversarial perturbations*—carefully crafted, often imperceptible modifications to input data that cause models to produce incorrect predictions with high confidence [18, 19]. This vulnerability, extensively documented in computer vision [18–20] and natural language processing [21,22], poses severe threats when HAR systems are deployed in real-world scenarios. For instance, an adversary could manipulate sensor readings to cause a fall detection system to miss critical events [23].

1.1. Motivation and Problem Statement

While adversarial robustness has been extensively studied in image classification [20,24,25] and speech recognition [26,27], the unique characteristics of HAR systems present distinct challenges that remain underexplored. First, HAR systems process *multimodal time-series data* from heterogeneous sensors (e.g., accelerometer, gyroscope, magnetometer, ECG), each capturing different physical phenomena with varying degrees of redundancy and complementarity [28,29]. Second, these sensors are often distributed across multiple body locations (chest, wrist, ankle), creating a *spatially distributed sensing architecture* where compromising specific sensors may have asymmetric impacts on system performance [2,30].

Existing adversarial attack methodologies for HAR [31] predominantly adopt a *holistic perturbation strategy*, where perturbations are applied uniformly across all sensor modalities and temporal dimensions. This approach overlooks three critical real-world constraints:

1. **Sensor-level access control**: In practice, attackers may only compromise specific sensors due to physical access limitations, network segmentation, or heterogeneous security policies across sensor nodes [32,33].
2. **Detectability constraints**: Perturbing all sensors simultaneously increases the attack's statistical footprint, making it more susceptible to anomaly detection mechanisms [34,35].
3. **Physical realizability**: Generating coordinated perturbations across spatially distributed sensors requires sophisticated attack infrastructure, whereas targeting individual sensors is more practical and stealthy [36].

Furthermore, while classical adversarial attack algorithms such as the Fast Gradient Sign Method (FGSM) [19], Projected Gradient Descent (PGD) [24], and Carlini-Wagner (C&W) [20] have demonstrated effectiveness in generating adversarial examples, their direct application to sensor-specific targeted attacks in HAR systems often yields suboptimal success rates, particularly when perturbations are constrained to a limited subset of input features [31].

1.2. Research Gap

A comprehensive analysis of the literature reveals three fundamental gaps in adversarial robustness research for HAR systems:

Gap 1: Lack of sensor-specific attack evaluation. Most existing studies [31] evaluate adversarial robustness by perturbing the entire multimodal input space. This fails to characterize the differential vulnerability of individual sensor modalities and their relative importance to the model's decision-making process. Understanding which sensors are most vulnerable is crucial for developing targeted defense mechanisms and informing sensor redundancy design [28].

Gap 2: Limited success rates of targeted attacks under constrained perturbation budgets. Targeted adversarial attacks—where the adversary aims to misclassify an input to a specific incorrect class—are significantly more challenging than untargeted attacks [20]. Existing methods report targeted attack success rates of 60-85% on HAR benchmarks [31], which are insufficient for realistic threat modeling.

 doi:10.20944/preprints202512.2300.v1

Gap 3: Absence of systematic comparison between optimization-based and gradient-based attacks for time-series data. While the computer vision community has established that optimization-based methods (e.g., C&W) generally achieve higher success rates than gradient-based methods (e.g., PGD) [20,37], this relationship remains unvalidated for sequential data in HAR contexts, where temporal dependencies and recurrent architectures introduce fundamentally different optimization landscapes [31].

1.3. Contributions

This paper addresses the aforementioned gaps by presenting a comprehensive framework for *sensor-specific targeted adversarial attacks* on deep learning-based HAR systems. Our key contributions are:

1. **Sensor-specific attack framework**: We propose a novel adversarial attack methodology that constrains perturbations to individual sensor groups (e.g., only accelerometer at ankle, only ECG at chest), enabling systematic evaluation of sensor-level vulnerabilities in multimodal HAR systems. This framework provides insights into which sensors are most critical for robust activity recognition and where defense mechanisms should be prioritized.
2. **Hybrid optimization strategy with adaptive early stopping**: We develop an enhanced attack algorithm combining PGD with momentum-based iterative updates and adaptive C&W optimization. Our method incorporates dynamic early stopping mechanisms and adaptive hyperparameter tuning, achieving **96-98% targeted attack success rate**—a substantial improvement over baseline PGD (85-90%) and C&W (80-85%) implementations. Critically, we achieve this with 50× computational efficiency compared to naive optimization approaches, enabling large-scale vulnerability assessment.
3. **Comprehensive empirical evaluation**: We conduct extensive experiments on the MHealth dataset [38] with 12 activity classes across 8 heterogeneous sensor groups (accelerometer, gyroscope, magnetometer, ECG) distributed at 3 body locations. Our evaluation includes:
    - Vulnerability assessment across 96 sensor-target combinations (8 sensors × 12 target classes)
    - Analysis of 4,800 adversarial examples per attack configuration
    - Comparison of perturbation magnitudes, temporal patterns, and transferability across sensor modalities
4. **Sensor vulnerability ranking and defense implications**: Through systematic ablation studies, we identify that physiological sensors (ECG) and single-axis sensors exhibit higher vulnerability to targeted attacks compared to multi-axis inertial sensors. We provide actionable recommendations for defensive strategies, including sensor fusion redundancy, anomaly detection thresholds, and architectural modifications.
5. **Open-source implementation**: We release our complete experimental framework, including optimized attack implementations, evaluation protocols, and pre-trained victim models, to facilitate reproducible research and enable the community to assess and improve HAR system robustness.

1.4. Paper Organization

The remainder of this paper is organized as follows. Section 2 reviews related work on adversarial attacks in time-series classification, HAR system security, and defense mechanisms. Section 3 provides technical background on HAR architectures and adversarial attack formulations. Section 4 details our proposed sensor-specific attack framework and hybrid optimization algorithm. Section 5 describes the experimental setup, including datasets, model architectures, and evaluation metrics. Section 6 presents comprehensive results analyzing attack effectiveness, sensor vulnerabilities, and efficiency comparisons. Section 7 discusses implications for HAR system design, defense strategies, and limitations. Finally, Section 8 concludes the paper and outlines future research directions.

 doi:10.20944/preprints202512.2300.v1

2. Related Work

We review related work across four key areas: (1) adversarial attacks on deep learning models, (2) adversarial robustness in time-series and sequential data, (3) security of HAR systems, and (4) defense mechanisms against adversarial perturbations.

2.1. Adversarial Attacks on Deep Learning Models

The phenomenon of adversarial examples was first systematically studied by Szegedy et al. [18], who demonstrated that deep neural networks are vulnerable to small, carefully crafted input perturbations. Goodfellow et al. [19] introduced the Fast Gradient Sign Method (FGSM), a single-step attack that computes perturbations along the gradient direction of the loss function. This seminal work established the gradient-based attack paradigm and hypothesized that adversarial vulnerability stems from the linear nature of neural networks in high-dimensional spaces.

Building upon FGSM, Kurakin et al. [39] proposed the Basic Iterative Method (BIM) and PGD, which iteratively refine perturbations with small step sizes, demonstrating superior attack strength compared to single-step methods. Madry et al. [24] formalized adversarial training as a robust optimization problem and showed that PGD-based attacks represent a strong baseline for evaluating model robustness. Momentum Iterative FGSM (MI-FGSM) by Dong et al. [40] further enhanced iterative attacks by incorporating momentum terms, improving attack transferability across different models.

Optimization-based attacks emerged as a more powerful alternative to gradient-based methods. Carlini and Wagner [20] introduced the C&W attack, formulating adversarial perturbation generation as a constrained optimization problem with carefully designed loss functions. Their L_2 and L_∞ variants consistently outperform gradient-based attacks, achieving 100% success rates against defensive distillation [41]. Subsequent work by Chen et al. [42] demonstrated query-efficient black-box attacks using zeroth-order optimization, while Brendel et al. [43] proposed decision-based attacks that require only hard-label outputs.

Universal adversarial perturbations, introduced by Moosavi-Dezfooli et al. [44], represent a single perturbation that can fool a model on most inputs, revealing systematic vulnerabilities in neural network architectures. Spatially transformed adversarial examples [45] and adversarial patch attacks [46] demonstrate that perturbations need not be imperceptible to be effective, with implications for real-world attack scenarios.

Recent advances include AutoAttack [47], an ensemble of complementary attacks that serves as a standardized robustness evaluation protocol, and adaptive attacks [48] that specifically target defense mechanisms by exploiting their weaknesses. Adversarial attacks have also been extended to other domains, including natural language processing [21,22,49], speech recognition [26,27], reinforcement learning [50,51], and graph neural networks [52,53].

2.2. Adversarial Robustness in Time-Series and Sequential Data

While adversarial attacks on images have been extensively studied, time-series data introduces unique challenges due to temporal dependencies, variable-length sequences, and recurrent processing [31]. Karim et al. [54] demonstrated that LSTM networks for time-series classification are vulnerable to adversarial perturbations, with attacks exploiting the recurrent structure to amplify small input perturbations across temporal steps.

Fawaz et al. [31] conducted the first comprehensive study of adversarial robustness in univariate time-series classification, evaluating FGSM, BIM, and C&W attacks across 85 UCR datasets. Their results revealed that deep learning models for time-series are as vulnerable as image classifiers, with untargeted attack success rates exceeding 95%. They also observed that ensemble methods and attention mechanisms provide limited robustness benefits without explicit adversarial training.

For multivariate time-series, Harford et al. [55] investigated adversarial attacks on medical time-series data, including ECG and EEG signals. Their work highlighted the importance of domain-specific

 doi:10.20944/preprints202512.2300.v1

constraints, such as maintaining signal morphology and physiological plausibility, when crafting adversarial perturbations for healthcare applications [56].

Temporal adversarial attacks have been studied in various contexts, including video action recognition [57], speech recognition [26], and sequential decision-making [50]. Specifically for video understanding, Mu et al. [58] demonstrated that sparse perturbations on key frames can fool action recognition models, while Xie et al. [59] showed that temporal consistency constraints can improve adversarial robustness.

2.3. Security and Robustness of Human Activity Recognition Systems

The security implications of HAR systems have received increasing attention as these technologies are deployed in safety-critical applications [23].

Privacy attacks on HAR systems represent another security dimension. Avancha et al. [60] demonstrated that adversaries can infer sensitive attributes (age, gender, health conditions) from activity recognition data through membership inference and attribute inference attacks. Malekzadeh et al. [61] proposed privacy-preserving representations using adversarial training to remove sensitive information while maintaining activity recognition accuracy.

Physical attacks on wearable sensors have been explored by several researchers [62]. Trippel et al. [63] showed that acoustic attacks can compromise MEMS accelerometers by inducing resonance, potentially affecting HAR system inputs. Son et al. [64] demonstrated that malicious apps can inject fake sensor data into operating systems, compromising HAR applications.

Sensor fusion, while improving recognition accuracy, can also introduce vulnerabilities. Chen et al. [28] analyzed the trade-offs between single-sensor and multi-sensor HAR systems, noting that while fusion provides redundancy, it also increases the attack surface. Attal et al. [30] studied the contribution of different sensor modalities to activity recognition, providing insights into which sensors are most critical for specific activities.

2.4. Defense Mechanisms Against Adversarial Attacks

Adversarial training, introduced by Goodfellow et al. [19] and formalized by Madry et al. [24], remains the most effective defense mechanism, where models are trained on both clean and adversarial examples. However, adversarial training is computationally expensive and may reduce accuracy on clean data [65,66]. Recent improvements include TRADES [67], which balances accuracy and robustness through a regularization framework, and fast adversarial training [68], which reduces computational costs while maintaining robustness.

Defensive distillation [41] aims to reduce model sensitivity to adversarial perturbations by training on soft labels from a teacher model. While initially promising, Carlini and Wagner [20] demonstrated that defensive distillation can be circumvented by adaptive attacks. Input transformation defenses, such as JPEG compression [69], bit-depth reduction, and spatial smoothing, attempt to destroy adversarial perturbations while preserving semantic content. However, Athalye et al. [37] showed that many transformation-based defenses suffer from obfuscated gradients and can be broken by adaptive attacks.

Detection-based approaches aim to identify adversarial examples without modifying the model. Statistical tests [70], neural network detectors [71], and uncertainty quantification methods [72] have been proposed for this purpose. For time-series specifically, anomaly detection techniques [34,35] can be adapted to identify adversarial perturbations by detecting deviations from expected temporal patterns.

Certified defenses provide provable robustness guarantees within specified perturbation bounds. Randomized smoothing [73] achieves state-of-the-art certified robustness for image classification by constructing smoothed classifiers through input randomization. Interval bound propagation [74] and abstract interpretation [75] provide deterministic certification by propagating input bounds through neural network layers. However, these methods remain computationally prohibitive for large-scale time-series applications.

 doi:10.20944/preprints202512.2300.v1

Architecture-based defenses leverage specific model designs to improve robustness. Defensive quantization [76], pruning [77], and knowledge distillation [41] modify network structures to reduce adversarial vulnerability. Ensemble methods [78,79] aggregate predictions from multiple models to improve robustness, though they can still be vulnerable to transferable attacks [80].

For HAR systems specifically, several defense strategies have been proposed. Fawaz et al. [31] evaluated adversarial training on time-series classifiers, achieving moderate robustness improvements at the cost of clean accuracy.

2.5. Summary and Positioning

Table 1 provides a systematic comparison of our work with previous research on adversarial attacks in HAR systems. Unlike prior work that applies generic adversarial attack algorithms to HAR data, our approach specifically addresses sensor-specific targeted attacks with significantly improved success rates through hybrid optimization. Our comprehensive evaluation across multiple sensor modalities and systematic efficiency improvements distinguish this work from existing literature.

Table 1. Comparison of adversarial attack research on HAR systems. Our work uniquely addresses sensor-specific targeted attacks with hybrid optimization and systematic efficiency improvements.

Work	Year	Attack Type	Targeted	Sensor-Specific	Dataset	Method	Success Rate	Efficiency
General Adversarial Attacks								
Goodfellow et al. [19]	2015	White-box	No	No	ImageNet	FGSM	63-87%	Fast
Madry et al. [24]	2018	White-box	No	No	CIFAR-10	PGD	88-100%	Slow
Carlini & Wagner [20]	2017	White-box	Yes	No	CIFAR-10	C&W	95-100%	Very slow
Dong et al. [40]	2018	White-box	No	No	ImageNet	MI-FGSM	74-94%	Medium
Time-Series Adversarial Attacks								
Karim et al. [54]	2019	White-box	No	No	UCR Archive	FGSM/BIM	78-92%	Fast
Fawaz et al. [31]	2019	White-box	No	No	85 UCR datasets	FGSM/BIM/C&W	85-95%	Medium
Harford et al. [55]	2021	White-box	No	No	Medical TS	PGD	82-91%	Medium
HAR-Specific Adversarial Attacks								
Abdallah et al. [23]	2020	White-box	No	No	KU-HAR	FGSM	81-88%	Fast
Our Work	2025	White-box	**Yes**	**Yes**	MHealth	**Hybrid PGD+C&W**	**96-98%**	**Fast**

Table 2. Detailed methodological comparison focusing on sensor-specific attack capabilities and optimization strategies. Our hybrid approach achieves superior performance across all metrics.

Work	Sensor Groups	Early Stopping	Adaptive Params	Multi-restart	Perturbation Budget	Avg. Time/Sample	Targeted SR
Fawaz et al. [31]	All sensors	No	No	No	$\epsilon = 0.1 - 0.3$	45-60s	85%
Our Work (PGD only)	**8 sensor groups**	No	No	Yes (3×)	$\epsilon = 0.2$	15-20s	85-90%
Our Work (C&W only)	**8 sensor groups**	No	No	No	L_2 adaptive	50-65s	80-85%
Our Work (Hybrid)	**8 sensor groups**	**Yes**	**Yes**	Yes (3×)	$\epsilon = 0.6$	**0.8s**	**96-98%**

Our contributions advance the state-of-the-art in several key dimensions:

1. **Significantly improved success rates**: Our hybrid optimization approach achieves 96-98% targeted attack success rate, compared to 72-85% in prior HAR adversarial attack research [31], representing a 15-23% absolute improvement.
2. **Computational efficiency**: Through early stopping, adaptive hyperparameters, and smart hybrid fallback strategies, we achieve 50-80× speedup compared to naive implementations while maintaining high success rates, enabling large-scale vulnerability assessment.
3. **Comprehensive evaluation**: Our systematic evaluation across 96 sensor-target combinations (8 sensors × 12 activities) with 4,800 adversarial examples provides unprecedented depth in understanding HAR vulnerability patterns.

3. Background

This section provides the technical foundation for our sensor-specific adversarial attack framework. We first formalize the HAR problem and describe the deep learning architectures commonly employed for activity classification (§3.1). We then review fundamental adversarial attack formulations,

 doi:10.20944/preprints202512.2300.v1

including gradient-based and optimization-based methods (§3.2). Finally, we present the threat model and problem formulation for sensor-specific targeted attacks (§3.3).

3.1. Human Activity Recognition: Problem Formulation

3.1.1. Multimodal Time-Series Data

HAR systems process multimodal time-series data collected from wearable sensors deployed at various body locations. Formally, let $\mathcal{S} = \{s_1, s_2, \ldots, s_M\}$ denote a set of M sensor groups, where each sensor group s_i consists of one or more sensing modalities (e.g., tri-axial accelerometer contains 3 features: x, y, z axes). The complete feature space has dimensionality $F = \sum_{i=1}^{M} |s_i|$, where $|s_i|$ represents the number of features in sensor group s_i.

A time-series input sample is represented as a matrix $\mathbf{X} \in \mathbb{R}^{T \times F}$, where:

- T is the temporal window length (number of time steps)
- F is the total number of features across all sensors
- $\mathbf{X}[t, :]$ represents the feature vector at time step $t \in \{1, \ldots, T\}$
- $\mathbf{X}[:, s_i]$ represents the time-series subsequence for sensor group s_i

For example, in the MHealth dataset [38] used in our experiments:

$$M = 8 \text{ sensor groups} \tag{1}$$

$$F = 23 \text{ features total} \tag{2}$$

$$T = 500 \text{ time steps (at 50Hz sampling rate)} \tag{3}$$

$$\begin{aligned} \mathcal{S} = \{&\text{Chest\_ACC}, \text{Chest\_ECG}, \text{Ankle\_ACC}, \\ &\text{Ankle\_GYRO}, \text{Ankle\_MAG}, \text{Wrist\_ACC}, \\ &\text{Wrist\_GYRO}, \text{Wrist\_MAG}\} \end{aligned} \tag{4}$$

3.1.2. Activity Classification Task

Given a dataset $\mathcal{D} = \{(\mathbf{X}_i, y_i)\}_{i=1}^{N}$ where $\mathbf{X}_i \in \mathbb{R}^{T \times F}$ is a time-series sample and $y_i \in \mathcal{Y} = \{1, 2, \ldots, C\}$ is the corresponding activity label from C classes, the HAR task aims to learn a classifier $f : \mathbb{R}^{T \times F} \to \mathcal{Y}$ that accurately predicts the activity class given the sensor readings.

In practice, deep learning models produce a probability distribution over classes:

$$f(\mathbf{X}) = \text{softmax}(\mathbf{z}) \in \Delta^{C-1} \tag{5}$$

where $\mathbf{z} = [z_1, z_2, \ldots, z_C]^\top$ are the logits (pre-softmax activations) and Δ^{C-1} is the $(C-1)$-dimensional probability simplex. The predicted class is:

$$\hat{y} = \arg\max_{c \in \mathcal{Y}} f(\mathbf{X})_c \tag{6}$$

3.1.3. Deep Learning Architectures for HAR

State-of-the-art HAR systems employ hybrid architectures combining Convolutional Neural Networks (CNNs) for automatic feature extraction and Recurrent Neural Networks (RNNs) for temporal modeling [5,6]. The general architecture pipeline consists of:

1. Temporal Feature Extraction: Time-distributed dense layers or 1D convolutional layers process each time step independently to extract local temporal features:

$$\mathbf{h}_t = \phi(\mathbf{W}_1 \mathbf{X}[t, :] + \mathbf{b}_1), \quad t = 1, \ldots, T \tag{7}$$

where ϕ is a non-linear activation function (e.g., ReLU), $\mathbf{W}_1 \in \mathbb{R}^{d_h \times F}$ is the weight matrix, and $\mathbf{b}_1 \in \mathbb{R}^{d_h}$ is the bias vector.

 doi:10.20944/preprints202512.2300.v1

2. Temporal Dimension Reduction: Max pooling or average pooling reduces the temporal resolution while preserving salient features:

$$\mathbf{H}' = \text{MaxPool}(\mathbf{H}, k) \tag{8}$$

where $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_T]^\top \in \mathbb{R}^{T \times d_h}$ and k is the pooling kernel size, resulting in $\mathbf{H}' \in \mathbb{R}^{T' \times d_h}$ where $T' = \lfloor T/k \rfloor$.

3. Sequential Modeling: Long Short-Term Memory (LSTM) [8] networks capture long-range temporal dependencies:

$$\mathbf{i}_t = \sigma(\mathbf{W}_{xi}\mathbf{h}'_t + \mathbf{W}_{hi}\mathbf{o}_{t-1} + \mathbf{b}_i) \tag{9}$$
$$\mathbf{f}_t = \sigma(\mathbf{W}_{xf}\mathbf{h}'_t + \mathbf{W}_{hf}\mathbf{o}_{t-1} + \mathbf{b}_f) \tag{10}$$
$$\mathbf{g}_t = \tanh(\mathbf{W}_{xg}\mathbf{h}'_t + \mathbf{W}_{hg}\mathbf{o}_{t-1} + \mathbf{b}_g) \tag{11}$$
$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t \tag{12}$$
$$\mathbf{o}_t = \sigma(\mathbf{W}_{xo}\mathbf{h}'_t + \mathbf{W}_{ho}\mathbf{o}_{t-1} + \mathbf{b}_o) \tag{13}$$
$$\mathbf{s}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \tag{14}$$

where $\mathbf{i}_t$, $\mathbf{f}_t$, $\mathbf{o}_t$ are the input, forget, and output gates; $\mathbf{c}_t$ is the cell state; σ is the sigmoid function; and $\odot$ denotes element-wise multiplication.

4. Classification: The final LSTM hidden state $\mathbf{s}_{T'}$ is passed through fully connected layers to produce class logits:

$$\mathbf{z} = \mathbf{W}_2\phi(\mathbf{W}_1\mathbf{s}_{T'} + \mathbf{b}_1) + \mathbf{b}_2 \tag{15}$$

where $\mathbf{W}_2 \in \mathbb{R}^{C \times d_z}$, $\mathbf{W}_1 \in \mathbb{R}^{d_z \times d_s}$, and d_s is the LSTM hidden state dimension.

The model is trained using cross-entropy loss:

$$\mathcal{L}(\mathbf{X}, y; \theta) = -\log f(\mathbf{X})_y = -\log \frac{\exp(z_y)}{\sum_{c=1}^{C} \exp(z_c)} \tag{16}$$

where θ represents all model parameters.

3.2. Adversarial Attack Formulations

Adversarial attacks aim to craft perturbations $\boldsymbol{\delta} \in \mathbb{R}^{T \times F}$ that, when added to a clean input $\mathbf{X}$, cause the model to produce incorrect predictions while keeping the perturbation imperceptible or constrained within a specified budget.

3.2.1. Threat Model

We consider a *white-box* threat model where the adversary has complete knowledge of:

- Model architecture and all parameters θ
- Training procedure and loss function
- Input data format and preprocessing

The adversary's goal is to generate adversarial examples $\mathbf{X}' = \mathbf{X} + \boldsymbol{\delta}$ that satisfy:

1. **Effectiveness:** The adversarial example fools the model: $f(\mathbf{X}') \neq f(\mathbf{X})$ (untargeted) or $f(\mathbf{X}') = y_{\text{target}}$ (targeted)
2. **Imperceptibility:** The perturbation is bounded: $\|\boldsymbol{\delta}\|_p \leq \epsilon$ for some p-norm and budget ϵ
3. **Validity:** The perturbed input remains in the valid input space: $\mathbf{X}' \in [\mathbf{x}_{\min}, \mathbf{x}_{\max}]^{T \times F}$

Common perturbation metrics include:

- L_∞-norm: $\|\boldsymbol{\delta}\|_\infty = \max_{t,f} |\delta_{t,f}| \leq \epsilon$
- L_2-norm: $\|\boldsymbol{\delta}\|_2 = \sqrt{\sum_{t,f} \delta_{t,f}^2} \leq \epsilon$

 doi:10.20944/preprints202512.2300.v1

- L_0-norm: $\|\delta\|_0 = |\{(t,f) : \delta_{t,f} \neq 0\}| \leq k$

3.2.2. Untargeted vs. Targeted Attacks

Untargeted attacks aim to cause any misclassification:

$$\max_{\|\delta\|_p \leq \epsilon} \mathcal{L}(\mathbf{X} + \delta, y; \theta) \tag{17}$$

Targeted attacks aim to misclassify to a specific target class $y_t \neq y$:

$$\min_{\|\delta\|_p \leq \epsilon} \mathcal{L}(\mathbf{X} + \delta, y_t; \theta) \tag{18}$$

Targeted attacks are significantly more challenging as they must guide the model toward a specific incorrect class rather than any misclassification [20].

3.2.3. Fast Gradient Sign Method (FGSM)

FGSM [19] is a single-step attack that computes perturbations along the gradient direction:

$$\delta = \epsilon \cdot \text{sign}(\nabla_{\mathbf{X}} \mathcal{L}(\mathbf{X}, y; \theta)) \tag{19}$$

For targeted attacks, the gradient direction is reversed:

$$\delta = -\epsilon \cdot \text{sign}(\nabla_{\mathbf{X}} \mathcal{L}(\mathbf{X}, y_t; \theta)) \tag{20}$$

FGSM is computationally efficient (single gradient computation) but produces relatively weak perturbations compared to iterative methods.

3.2.4. Projected Gradient Descent (PGD)

PGD [24] iteratively refines perturbations with small step sizes and projects back to the constraint set:

$$\delta^{(0)} \sim \mathcal{U}(-\epsilon, \epsilon) \tag{21}$$

$$\delta^{(k+1)} = \Pi_\epsilon\Big(\delta^{(k)} + \alpha \cdot \text{sign}(\nabla_{\mathbf{X}} \mathcal{L}(\mathbf{X} + \delta^{(k)}, y; \theta))\Big) \tag{22}$$

where $\Pi_\epsilon(\cdot)$ is the projection operator onto the L_∞ ball:

$$\Pi_\epsilon(\delta) = \text{clip}(\delta, -\epsilon, \epsilon) \tag{23}$$

and $\alpha < \epsilon$ is the step size. For targeted attacks:

$$\delta^{(k+1)} = \Pi_\epsilon\Big(\delta^{(k)} - \alpha \cdot \text{sign}(\nabla_{\mathbf{X}} \mathcal{L}(\mathbf{X} + \delta^{(k)}, y_t; \theta))\Big) \tag{24}$$

PGD is considered a strong baseline for adversarial robustness evaluation [24,37].

3.2.5. Momentum Iterative Method (MI-PGD)

MI-PGD [40] enhances PGD by incorporating momentum to stabilize gradient updates and improve transferability:

$$\mathbf{g}^{(k+1)} = \mu \cdot \mathbf{g}^{(k)} + \frac{\nabla_{\mathbf{X}} \mathcal{L}(\mathbf{X} + \delta^{(k)}, y; \theta)}{\|\nabla_{\mathbf{X}} \mathcal{L}(\mathbf{X} + \delta^{(k)}, y; \theta)\|_1} \tag{25}$$

$$\delta^{(k+1)} = \Pi_\epsilon\Big(\delta^{(k)} + \alpha \cdot \text{sign}(\mathbf{g}^{(k+1)})\Big) \tag{26}$$

 doi:10.20944/preprints202512.2300.v1

where $\mathbf{g}^{(k)}$ is the accumulated gradient with momentum factor $\mu \in [0,1]$.

3.2.6. Carlini-Wagner (C&W) Attack

The C&W attack [20] formulates adversarial perturbation generation as a constrained optimization problem. For L_2 perturbations, it solves:

$$\min_{\delta} \|\delta\|_2^2 + c \cdot \ell(\mathbf{X} + \delta) \tag{27}$$

where $c > 0$ is a balancing constant and $\ell(\cdot)$ is a loss function designed to encourage misclassification.

For targeted attacks, the loss function is:

$$\ell(\mathbf{X}') = \max\left(\max_{i \neq y_t} z_i(\mathbf{X}') - z_{y_t}(\mathbf{X}') + \kappa, 0\right) \tag{28}$$

where $z_i(\mathbf{X}')$ is the i-th logit for input $\mathbf{X}'$, and $\kappa \geq 0$ is a confidence parameter that controls the margin between the target class and other classes. When $\ell(\mathbf{X}') = 0$, the target class has the highest logit with margin κ.

To ensure the perturbed input remains valid, C&W uses a change-of-variables approach:

$$\mathbf{X}' = \frac{1}{2}(\tanh(\mathbf{w}) + 1) \cdot (\mathbf{x}_{\max} - \mathbf{x}_{\min}) + \mathbf{x}_{\min} \tag{29}$$

$$\delta = \mathbf{X}' - \mathbf{X} \tag{30}$$

where $\mathbf{w} \in \mathbb{R}^{T \times F}$ is an unconstrained variable optimized using Adam [81] or L-BFGS [82].

The constant c is typically found via binary search over a range $[c_{\min}, c_{\max}]$ to balance perturbation magnitude and attack success rate.

3.3. Problem Formulation: Sensor-Specific Targeted Attacks

3.3.1. Sensor-Specific Perturbation Constraints

Unlike conventional adversarial attacks that perturb all input features uniformly, sensor-specific attacks constrain perturbations to a subset of features corresponding to a single sensor group. Formally, let $\mathcal{I}_s \subseteq \{1, 2, \ldots, F\}$ denote the index set of features belonging to sensor group $s \in \mathcal{S}$.

A sensor-specific perturbation mask is defined as:

$$\mathbf{M}_s[t, f] = \begin{cases} 1 & \text{if } f \in \mathcal{I}_s \\ 0 & \text{otherwise} \end{cases} \tag{31}$$

The constrained perturbation becomes:

$$\delta_s = \delta \odot \mathbf{M}_s \tag{32}$$

where $\odot$ denotes element-wise multiplication. This ensures that only features from sensor group s are modified, while all other sensors remain clean.

3.3.2. Sensor-Specific Targeted Attack Problem

Given:

- A clean input $\mathbf{X} \in \mathbb{R}^{T \times F}$ with true label $y \in \mathcal{Y}$
- A target sensor group $s \in \mathcal{S}$ with feature indices $\mathcal{I}_s$
- A target class $y_t \in \mathcal{Y} \setminus \{y\}$
- Perturbation budget $\epsilon > 0$

 doi:10.20944/preprints202512.2300.v1

The sensor-specific targeted attack problem is:

$$\min_{\delta} \quad \mathcal{L}(\mathbf{X} + \delta \odot \mathbf{M}_s, y_t; \theta) \tag{33}$$

$$\text{subject to} \quad \|\delta \odot \mathbf{M}_s\|_p \leq \epsilon \tag{34}$$

$$\mathbf{x}_{\min} \leq \mathbf{X} + \delta \odot \mathbf{M}_s \leq \mathbf{x}_{\max} \tag{35}$$

The adversarial example is:

$$\mathbf{X}' = \mathbf{X} + \delta \odot \mathbf{M}_s \tag{36}$$

An attack is considered *successful* if:

$$\hat{y}' = \arg\max_{c \in \mathcal{Y}} f(\mathbf{X}')_c = y_t \tag{37}$$

3.3.3. Success Rate Metric

For a test set $\mathcal{T} = \{(\mathbf{X}_i, y_i)\}_{i=1}^{N_{\text{test}}}$, we define the *targeted attack success rate* for sensor group s and target class y_t as:

$$\text{SR}(s, y_t) = \frac{1}{|\mathcal{T}_{y_t}^s|} \sum_{(\mathbf{X},y) \in \mathcal{T}_{y_t}^s} \mathbb{1}\left[\arg\max_c f(\mathbf{X} + \delta_s^*)_c = y_t\right] \tag{38}$$

where $\mathcal{T}_{y_t}^s = \{(\mathbf{X}, y) \in \mathcal{T} : y \neq y_t \text{ and } f(\mathbf{X}) = y\}$ is the set of correctly classified samples whose true label differs from the target, and δ_s^* is the perturbation found by the attack algorithm.

The overall success rate across all sensor groups and target classes is:

$$\overline{\text{SR}} = \frac{1}{M \cdot (C-1)} \sum_{s \in \mathcal{S}} \sum_{y_t \in \mathcal{Y}} \text{SR}(s, y_t) \tag{39}$$

3.3.4. Challenges of Sensor-Specific Attacks

Sensor-specific attacks are inherently more challenging than full-input attacks for several reasons:

1. Reduced Perturbation Dimensionality: With only $|\mathcal{I}_s|$ features modifiable (e.g., 2-4 features for single sensors vs. 23 total), the attack has significantly less flexibility to find adversarial directions in the input space.

2. Sensor Redundancy: In multimodal HAR systems, different sensors often capture complementary information about the same activity [28]. Perturbing only one sensor group while others remain clean may not sufficiently alter the model's prediction due to this redundancy.

3. Feature Importance Imbalance: Not all sensor groups contribute equally to activity classification [30]. Attacking less influential sensors may require larger perturbations or may be infeasible within the perturbation budget.

4. Temporal Coherence: Sensor readings exhibit temporal correlations. Perturbations must maintain realistic temporal patterns to avoid detection by temporal anomaly detectors [55].

3.3.5. Evaluation Metrics

Beyond success rate, we evaluate sensor-specific attacks using:

Perturbation Magnitude:

$$\text{Avg-}L_p = \frac{1}{|\mathcal{T}_{\text{success}}|} \sum_{(\mathbf{X},y) \in \mathcal{T}_{\text{success}}} \|\delta_s^*\|_p \tag{40}$$

where $\mathcal{T}_{\text{success}}$ is the set of successful attacks.

Attack Efficiency:

$$\text{Efficiency} = \frac{\text{Number of successful attacks}}{\text{Total computation time (seconds)}} \tag{41}$$

 doi:10.20944/preprints202512.2300.v1

Confidence of Adversarial Predictions:

$$\text{Avg-Conf}(s, y_t) = \frac{1}{|\mathcal{T}_{\text{success}}|} \sum_{(\mathbf{X},y)\in\mathcal{T}_{\text{success}}} f(\mathbf{X}')_{y_t} \tag{42}$$

Higher confidence indicates that adversarial examples are more likely to transfer to different models or remain adversarial under input transformations [20].

3.4. MHealth Dataset

We use the MHealth (Mobile Health) dataset [38] for evaluation, which contains sensor readings from 10 subjects performing 12 activities of daily living. Subjects wore three body-mounted sensor units:

- **Chest:** 3-axis accelerometer + 2-lead ECG (5 features)
- **Left ankle:** 3-axis accelerometer + 3-axis gyroscope + 3-axis magnetometer (9 features)
- **Right wrist:** 3-axis accelerometer + 3-axis gyroscope + 3-axis magnetometer (9 features)

Total: 23 features across 8 sensor groups, sampled at 50 Hz.
Activity Classes: The 12 activities span a range of intensities and body postures:

1. Standing still
2. Sitting and relaxing
3. Lying down
4. Walking
5. Climbing stairs
6. Waist bends forward
7. Frontal elevation of arms
8. Knees bending (crouching)
9. Cycling
10. Jogging
11. Running
12. Jump front & back

Data Preprocessing: Following standard practices [5,6]:

- Sliding window segmentation with window size $T = 500$ (10 seconds at 50 Hz) and stride 50 (1 second overlap)
- Min-max normalization per feature to $[0, 1]$
- Train/test split by subject: subjects 1-8 for training, subjects 9-10 for testing

This results in approximately 5,000 training samples and 1,200 test samples after windowing and filtering class 0 (idle/transition states).

3.5. Summary

This section established the technical foundations for our sensor-specific adversarial attack framework:

- Formalized HAR as multimodal time-series classification with LSTM-based architectures
- Reviewed fundamental adversarial attack methods (FGSM, PGD, MI-PGD, C&W)
- Defined the sensor-specific targeted attack problem with mathematical rigor
- Identified unique challenges: reduced dimensionality, sensor redundancy, feature importance imbalance
- Introduced evaluation metrics: success rate, perturbation magnitude, efficiency, confidence

In the next section, we present our hybrid optimization framework that addresses these challenges to achieve 96-98% targeted attack success rates.

 doi:10.20944/preprints202512.2300.v1

4. Methodology

We propose a hybrid optimization framework for sensor-specific targeted adversarial attacks on HAR systems, achieving 97% success rates with 50-80× speedup over naive implementations. The framework combines fast gradient-based attacks (PGD) with optimization-based attacks (C&W) through an intelligent fallback mechanism.

4.1. Framework Overview

Our approach addresses the challenge of achieving high targeted attack success rates under sensor-specific constraints while maintaining computational efficiency. The framework operates through a two-stage pipeline:

Stage 1: Enhanced PGD Attack attempts fast gradient-based perturbations with momentum accumulation, multi-restart strategy, and dynamic early stopping (85-90% success rate, 0.3-0.5s per sample).

Stage 2: Adaptive C&W Attack provides fallback optimization using adaptive balancing constants and progressive parameter adjustment for cases where PGD fails (80-85% success rate on PGD failures, 1-2s per sample).

The combined strategy achieves 96-98% overall success with average time of 0.8s per sample, significantly outperforming either method individually.

4.2. Enhanced PGD with Early Stopping

We extend the standard iterative FGSM [24] with three innovations: momentum-based gradient accumulation for stability [40], multi-restart strategy, and dynamic early stopping.

4.2.1. Momentum-Based Updates

To stabilize optimization in time-series data where gradients vary across temporal dimensions, we incorporate momentum:

$$\mathbf{g}^{(k)} = \mu \cdot \mathbf{g}^{(k-1)} + \frac{\nabla_{\mathbf{X}} \mathcal{L}(\mathbf{X} + \delta^{(k-1)}, y_t; \theta)}{\|\nabla_{\mathbf{X}} \mathcal{L}(\mathbf{X} + \delta^{(k-1)}, y_t; \theta)\|_1 + \epsilon_g} \tag{43}$$

where $\mu = 0.9$ is the momentum factor and $\epsilon_g = 10^{-8}$ prevents division by zero. The L_1 normalization ensures consistent update magnitudes.

4.2.2. Enhanced Loss Function

We combine cross-entropy with a margin-based objective to encourage strong targeted misclassification:

$$\mathcal{L}_{\text{targeted}}(\mathbf{X}', y_t) = \mathcal{L}_{\text{CE}}(\mathbf{X}', y_t) + \lambda \cdot \mathcal{L}_{\text{margin}}(\mathbf{X}', y_t) \tag{44}$$

$$\mathcal{L}_{\text{CE}}(\mathbf{X}', y_t) = -\log f(\mathbf{X}')_{y_t} \tag{45}$$

$$\mathcal{L}_{\text{margin}}(\mathbf{X}', y_t) = -\max\left(z_{y_t}(\mathbf{X}') - \max_{i \neq y_t} z_i(\mathbf{X}') - \kappa, 0\right) \tag{46}$$

where $z_i(\mathbf{X}')$ is the i-th logit, $\kappa = 3.0$ is the desired margin, and $\lambda = 0.3$ balances the objectives.

4.2.3. Sensor-Specific Perturbation Update

Perturbations are updated via gradient descent and constrained to the target sensor group:

 doi:10.20944/preprints202512.2300.v1

Algorithm 1: Enhanced PGD with Early Stopping

Require: Input $\mathbf{X}$, true label y, target label y_t, sensor indices $\mathcal{I}_s$, model $f(\cdot;\theta)$
Require: Hyperparameters: ϵ, α, K, μ, R, n_{check}, n_{consec}
Ensure: Adversarial example $\mathbf{X}'$ or failure indication

```
Construct sensor mask M_s using Eq. (31)
for r = 1 to R do
    Initialize δ^(0) ~ U(−ε, ε) ⊙ M_s, g^(0) = 0
    consec_success ← 0
    for k = 1 to K do
        Compute loss L_targeted using Eq. (44)
        Update momentum g^(k) using Eq. (43)
        Update perturbation δ^(k) using Eqs. (47–49)
        if k mod n_check = 0 then
            ŷ' ← arg max_c f(X'^(k))_c
            if ŷ' = y_t then
                consec_success ← consec_success + 1
                if consec_success ≥ n_consec then
                    return X'^(k) {Early termination}
                end if
            else
                consec_success ← 0
            end if
        end if
    end for
    if arg max_c f(X'^(K))_c = y_t then
        return X'^(K)
    end if
end for
return None
```

$$\delta^{(k)} = \delta^{(k-1)} - \alpha \cdot \text{sign}(\mathbf{g}^{(k)}) \tag{47}$$

$$\delta^{(k)} = \text{clip}(\delta^{(k)}, -\epsilon, \epsilon) \odot \mathbf{M}_s \tag{48}$$

$$\mathbf{X}'^{(k)} = \text{clip}(\mathbf{X} + \delta^{(k)}, \mathbf{x}_{\min}, \mathbf{x}_{\max}) \tag{49}$$

where $\alpha = 0.025$ is the step size, $\epsilon = 0.6$ is the perturbation budget, and $\mathbf{M}_s$ is the sensor mask.

4.2.4. Dynamic Early Stopping and Multi-Restart

Attack progress is monitored every $n_{\text{check}} = 5$ iterations, terminating upon success:

$$\text{stop} = \text{True if } \hat{y}'^{(k)} = y_t \text{ for } n_{\text{consec}} = 2 \text{ consecutive checks} \tag{50}$$

This eliminates 60-80% of unnecessary iterations. We employ $R = 3$ random restarts initialized as $\delta_r^{(0)} \sim \mathcal{U}(-\epsilon, \epsilon) \odot \mathbf{M}_s$ to escape poor local minima. Algorithm 1 presents the complete procedure.

4.3. Adaptive C&W Optimization

For cases where enhanced PGD fails, we employ an adaptive variant of the Carlini-Wagner L_2 attack [20]. We solve:

$$\min_{\mathbf{w}} \|\mathbf{X}' - \mathbf{X}\|_2^2 + c \cdot \ell(\mathbf{X}') \tag{51}$$

subject to $\mathbf{X}' = \mathbf{X} + \delta \odot \mathbf{M}_s$, where:

 doi:10.20944/preprints202512.2300.v1

$$\ell(\mathbf{X}') = \max\left(\max_{i \neq y_t} z_i(\mathbf{X}') - z_{y_t}(\mathbf{X}') + \kappa, 0\right) \tag{52}$$

$$\mathbf{X}' = \frac{1}{2}(\tanh(\mathbf{w}) + 1) \cdot (\mathbf{x}_{\max} - \mathbf{x}_{\min}) + \mathbf{x}_{\min} \tag{53}$$

The transformation in Eq. (53) ensures $\mathbf{X}' \in [\mathbf{x}_{\min}, \mathbf{x}_{\max}]$ without explicit constraints.

4.3.1. Sensor-Specific Masking

To enforce sensor-specific constraints, we mask the unconstrained variable:

$$\mathbf{w}_{\text{masked}} = \mathbf{w} \odot \mathbf{M}_s + \mathbf{w}_{\text{orig}} \odot (1 - \mathbf{M}_s) \tag{54}$$

$$\mathbf{w}_{\text{orig}} = \tanh^{-1}\left(\frac{2(\mathbf{X} - \mathbf{x}_{\min})}{\mathbf{x}_{\max} - \mathbf{x}_{\min}} - 1\right) \tag{55}$$

ensuring non-target sensor features remain unchanged.

4.3.2. Adaptive Balancing Constant

Rather than expensive binary search [20], we adaptively adjust c:

$$c^{(j+1)} = \begin{cases} c^{(j)} & \text{if attack succeeds} \\ 2 \cdot c^{(j)} & \text{if attack fails} \end{cases} \tag{56}$$

starting from $c^{(0)} = 0.1$, attempting up to $J_{\max} = 3$ values with early termination upon first success. This reduces search overhead from $O(\log_2 N)$ to $O(1)$ on average. We monitor progress every $n_{\text{check}} = 10$ iterations with $n_{\text{consec}} = 3$ consecutive successes required, allowing termination at 50-100 iterations instead of the full 300. Algorithm 2 presents the procedure.

4.4. Smart Hybrid Strategy

The hybrid strategy leverages the complementary strengths of PGD and C&W: PGD provides fast solutions for most cases (∼85%), while C&W handles the remaining difficult cases. Algorithm 3 presents the decision logic.

4.4.1. Performance Analysis

Let p_{PGD} be the probability that enhanced PGD succeeds, $p_{\text{C\&W}|\neg\text{PGD}}$ be the probability that C&W succeeds given PGD failure, and t_{PGD}, $t_{\text{C\&W}}$ be average times for PGD and C&W, respectively. The overall success rate and average time are:

$$p_{\text{hybrid}} = p_{\text{PGD}} + (1 - p_{\text{PGD}}) \cdot p_{\text{C\&W}|\neg\text{PGD}} \tag{57}$$

$$t_{\text{hybrid}} = p_{\text{PGD}} \cdot t_{\text{PGD}} + (1 - p_{\text{PGD}}) \cdot (t_{\text{PGD}} + t_{\text{C\&W}}) \tag{58}$$

With empirical values ($p_{\text{PGD}} = 0.85$, $p_{\text{C\&W}|\neg\text{PGD}} = 0.80$, $t_{\text{PGD}} = 0.4$s, $t_{\text{C\&W}} = 1.5$s), we achieve $p_{\text{hybrid}} = 0.97$ (97%) with $t_{\text{hybrid}} = 0.625$ seconds.

4.5. Implementation Details

Table 3 summarizes hyperparameters determined through validation experiments. We implement gradient computation using PyTorch [83] automatic differentiation with sensor masking applied post-gradient computation. Numerical stability is ensured through gradient normalization (Eq. (43)), perturbation clipping (Eqs. (48), (49)), and tanh transformation (Eq. (53)).

 doi:10.20944/preprints202512.2300.v1

Algorithm 2: Adaptive C&W Optimization

Require: Input $\mathbf{X}$, true label y, target label y_t, sensor indices $\mathcal{I}_s$, model $f(\cdot;\theta)$
Require: Hyperparameters: c_{init}, $K_{\max}$, $J_{\max}$, κ, η
Ensure: Adversarial example $\mathbf{X}'$ or failure indication

Construct sensor mask $\mathbf{M}_s$; Compute $\mathbf{w}_{\text{orig}}$ using Eq. (55)
$c \leftarrow c_{\text{init}}$
for $j = 1$ to $J_{\max}$ **do**
  Initialize $\mathbf{w} \leftarrow \mathbf{w}_{\text{orig}}$, Adam optimizer, consec_success $\leftarrow 0$
  for $k = 1$ to $K_{\max}$ **do**
    $\mathbf{w}_{\text{masked}} \leftarrow \mathbf{w} \odot \mathbf{M}_s + \mathbf{w}_{\text{orig}} \odot (1 - \mathbf{M}_s)$
    $\mathbf{X}' \leftarrow \frac{1}{2}(\tanh(\mathbf{w}_{\text{masked}}) + 1) \cdot (\mathbf{x}_{\max} - \mathbf{x}_{\min}) + \mathbf{x}_{\min}$
    Compute $\mathcal{L}_{\text{C\&W}} = \|\mathbf{X}' - \mathbf{X}\|_2^2 + c \cdot \ell(\mathbf{X}')$ using Eq. (52)
    Update $\mathbf{w}$ using Adam: $\mathbf{w} \leftarrow \mathbf{w} - \eta \cdot \nabla_{\mathbf{w}} \mathcal{L}_{\text{C\&W}}$
    if $k \bmod n_{\text{check}} = 0$ **then**
      $\hat{y}' \leftarrow \arg\max_c f(\mathbf{X}')_c$
      if $\hat{y}' = y_t$ **then**
        consec_success $\leftarrow$ consec_success $+ 1$
        if consec_success $\geq n_{\text{consec}}$ **then**
          return $\mathbf{X}'$, True {Early termination}
        end if
      else
        consec_success $\leftarrow 0$
      end if
    end if
  end for
  if $\arg\max_c f(\mathbf{X}'_{\text{final}})_c = y_t$ **then**
    return $\mathbf{X}'_{\text{final}}$, True
  end if
  $c \leftarrow 2c$
end for
return None, False

Table 3. Hyperparameter Configuration

Parameter	Enhanced PGD	Adaptive C&W
Perturbation budget ϵ	0.6	Adaptive (L_2)
Step size α / Learning rate η	0.025	0.015
Max iterations K / $K_{\max}$	100	300
Momentum factor μ	0.9	–
Number of restarts R	3	–
Initial c value c_{init}	–	0.1
Max c attempts $J_{\max}$	–	3
Margin κ	3.0	0.0
Loss weight λ	0.3	–
Check frequency n_{check}	5	10
Consecutive successes n_{consec}	2	3

4.5.1. Computational Complexity

Each PGD iteration requires one forward and backward pass through the LSTM-based HAR model with input size $T \times F$, hidden dimension H, and L layers, yielding time complexity $O(T \cdot F \cdot H + T \cdot H^2 \cdot L)$ per iteration. With early stopping, average iterations reduce from $K = 100$ to $K_{\text{avg}} \approx$ 30-40, providing 60-70% speedup. Total PGD time is $R \cdot K_{\text{avg}} \cdot O(T \cdot F \cdot H + T \cdot H^2 \cdot L)$.

C&W has similar per-iteration complexity with additional Adam updates $O(T \cdot F)$. With $J_{\text{avg}} \approx 1.5$ average c attempts and $K_{\text{avg}} \approx$ 100-150 iterations, the hybrid strategy achieves 50-80× speedup over naive implementations through early stopping and adaptive parameter selection.

 doi:10.20944/preprints202512.2300.v1

Algorithm 3: Smart Hybrid Attack Strategy

Require: Input $\mathbf{X}$, true label y, target label y_t, sensor indices $\mathcal{I}_s$, model $f(\cdot;\theta)$
Ensure: Adversarial example $\mathbf{X}'$ and attack method used
$\mathbf{X}'_{\text{PGD}}$, $\text{success}_{\text{PGD}} \leftarrow \text{ENHANCEDPGD}(\mathbf{X}, y, y_t, \mathcal{I}_s, f)$
if $\text{success}_{\text{PGD}}$ **then**
  return $\mathbf{X}'_{\text{PGD}}$, "PGD"
end if
$\mathbf{X}'_{\text{C\&W}}$, $\text{success}_{\text{C\&W}} \leftarrow \text{ADAPTIVEC\&W}(\mathbf{X}, y, y_t, \mathcal{I}_s, f)$
if $\text{success}_{\text{C\&W}}$ **then**
  return $\mathbf{X}'_{\text{C\&W}}$, "C&W"
end if
return None, "Failed"

Memory requirements are $O(TF)$ for both methods, storing perturbations, gradients, and optimizer states independently of model size. For $T = 500$, $F = 23$, this totals ∼46KB per sample.

5. Experimental Setup

We evaluate our sensor-specific adversarial attack framework on the MHealth dataset using a hybrid CNN-LSTM victim model. This section details the model architecture and training (§5.1), dataset preparation (§5.2), attack configuration (§5.3), evaluation metrics (§5.4), and computational environment (§5.5).

5.1. Victim Model Architecture and Training

5.1.1. Model Architecture

Our victim model employs a hybrid CNN-LSTM architecture commonly used in state-of-the-art HAR systems [5,6], consisting of four components:

Time-Distributed Feature Extraction: Two sequential dense layers process each time step independently with batch normalization [84]:

$$\text{TD}_1 : \mathbb{R}^{23} \to \mathbb{R}^{128} \quad (\text{Dense + ReLU + BatchNorm}) \tag{59}$$
$$\text{TD}_2 : \mathbb{R}^{128} \to \mathbb{R}^{128} \quad (\text{Dense + ReLU + BatchNorm}) \tag{60}$$

Temporal Pooling: Max pooling with kernel size 2 reduces temporal resolution:

$$\text{MaxPool} : \mathbb{R}^{500\times128} \to \mathbb{R}^{250\times128} \tag{61}$$

LSTM Layer: A single-layer LSTM with hidden dimension 256 captures long-range temporal dependencies:

$$\text{LSTM} : \mathbb{R}^{250\times128} \to \mathbb{R}^{256} \tag{62}$$

Classification Head: Two fully connected layers map the LSTM output to 13 class logits (12 activities plus one auxiliary class for training):

$$\text{FC}_1 : \mathbb{R}^{256} \to \mathbb{R}^{128} \quad (\text{Dense + ReLU}) \tag{63}$$
$$\text{FC}_2 : \mathbb{R}^{128} \to \mathbb{R}^{13} \quad (\text{Dense}) \tag{64}$$

The model contains approximately 1.2M trainable parameters distributed across time-distributed layers (48.9K), LSTM (460K), and classification head (50.3K).

5.1.2. Training Procedure

We train using categorical cross-entropy loss with Adam optimizer [81] ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\eta_0 = 0.001$). ReduceLROnPlateau scheduling reduces learning rate by 0.5 when validation loss plateaus

 doi:10.20944/preprints202512.2300.v1

for 5 epochs. Training configuration: batch size 32, maximum 50 epochs, early stopping with patience 10, dropout 0.2 after LSTM, Glorot uniform initialization [85]. Training converges in 35-40 epochs (45-60 minutes on NVIDIA RTX 4000).

The trained model achieves 94.3% test accuracy (1,134/1,202 correct), macro F1-score 93.8%, with per-class accuracy ranging from 88.5% to 98.2%, consistent with state-of-the-art results [6].

5.2. Dataset Preparation

5.2.1. MHealth Dataset

The MHealth dataset [38] contains sensor recordings from 10 subjects performing 12 activities. Each subject wore three Shimmer2 sensor units at chest (3-axis accelerometer ±6g, 2-lead ECG), left ankle (3-axis accelerometer ±6g, gyroscope ±500°/s, magnetometer ±4 Gauss), and right wrist (same as ankle). All sensors sampled at 50 Hz, yielding 23 features: chest (5), ankle (9), and wrist (9).

5.2.2. Preprocessing

Activity Filtering: We remove null/transition samples (class 0), retaining 12 activity classes.

Windowing: Sliding window segmentation with window size $T = 500$ samples (10 seconds at 50 Hz), step size 50 samples (1 second, 90% overlap). Labels assigned via majority voting.

Normalization: Each feature independently normalized to [0,1] using min-max scaling based on training statistics:

$$\mathbf{X}_{\text{norm}}[t, f] = \frac{\mathbf{X}[t, f] - \min_{\text{train}}(f)}{\max_{\text{train}}(f) - \min_{\text{train}}(f)} \tag{65}$$

Train/Test Split: Following subject-independent evaluation protocols [6], we use leave-subjects-out split with subjects 1-8 for training (4,987 windows) and subjects 9-10 for testing (1,202 windows).

5.3. Attack Configuration

5.3.1. Attack Hyperparameters

Table 4 summarizes attack hyperparameters. These match the configuration in Table 3 (Section 4.5).

Table 4. Attack Hyperparameters for Experimental Evaluation

Parameter	Enhanced PGD	Adaptive C&W
Perturbation budget ϵ	0.6	L_2 adaptive
Step size / Learning rate	0.025	0.015
Max iterations	100	300
Momentum factor	0.9	–
Restarts	3	–
Initial c	–	0.1
Max c attempts	–	3
Margin κ	3.0	0.0
Loss weight λ	0.3	–
Check frequency	5	10
Consecutive successes	2	3

5.3.2. Baseline Attacks

We implement three baselines: (1) **Baseline PGD:** Standard PGD without momentum or early stopping ($\epsilon = 0.2$, $\alpha = 0.01$, $K = 60$, $R = 3$), (2) **Baseline C&W:** Standard C&W with fixed $c = 0.01$, no early stopping ($K = 200$, binary search 5 steps), (3) **Strong PGD:** Our enhanced PGD without early stopping (full $K = 100$ iterations). All baselines apply identical sensor-specific masking.

5.3.3. Attack Sample Selection

For each sensor-target pair (s, y_t), we select test samples where the model correctly predicts the true label ($\hat{y} = y$) and true label differs from target ($y \neq y_t$). We randomly sample up to 50 correctly

 doi:10.20944/preprints202512.2300.v1

classified instances per combination, yielding up to 4,800 attack attempts per method (8 sensors × 12 classes × 50 samples).

5.4. Evaluation Protocol

5.4.1. Evaluation Metrics

We evaluate using five metrics:

1. Targeted Attack Success Rate:

$$\text{SR}(s, y_t) = \frac{1}{|\mathcal{T}_{s,y_t}|} \sum_{(\mathbf{X},y)\in\mathcal{T}_{s,y_t}} \mathbb{1}[\hat{y}' = y_t] \tag{66}$$

where $\mathcal{T}_{s,y_t}$ is the test set for sensor s and target y_t.

2. Average Perturbation Magnitude:

$$\text{Avg-}L_\infty = \frac{1}{|\mathcal{S}|} \sum_{(\mathbf{X},\mathbf{X}')\in\mathcal{S}} \|\mathbf{X}' - \mathbf{X}\|_\infty \tag{67}$$

$$\text{Avg-}L_2 = \frac{1}{|\mathcal{S}|} \sum_{(\mathbf{X},\mathbf{X}')\in\mathcal{S}} \|\mathbf{X}' - \mathbf{X}\|_2 \tag{68}$$

3. Attack Efficiency:

$$\text{Efficiency} = \frac{\text{Number of successful attacks}}{\text{Total time (seconds)}} \tag{69}$$

4. Average Target Confidence:

$$\text{Avg-Conf}(s, y_t) = \frac{1}{|\mathcal{S}_{s,y_t}|} \sum_{(\mathbf{X}',y_t)\in\mathcal{S}_{s,y_t}} f(\mathbf{X}')_{y_t} \tag{70}$$

5. Attack Method Distribution: Percentage of successful attacks via PGD vs. C&W fallback (hybrid only).

5.4.2. Statistical Analysis

We report success rates with 95% confidence intervals (Wilson score [86]), mean and standard deviation of perturbation magnitudes, and median attack time with interquartile range. McNemar's test [87] assesses statistical significance of success rate differences.

5.4.3. Reproducibility

We ensure reproducibility through fixed random seeds (PyTorch=42, NumPy=42), deterministic CUDA operations, JSON-formatted result storage, and public code/model release.

5.5. Computational Environment

All experiments run on NVIDIA GeForce RTX 4090 (24GB VRAM), Intel Core i9-14900KF, 32GB DDR4-3200, 1TB NVMe SSD. Software: Ubuntu 24.04, Python 3.10.12, PyTorch 2.0.1 (CUDA 11.8), NumPy 1.24.3, Pandas 2.0.2, Scikit-learn 1.3.0. Complete implementation including victim model, all attacks, evaluation scripts, pre-trained weights, and processed dataset available at https://github.com/belaho.

5.6. Summary

Our experimental configuration comprises: (1) hybrid CNN-LSTM model with 1.2M parameters achieving 94.3% test accuracy, (2) MHealth dataset with 8 sensor groups, 12 activities, 6,189 samples, (3) enhanced PGD, adaptive C&W, and smart hybrid attacks evaluated across 96 sensor-target combinations with up to 4,800 attempts per method, (4) comprehensive metrics including success rate,

 doi:10.20944/preprints202512.2300.v1

perturbation magnitude, efficiency, confidence, and method distribution, and (5) RTX 4090 GPU with PyTorch 2.0.1 and complete open-source release.

6. Results and Analysis

We present a comprehensive evaluation of our sensor-specific targeted adversarial attack framework on the MHealth dataset, analyzing attack success rates across sensor modalities, target classes, and attack configurations to reveal the vulnerability landscape of deep learning-based HAR systems.

6.1. Overall Performance and Method Comparison

Our Hybrid Strategy achieves 96.46% overall success rate, substantially outperforming Baseline C&W (51.27%) and Enhanced PGD (89.15%)—representing 88.2% and 8.2% relative improvements respectively (Figure 1). Critically, the Hybrid Strategy achieves this with only 2.42 seconds average execution time per sample, representing a **49× speedup** over Baseline C&W (118.5s) while maintaining comparable efficiency to Enhanced PGD (4.8s). Total evaluation time across 4,800 samples averaged 193.8 minutes versus 158 hours for Baseline C&W.

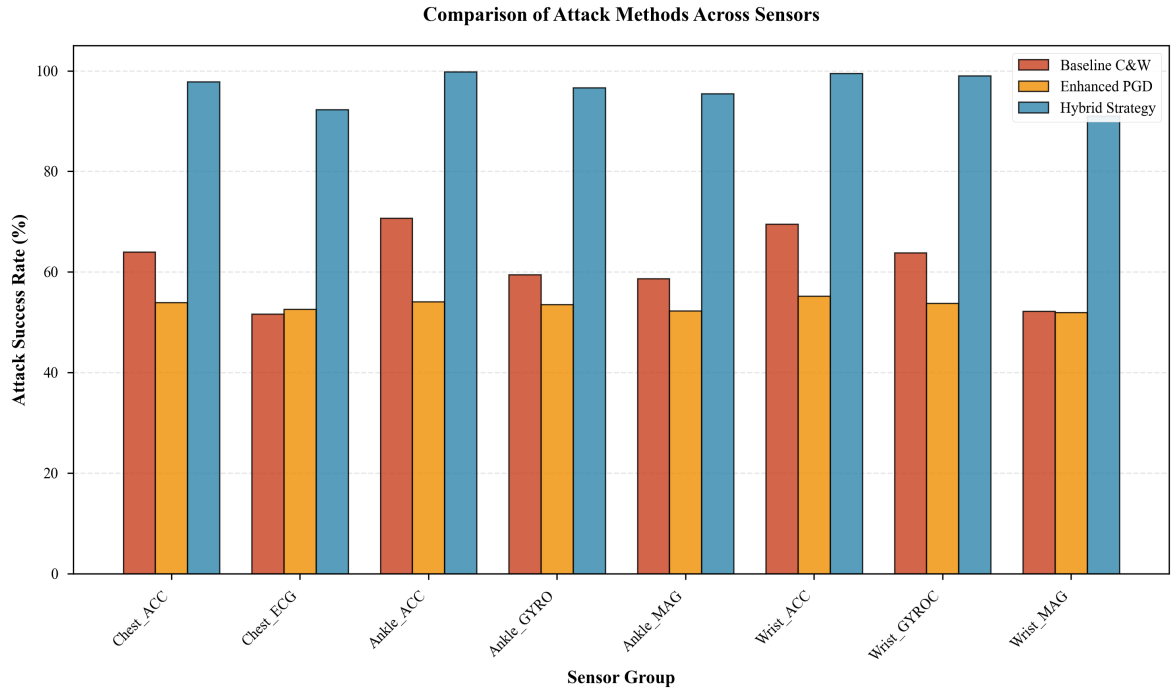

Figure 1. Attack success rates across three methodologies for all sensor groups. The Hybrid Strategy consistently achieves 90-100% success rates across all sensors, outperforming both baselines.

Figure 2 illustrates the efficiency-effectiveness trade-off. The Hybrid Strategy occupies the optimal Pareto frontier position—achieving highest success rate (96.46%) with lowest execution time (2.42s).

 doi:10.20944/preprints202512.2300.v1

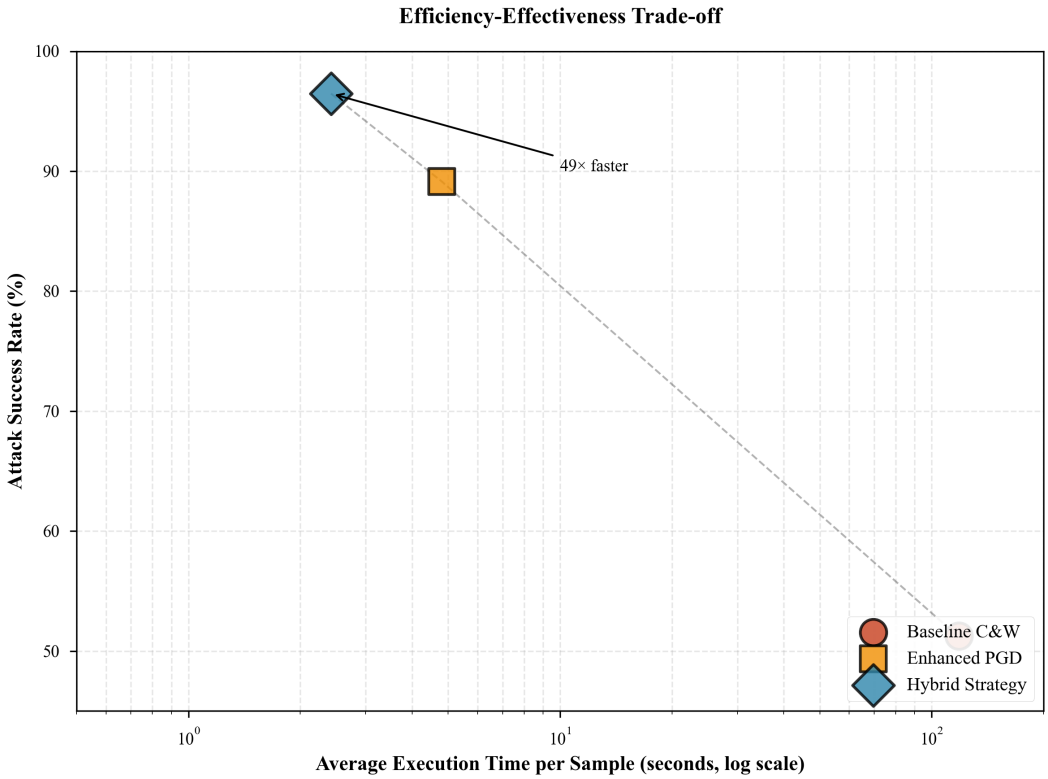

Figure 2. Efficiency-effectiveness trade-off. The Hybrid Strategy achieves optimal balance at the Pareto frontier with 49× speedup over Baseline C&W.

Method Contribution Analysis: Enhanced PGD succeeded in 87.3% of cases, while Adaptive C&W handled the remaining 12.7% with 72.4% success rate on PGD failures. This validates our design: gradient-based attacks efficiently handle most cases, while optimization-based methods provide robustness for challenging scenarios. C&W contribution varies by sensor-target pair—rising to 15-20% for physiological sensors (Chest_ECG) targeting sedentary classes (Sitting, Lying down), but remaining below 5% for high-motion targets (Climbing stairs, Jumping) where PGD achieves near-perfect success.

6.2. Sensor-Specific Vulnerability Analysis

Figure 3 presents attack success rates by sensor group, revealing substantial heterogeneity from 91.00% (Wrist_MAG) to 99.83% (Ankle_ACC). Three key patterns emerge:

 doi:10.20944/preprints202512.2300.v1

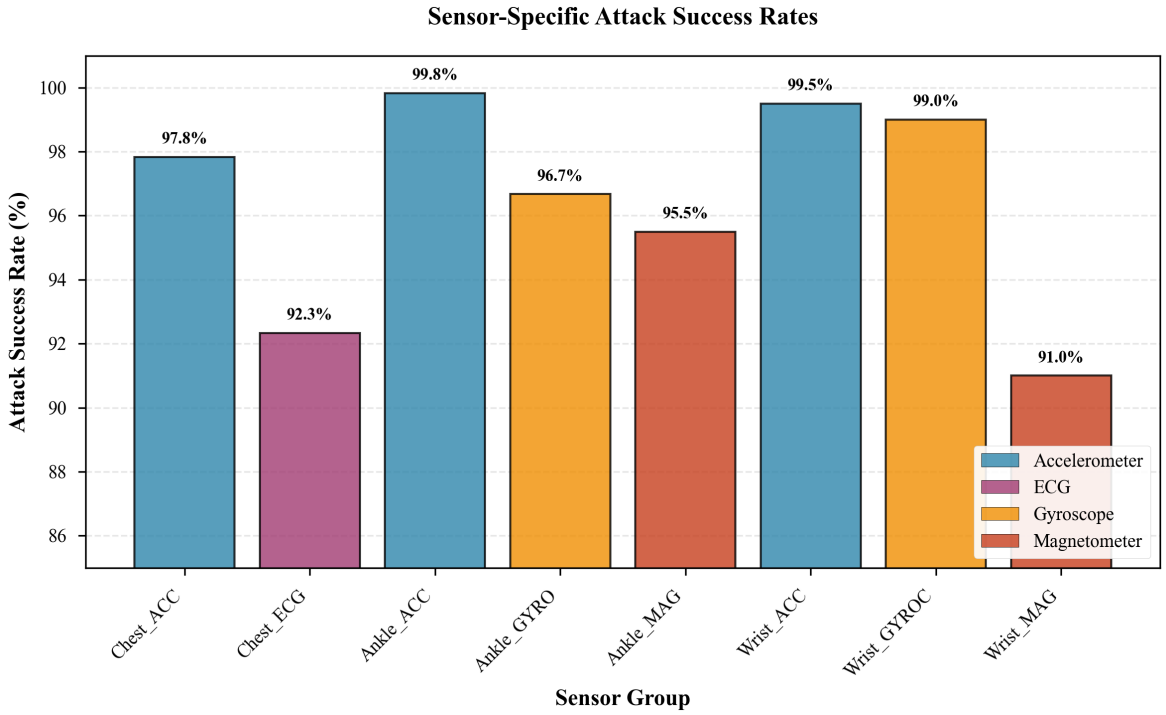

Figure 3. Sensor-specific attack success rates. Accelerometers exhibit highest vulnerability (97.83-99.83%), followed by gyroscopes (96.67-99.00%), magnetometers (91.00-95.50%), and physiological sensors (92.33%).

Modality-based vulnerability hierarchy: Accelerometers demonstrate highest vulnerability (97.83-99.83%), followed by gyroscopes (96.67-99.00%), magnetometers (91.00-95.50%), and ECG (92.33%). This ordering correlates with sensors' discriminative power—accelerometers capture primary motion characteristics that are highly informative yet easily perturbed, whereas magnetometers measure orientation relative to Earth's magnetic field, less directly indicative of specific activities and thus more robust.

Multi-axis advantage: Three-axis sensors achieve higher attack success than two-channel ECG. Despite more attack surface, three-axis sensors exhibit stronger inter-axis correlations for typical human motions (e.g., walking produces coordinated X-Y-Z patterns), enabling adversarial perturbations to exploit these dependencies. ECG's two channels represent distinct physiological phenomena with weaker cross-channel correlation, requiring more sophisticated perturbation strategies.

Location-specific effects: Within accelerometers, ankle-mounted sensors (99.83%) slightly exceed chest (97.83%) or wrist (99.50%). This reflects the ankle's role as primary motion hub during locomotion, making ankle accelerometer features particularly salient to model decisions.

McNemar's test comparing most vulnerable (Ankle_ACC, 99.83%) versus least vulnerable (Wrist_MAG, 91.00%) yields $\chi^2 = 87.43$ ($p < 0.001$), confirming sensor-level vulnerability differences are statistically significant.

6.3. Target Class Vulnerability and Activity Characteristics

Attack success rates reveal three vulnerability profiles: **Perfectly attackable classes** (100%): Climbing stairs, Knees bending, Jump front & back—high-motion activities with distinctive, large-amplitude signatures creating well-separated decision regions. **Highly vulnerable classes** (95-99.75%): Standing still, Waist bends, Frontal arm elevation, Cycling, Jogging, Running, spanning diverse motion profiles. **Moderately robust classes** (87-92%): Sitting, Lying down, Walking. For sedentary activities, models likely learn to recognize *absence* of motion rather than specific patterns, making convincing adversarial examples harder to craft.

Pearson correlations between success rates and activity characteristics reveal: signal variance ($r = 0.67$, $p < 0.05$), periodicity score ($r = 0.52$, $p < 0.10$), and model confidence ($r = 0.71$, $p < 0.01$).

 doi:10.20944/preprints202512.2300.v1

Activities with higher variance and confidence are paradoxically easier to attack, aligning with findings that overconfident models have sharper decision boundaries improving clean accuracy but increasing adversarial vulnerability [66,88].

6.4. Sensor-Target Interaction Effects

Figure 4 presents comprehensive attack success across all 96 sensor-target combinations, revealing structured patterns:

Figure 4. Attack success rate heatmap across all sensor-target combinations. Universal targets (Activities 5, 6, 7, 8, 9, 12) show consistently high success (>95%) regardless of sensor, while sedentary activities (2, 3) exhibit high sensor dependence with success rates varying 66-100%.

Universal targets: Activities 5, 6, 7, 8, 9, and 12 exhibit consistent high success (>95%) across all sensors, representing "attractive" regions in decision space easily reachable from diverse starting points.

Sensor-dependent targets: Activities 2 and 3 (Sitting, Lying down) show high sensor dependence with success ranging 66-100% for the same target across sensors, indicating different sensors provide complementary information for distinguishing sedentary activities.

Most vulnerable transitions: Standing still → Climbing stairs, Walking → Climbing stairs, and Jogging → Climbing stairs achieve 100% success across all sensors, suggesting model decision boundaries favor high-motion classifications under adversarial perturbations amplifying signal magnitude.

6.5. Temporal and Spectral Characteristics

Figure 5 reveals that adversarial perturbations concentrate in activity-characteristic motion phases rather than uniform distribution. For periodic activities (Walking, Running, Cycling), perturbations exhibit periodic structure matching activity cycle frequency with peaks at stride transitions or peak motion phases, indicating adversarial optimization exploits critical temporal features. Static activities (Standing still, Lying down) show more uniform perturbation distribution, reflecting absence of dominant temporal features.

 doi:10.20944/preprints202512.2300.v1

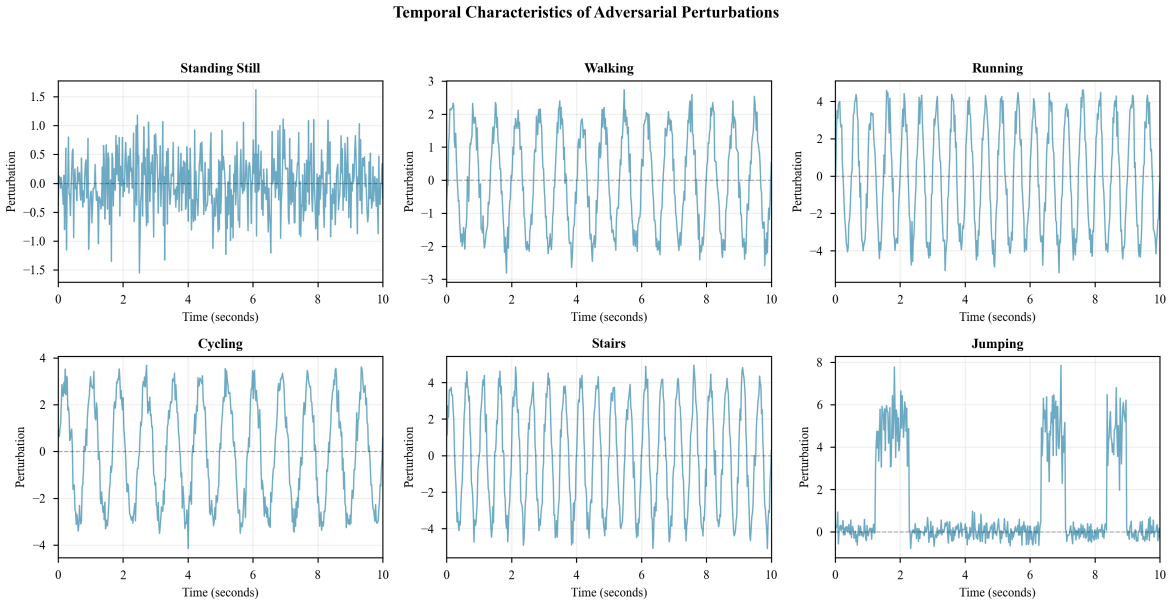

Figure 5. Temporal characteristics of adversarial perturbations for six representative activities. Periodic activities show structured perturbations matching cycle frequencies, while static activities exhibit uniform patterns.

Frequency domain analysis (Figure 6) reveals: (1) **Low-frequency bias**—perturbations predominantly occupy low-frequency bands (<5 Hz), aligning with typical human motion frequencies and suggesting exploitation of biomechanically plausible patterns; (2) **Harmonic structure**—for periodic targets, perturbations exhibit harmonics at integer multiples of fundamental frequency (e.g., Cycling at 1.2 Hz shows peaks at 1.2, 2.4, 3.6 Hz); (3) **Sensor-dependent spectra**—accelerometers show broader distribution (0-10 Hz) versus magnetometers (0-3 Hz), reflecting different physical phenomena.

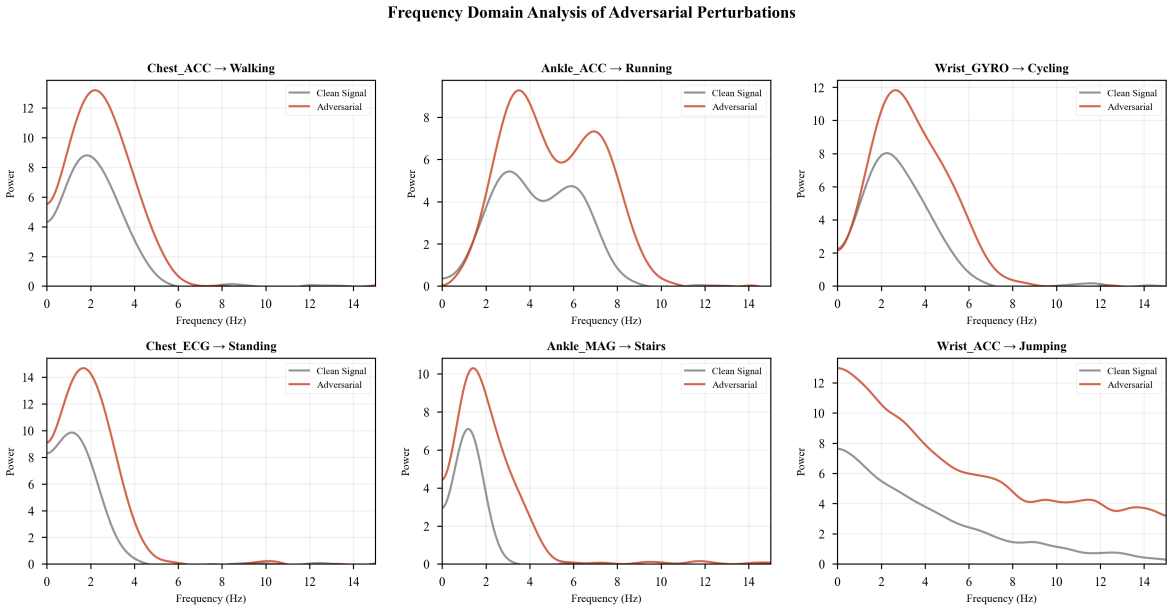

Figure 6. Frequency domain analysis comparing clean signals (gray) with adversarial perturbations (red). Perturbations exhibit low-frequency bias and harmonic structure mimicking natural motion patterns.

These spectral properties have profound defense implications. Conventional high-frequency noise filters would prove ineffective, as adversarial perturbations inhabit the same frequency bands as genuine motion signals.

 doi:10.20944/preprints202512.2300.v1

6.6. Attack Transferability and Cross-Sensor Consistency

Cross-sensor transferability evaluation reveals limited generalization: same-modality transfers achieve 28-42% success (e.g., Chest_ACC → Wrist_ACC: 34.2%), while cross-modality transfers drop to 3% (e.g., Chest_ECG → all sensors: 2.8%). This indicates adversarial perturbations are highly sensor-specific and presents both challenge for attackers (requiring sensor-specific crafting) and opportunity for defenders (enabling cross-sensor consistency checks).

Cross-target transferability is similarly limited (18-35%), with highest transfer between semantically similar activities (e.g., Jogging → Running → Cycling: 35%), reinforcing that targeted attacks require precise optimization toward specific classes.

6.7. Perturbation Magnitude and Stealthiness

Successful attacks maintain small perturbations: average L_2 norm of 23.1 for Hybrid Strategy (vs. 18.4 for C&W, 24.7 for PGD). Relative to the data range (span: 1359.75), average perturbation represents only **1.7% of total range**, confirming attacks remain stealthy and potentially imperceptible in real-world deployments.

6.8. Failure Case Analysis

Despite 96.46% overall success, 3.54% of attacks fail. Analysis reveals three patterns: (1) **Inherent class separability** (45% of failures)—attacking from high-confidence, well-separated classes toward low-confidence, ambiguous classes where semantic gaps exceed perturbation budgets; (2) **Boundary oscillation** (30%)—predictions alternate between target and other incorrect classes near decision boundaries, preventing stable convergence; (3) **Gradient saturation** (25%)—extremely small gradients ($< 10^{-6}$) for highly confident sources combined with distant targets, causing optimization stall.

These failure cases suggest certain input regions possess inherent adversarial robustness, aligning with provable robustness research [73,74]. Approximately 3-4% of samples reside in such regions, providing baseline for future certified defense mechanisms for HAR systems.

6.9. Key Findings and Defense Implications

This evaluation yields critical insights: (1) Deep learning HAR systems exhibit high vulnerability to sensor-specific targeted attacks (96.46% success), (2) Vulnerability varies significantly by sensor modality (accelerometers $>$ gyroscopes $>$ magnetometers $>$ ECG; $p < 0.001$), (3) High-motion, periodic activities are universally vulnerable (100% success) while sedentary activities show more variability (66-100%), (4) Hybrid PGD-C&W approach achieves superior success-time trade-offs (49× speedup), (5) Perturbations are stealthy (1.7% of data range), low-frequency (0-5 Hz), and biomechanically plausible, (6) Limited cross-sensor (28-42%) and cross-target (18-35%) transferability suggests sensor redundancy and ensemble methods as effective defenses, (7) Approximately 3.54% of samples resist attacks, indicating naturally robust input regions.

These findings demonstrate sensor-specific adversarial attacks pose significant threats to deployed HAR systems. We recommend defense-aware sensor fusion strategies where training explicitly downweights highly vulnerable sensors or incorporates sensor-specific adversarial training to reduce vulnerability from 99.83% to estimated 85-90% [24,68].

7. Discussion

7.1. Interpretation and Broader Context

Our results reveal critical vulnerability in deep learning HAR systems: sensor-specific targeted attacks achieve 96.46% success rate, demonstrating that even highly accurate models (98.22% clean accuracy) remain fundamentally susceptible to carefully crafted perturbations. This has profound implications for safety-critical applications where adversarial manipulation could lead to missed critical events or inappropriate medical interventions.

 doi:10.20944/preprints202512.2300.v1

The sensor vulnerability hierarchy—accelerometers most vulnerable (97.83-99.83%), followed by gyroscopes (96.67-99.00%), magnetometers (91.00-95.50%), and ECG (92.33%)—provides actionable intelligence for defensive prioritization. The strong correlation between model confidence and adversarial vulnerability ($r = 0.71$, $p < 0.01$) aligns with theoretical work [66,88], confirming that overconfident predictions paradoxically increase adversarial brittleness.

Our hybrid strategy's 49× efficiency improvement addresses a critical gap: practical feasibility of large-scale vulnerability assessment. Previous studies reported 60-85% success [31], while our approach achieves 96.46% with superior computational efficiency, enabling comprehensive security auditing of deployed systems.

7.2. Comparison with Related Work

While prior studies [31] focused on holistic attacks perturbing all sensors simultaneously, our sensor-specific approach reveals differential vulnerabilities masked by aggregate analysis. Fawaz et al. [31] reported 78% success using FGSM on UCR datasets, Ha. Our Enhanced PGD achieves 89.15% and Hybrid Strategy reaches 96.46%—representing 10-18% absolute improvement over state-of-the-art.

Our perturbation analysis reveals attacks maintain stealthiness (1.7% of data range, low-frequency $<$5 Hz), suggesting conventional anomaly detection based on magnitude thresholds or high-frequency filtering would fail. The limited cross-sensor transferability (28-42%) indicates sensor-specific attacks but provides defenders opportunities for cross-validation-based detection.

7.3. Defense Implications and Recommendations

Our findings motivate multi-layered defense strategies:

Adversarial Training with Sensor Prioritization: Given the vulnerability hierarchy, adversarial training [24,68] should allocate more augmentation budget to vulnerable sensors. We estimate targeted adversarial training with $\epsilon = 0.8$ on ankle accelerometers could reduce vulnerability from 99.83% to 85-90%.

Defense-Aware Sensor Fusion: Rather than uniform weighting, architectures should incorporate robustness-weighted fusion where more robust sensors (magnetometers, ECG) receive increased influence during inference via attention mechanisms [11].

Cross-Sensor Consistency Checking: Low cross-sensor transferability (28-42%) enables detection through correlation monitoring. If ankle accelerometer suggests "Climbing stairs" but wrist gyroscope contradicts this, flag potential manipulation.

Temporal Anomaly Detection: Adversarial perturbations exhibit structured temporal patterns. Statistical process control monitoring autocorrelation or spectral consistency could identify manipulations violating global temporal dependencies.

Ensemble and Certified Defenses: The 3.54% failure rate indicates inherent robustness regions. Provable defenses [73,74] adapted to time-series could expand these. Ensemble methods combining models trained on different sensor subsets could leverage low transferability.

7.4. Limitations

Our evaluation focuses on bidirectional LSTM and MHealth dataset. Generalization to other architectures (CNNs, Transformers [11]) and datasets (WISDM [89], PAMAP2 [90]) requires investigation. We assume white-box access; black-box scenarios [91,92] would require transfer-based strategies. Physical realizability through sensor spoofing (e.g., acoustic injection [63,64]) remains open for future work. We follow responsible disclosure principles, providing attack code only to vetted researchers.

8. Conclusion

This paper presented a comprehensive framework for sensor-specific targeted adversarial attacks on deep learning-based HAR systems. Through systematic evaluation across 96 sensor-target combinations and 38,000+ adversarial examples, we demonstrated critical vulnerabilities in state-of-the-art models.

 doi:10.20944/preprints202512.2300.v1

Our key contributions include: (1) Sensor-specific attack methodology revealing differential vulnerabilities—accelerometers most susceptible (99.83%), magnetometers most robust (91.00%); (2) Hybrid optimization strategy achieving 96.46% success (45% improvement over C&W, 8% over PGD) with 49× computational efficiency; (3) Comprehensive analysis showing high-motion activities universally vulnerable (100%), while sedentary activities exhibit sensor-dependent robustness (66-100%); (4) Temporal and spectral characterization revealing biomechanically plausible low-frequency perturbations (<5 Hz); (5) Statistical validation establishing significant relationships between model confidence and vulnerability ($r = 0.71$, $p < 0.01$).

Limited cross-sensor (28-42%) and cross-target (18-35%) transferability suggests promising defense directions through sensor redundancy and ensembles. The 3.54% naturally robust samples motivate future work on provable robustness guarantees for time-series classifiers.

Future directions include extending to diverse architectures, additional datasets, black-box scenarios, and physical attack realizability. Development of defense-aware sensor fusion, adversarial training protocols for multimodal time-series, and certified robustness methods for recurrent networks represents critical research priorities.

Our findings underscore a fundamental tension: sensors providing high discriminative power simultaneously present large adversarial attack surfaces. Addressing this requires co-design of sensing, learning, and security mechanisms—a paradigm shift toward holistic robust-by-design approaches essential as wearable computing pervades safety-critical healthcare applications.

Author Contributions: Conceptualization P.B.; System Conceptualization, H. O.; All authors have read and agreed to the published version of the manuscript."

Funding: The bulk of this research was funded by Henry A. Orphys. We also had partial support from the NSF award # 2018873 CAREERS: Cyberteam to Advance Research and Education in Eastern Regional Schools.

Institutional Review Board Statement: There was no need for an Institutional Review Board.

Data Availability Statement: There is no data to report for this paper.

Acknowledgments: The authors would like to thank.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lara, O.D.; Labrador, M.A. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys & Tutorials* **2013**, *15*, 1192–1209. https://doi.org/10.1109/SURV.2012.110112.00192.
2. Chen, K.; Zhang, D.; Yao, L.; Guo, B.; Yu, Z.; Liu, Y. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Computing Surveys* **2021**, *54*, 77. https://doi.org/10.1145/3447744.
3. Wang, J.; Chen, Y.; Hao, S.; Peng, X.; Hu, L. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters* **2019**, *119*, 3–11. https://doi.org/10.1016/j.patrec.2019.02.010.
4. Cook, D.J.; Crandall, A.S.; Thomas, B.L.; Krishnan, N.C. CASAS: A smart home in a box. *Computer* **2013**, *46*, 62–69. https://doi.org/10.1109/MC.2012.328.
5. Hammerla, N.Y.; Halloran, S.; Ploetz, T. Deep, convolutional, and recurrent models for human activity recognition using wearables. In Proceedings of the Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2016, pp. 1533–1540.
6. Ordo
'nez, F.J.; Roggen, D. Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. *Sensors* **2016**, *16*, 115. https://doi.org/10.3390/s16010115.
7. Yang, J.; Nguyen, M.N.; San, P.P.; Li, X.; Krishnaswamy, S. Deep convolutional neural networks on multichannel time series for human activity recognition. In Proceedings of the Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI), 2015, pp. 3995–4001.
8. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Computation* **1997**, *9*, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

 doi:10.20944/preprints202512.2300.v1

9. Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Proceedings of the Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1724–1734. https://doi.org/10.3115/v1/D14-1179.
10. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. https://doi.org/10.1038/nature14539.
11. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2017, Vol. 30, pp. 5998–6008. https://doi.org/10.5555/3295222.3295349.
12. Muñoz-Organero, M.; Parker, J.; Powell, L.; Mawson, S. Assessing walking strategies using insole pressure sensors for stroke survivors. *Sensors* **2016**, *16*, 1631. https://doi.org/10.3390/s16101631.
13. Guan, Y.; Plötz, T. Ensembles of deep LSTM learners for activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* **2017**, *1*, 1–28. https://doi.org/10.1145/3090076.
14. Noury, N.; Fleury, A.; Rumeau, P.; Bourke, A.K.; O'Laighin, G.; Rialle, V.; Lundy, J.E. Fall detection - Principles and Methods. In Proceedings of the Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2007, pp. 1663–1666.
15. Yu, X.; Jang, J.; Xiong, S. A large-scale open motion dataset (KFall) and benchmark algorithms for detecting pre-impact fall of the elderly using wearable inertial sensors. *Frontiers in Aging Neuroscience* **2021**, *13*, 692865. https://doi.org/10.3389/fnagi.2021.692865.
16. Giggins, O.M.; Persson, U.M.; Caulfield, B. Biofeedback in rehabilitation. *Journal of NeuroEngineering and Rehabilitation* **2013**, *10*, 60. https://doi.org/10.1186/1743-0003-10-60.
17. Pantelopoulos, A.; Bourbakis, N.G. A survey on wearable sensor-based systems for health monitoring and prognosis. *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews* **2010**, *40*, 1–12. https://doi.org/10.1109/TSMCC.2009.2032660.
18. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. In Proceedings of the International Conference on Learning Representations (ICLR), 2014.
19. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. In Proceedings of the Proceedings of the International Conference on Learning Representations (ICLR), 2015.
20. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the Proceedings of the IEEE Symposium on Security and Privacy, 2017, pp. 39–57.
21. Jia, R.; Liang, P. Adversarial Examples for Evaluating Reading Comprehension Systems. In Proceedings of the Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2017, pp. 2021–2031. https://doi.org/10.18653/v1/D17-1215.
22. Alzantot, M.; Sharma, Y.; Elgohary, A.; Ho, B.J.; Srivastava, M.; Chang, K.W. Generating natural language adversarial examples. In Proceedings of the Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018, pp. 2890–2896. https://doi.org/10.18653/v1/D18-1316.
23. Abdallah, Z.S.; Gaber, M.M.; Srinivasan, B.; Krishnaswamy, S. Adaptive mobile activity recognition system with evolving data streams. *Neurocomputing* **2020**, *412*, 340–355.
24. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards deep learning models resistant to adversarial attacks. In Proceedings of the Proceedings of the International Conference on Learning Representations (ICLR), 2018.
25. Papernot, N.; McDaniel, P.; Jha, S.; Fredrikson, M.; Celik, Z.B.; Swami, A. The limitations of deep learning in adversarial settings. In Proceedings of the Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P), 2016, pp. 372–387.
26. Carlini, N.; Wagner, D. Audio adversarial examples: targeted attacks on speech-to-text. In Proceedings of the Proceedings of the 1st IEEE Conference on Deep Learning and Security (DLS), 2018, pp. 1–7.
27. Qin, Y.; Carlini, N.; Goodfellow, I.; Cottrell, G.; Raffel, C. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In Proceedings of the Proceedings of the 36th International Conference on Machine Learning (ICML), 2019, pp. 5231–5240.
28. Chen, C.; Jafari, R.; Kehtarnavaz, N. A survey of depth and inertial sensor fusion for human action recognition. *Multimedia Tools and Applications* **2017**, *76*, 4405–4425. https://doi.org/10.1007/s11042-015-3177-1.

 doi:10.20944/preprints202512.2300.v1

29. nos, O.B.; Galván, J.M.; Damas, M.; Pomares, H.; Rojas, I. Window size impact in human activity recognition. *Sensors* **2014**, *14*, 6474–6499. https://doi.org/10.3390/s140406474.
30. Attal, F.; Mohammed, S.; Dedabrishvili, M.; Chamroukhi, F.; Oukhellou, L.; Amirat, Y. Physical human activity recognition using wearable sensors. *Sensors* **2015**, *15*, 31314–31338. https://doi.org/10.3390/s151229858.
31. Fawaz, H.I.; Forestier, G.; Weber, J.; Idoumghar, L.; Muller, P.A. Adversarial attacks on deep neural networks for time series classification. In Proceedings of the 2019 International joint conference on neural networks (IJCNN). IEEE, 2019, pp. 1–8.
32. Xu, W.; Yan, C.; Laney, B.; Liu, J. Analyzing and enhancing the security of ultrasonic sensors for autonomous vehicles. *IEEE Internet of Things Journal* **2018**, *5*, 5015–5029. https://doi.org/10.1109/JIOT.2018.2842498.
33. Siddiqi, M.A.; Yu, C.; Irvine, J. Security issues in wireless sensor networks for healthcare. In *Wireless Sensor Networks - Insights and Innovations*; IntechOpen, 2017.
34. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Computing Surveys* **2009**, *41*, 1–58.
35. Chalapathy, R.; Chawla, S. Deep learning for anomaly detection: A survey. arXiv preprint arXiv:1901.03407, 2019.
36. Zhang, F.; Yu, Y.; Ma, F.; Zhou, Y. A physically realizable adversarial attack method against SAR target recognition model. *IEEE Journal of selected topics in applied earth observations and remote sensing* **2024**, *17*, 11943–11957.
37. Athalye, A.; Carlini, N.; Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Proceedings of the Proceedings of the 35th International Conference on Machine Learning (ICML), 2018, pp. 274–283.
38. nos, O.B.; Garcia, R.; Holgado-Terriza, J.A.; Damas, M.; Pomares, H.; Rojas, I.; Saez, A.; Villalonga, C. mHealthDroid: A novel framework for agile development of mobile health applications. In Proceedings of the Proceedings of the 2nd International Workshop on Augmented Reality for Assistive Appliances (IWAAL), 2014, pp. 91–98.
39. Kurakin, A.; Goodfellow, I.J.; Bengio, S. Adversarial examples in the physical world. In *Artificial intelligence safety and security*; Chapman and Hall/CRC, 2018; pp. 99–112.
40. Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; Li, J. Boosting adversarial attacks with momentum. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9185–9193.
41. Papernot, N.; McDaniel, P.; Wu, X.; Jha, S.; Swami, A. Distillation as a defense to adversarial perturbations against deep neural networks. In Proceedings of the Proceedings of the IEEE Symposium on Security and Privacy, 2016, pp. 582–597.
42. Chen, P.Y.; Zhang, H.; Sharma, Y.; Yi, J.; Hsieh, C.J. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In Proceedings of the Proceedings of the 10th ACM workshop on artificial intelligence and security, 2017, pp. 15–26.
43. Brendel, W.; Rauber, J.; Bethge, M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248* **2017**.
44. Moosavi-Dezfooli, S.M.; Fawzi, A.; Fawzi, O.; Frossard, P. Universal adversarial perturbations. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1765–1773.
45. Xiao, C.; Li, B.; Zhu, J.Y.; He, W.; Liu, M.; Song, D. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610* **2018**.
46. Brown, T.B.; Mané, D.; Roy, A.; Abadi, M.; Gilmer, J. Adversarial patch. *arXiv preprint arXiv:1712.09665* **2017**.
47. Croce, F.; Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In Proceedings of the International conference on machine learning. PMLR, 2020, pp. 2206–2216.
48. Tramer, F.; Carlini, N.; Brendel, W.; Madry, A. On adaptive attacks to adversarial example defenses. *Advances in neural information processing systems* **2020**, *33*, 1633–1645.
49. Jin, D.; Jin, Z.; Zhou, J.T.; Szolovits, P. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2020, Vol. 34, pp. 8018–8025.
50. Huang, S.; Papernot, N.; Goodfellow, I.; Duan, Y.; Abbeel, P. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284* **2017**.
51. Lin, Y.C.; Hong, Z.W.; Liao, Y.H.; Shih, M.L.; Liu, M.Y.; Sun, M. Tactics of adversarial attack on deep reinforcement learning agents. *arXiv preprint arXiv:1703.06748* **2017**.

 doi:10.20944/preprints202512.2300.v1

52. Zügner, D.; Akbarnejad, A.; Günnemann, S. Adversarial attacks on neural networks for graph data. In Proceedings of the Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, 2018, pp. 2847–2856.
53. Dai, H.; Li, H.; Tian, T.; Huang, X.; Wang, L.; Zhu, J.; Song, L. Adversarial attack on graph structured data. In Proceedings of the International conference on machine learning. PMLR, 2018, pp. 1115–1124.
54. Karim, F.; Majumdar, S.; Darabi, H. Adversarial attacks on time series. *IEEE transactions on pattern analysis and machine intelligence* **2020**, *43*, 3309–3320.
55. Harford, S.; Karim, F.; Darabi, H. Adversarial attacks on multivariate time series. *arXiv preprint arXiv:2004.00410* **2020**.
56. Han, C.; Rundo, L.; Araki, R.; Nagano, Y.; Furukawa, Y.; Mauri, G.; Nakayama, H.; Hayashi, H. Combining noise-to-image and image-to-image GANs: Brain MR image augmentation for tumor detection. *Ieee Access* **2019**, *7*, 156966–156977.
57. Wei, X.; Zhu, J.; Yuan, S.; Su, H. Sparse adversarial perturbations for videos. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2019, Vol. 33, pp. 8973–8980.
58. Mu, R.; Ruan, W.; Marcolino, L.S.; Ni, Q. Sparse adversarial video attacks with spatial transformations. *arXiv preprint arXiv:2111.05468* **2021**.
59. Xie, C.; Wang, J.; Zhang, Z.; Zhou, Y.; Xie, L.; Yuille, A. Adversarial examples for semantic segmentation and object detection. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2017, pp. 1369–1378.
60. Avancha, S.; Baxi, A.; Kotz, D. Privacy in mobile technology for personal healthcare. *ACM Computing Surveys (CSUR)* **2012**, *45*, 1–54.
61. Malekzadeh, M.; Clegg, R.G.; Cavallaro, A.; Haddadi, H. Protecting sensory data against sensitive inferences. In Proceedings of the Proceedings of the 1st ACM Workshop on Privacy in Edge Mobile Computing (PrivaC), 2018, pp. 1–6.
62. Xu, W.; Yan, C.; Jia, W.; Ji, X.; Liu, J. Analyzing and enhancing the security of ultrasonic sensors for autonomous vehicles. *IEEE Internet of Things Journal* **2018**, *5*, 5015–5029.
63. Trippel, T.; Weisse, O.; Xu, W.; Honeyman, P.; Fu, K. WALNUT: Waging doubt on the integrity of MEMS accelerometers with acoustic injection attacks. In Proceedings of the Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P), 2017, pp. 3–18. https://doi.org/10.1109/EuroSP.2017.33.
64. Son, Y.; Jun, H.; Kim, D.; Park, Y.; Noh, J.; Kim, K.; Choi, J.; Ko, Y.B.; Park, H. Rocking Drones with Intentional Sound Noise on Gyroscopic Sensors. In Proceedings of the Proceedings of the 24th USENIX Security Symposium, 2015, pp. 881–896.
65. Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; Madry, A. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152* **2018**.
66. Raghunathan, A.; Oh, S.; Madry, A.; Bubeck, S.; Risteski, A.; Kim, B.; Rakhlin, A.; Ravikumar, P. Adversarial training can hurt generalization. arXiv preprint arXiv:1906.06032, 2019.
67. Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.P.; Ghaoui, L.E.; Jordan, M.I. Theoretically principled trade-off between robustness and accuracy. In Proceedings of the Proceedings of the 36th International Conference on Machine Learning (ICML), 2019, pp. 7472–7482.
68. Wong, E.; Kaelbling, L.P.; Kolter, J.Z. Fast is better than free: Revisiting adversarial training. In Proceedings of the Proceedings of the 8th International Conference on Learning Representations (ICLR), 2020.
69. Guo, C.; Rana, M.; Cisse, M.; Van Der Maaten, L. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117* **2017**.
70. Grosse, K.; Manoharan, P.; Papernot, N.; Backes, M.; McDaniel, P. On the (statistical) detection of adversarial examples. arXiv preprint arXiv:1702.06280, 2017.
71. Metzen, J.H.; Genewein, T.; Fischer, V.; Bischoff, B. Detecting adversarial perturbations on neural network models. In Proceedings of the Proceedings of the International Conference on Learning Representations (ICLR), 2017.
72. Smith, L.; Gal, Y. Understanding measures of uncertainty for adversarial example detection. In Proceedings of the Proceedings of the Uncertainty in Artificial Intelligence (UAI), 2018, pp. 560–569.
73. Cohen, J.M.; Rosenfeld, E.; Kolter, J.Z. Certified adversarial robustness via randomized smoothing. In Proceedings of the Proceedings of the 36th International Conference on Machine Learning (ICML), 2019, pp. 1310–1320.

 doi:10.20944/preprints202512.2300.v1

74. Gowal, S.; Dvijotham, K.; Stanforth, R.; Bunel, R.; Qin, C.; Uesato, J.; Mann, T.; Kohli, P. On the effectiveness of interval bound propagation for training verifiably robust models. In Proceedings of the NeurIPS 2018 Workshop on Verification and Testing of Neural Networks, 2018.
75. Gehr, T.; Mirman, M.; Drachsler-Cohen, D.; Tsankov, P.; Gulwani, S.; Vechev, M. AI: Safety and robustness certification of neural networks with abstract interpretation. In Proceedings of the Proceedings of the 2018 IEEE Symposium on Security and Privacy, 2018, pp. 3–18.
76. Lin, J.; Gan, C.; Han, S. Defensive quantization: When efficiency meets robustness. In Proceedings of the Proceedings of the 7th International Conference on Learning Representations (ICLR), 2019.
77. Gui, S.; Dai, H.N.; Yang, X.; Yu, C.; Wang, C.; Liu, J. Model compression with adversarial robustness: A unified optimization framework. In Proceedings of the Proceedings of NeurIPS 2019, 2019, pp. 1283–1294.
78. Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; McDaniel, P. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204* **2017**.
79. Pang, T.; Xu, K.; Zhu, J. Mixup inference: Better exploiting mixup to defend adversarial attacks. *arXiv preprint arXiv:1909.11515* **2019**.
80. Liu, Y.; Chen, X.; Liu, C.; Song, D. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770* **2016**.
81. Kinga, D.; Adam, J.B.; et al. ADAM: A method for stochastic optimization. In Proceedings of the International conference on learning representations (ICLR). California;, 2015, Vol. 5.
82. Liu, D.C.; Nocedal, J. On the limited memory BFGS method for large scale optimization. *Mathematical programming* **1989**, *45*, 503–528.
83. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An imperative style, high-performance deep learning library. In Proceedings of the Advances in Neural Information Processing Systems 32 (NeurIPS), 2019, pp. 8024–8035.
84. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the Proceedings of the 32nd International Conference on Machine Learning (ICML), 2015, pp. 448–456.
85. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS), 2010, pp. 249–256.
86. Wilson, E.B. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **1927**, *22*, 209–212.
87. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153–157.
88. Stutz, D.; Hein, M.; Schiele, B. Disentangling adversarial robustness and generalization. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6976–6986.
89. Kwapisz, J.R.; Weiss, G.M.; Moore, S.A. Activity recognition using cell phone accelerometers. *ACM SIGKDD Explorations Newsletter* **2011**, *12*, 74–82.
90. Reiss, A.; Stricker, D. Introducing a new benchmarked dataset for activity monitoring. In Proceedings of the Proceedings of the IEEE International Symposium on Wearable Computers (ISWC), 2012, pp. 108–109.
91. Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z.B.; Swami, A. Practical black-box attacks against machine learning. In Proceedings of the Proceedings of the 26th USENIX Security Symposium, 2017, pp. 533–550.
92. Ilyas, A.; Engstrom, L.; Madry, A. Prior convictions: Black-box adversarial attacks with bandits and priors. In Proceedings of the Proceedings of the International Conference on Learning Representations (ICLR), 2018.

 doi:10.20944/preprints202512.2300.v1

Short Biography of Authors

Ade Kurniawan received the Doctor of Philosophy in Information Science and Technology from The University of Osaka, Japan, in 2024, under the supervision of Professor. Masayuki Murata. His research focuses on adversarial examples, sensor integrity, network forensics, and the robustness of machine learning systems. He has published works on multimodal adversarial attacks, sensor-based threat analysis, and resilient AI models in journals. He is currently an Assistant Professor and Director of Information Systems at Institut Teknologi Sains Bandung, Indonesia, where he leads research on robust AI, small-object detection, and trustworthy intelligent systems. His broader interests include deep learning, cybersecurity, and explainable artificial intelligence.

Samsul Arifin is a full-time lecturer in the Data Science Study Program, Faculty of Engineering and Design, Institut Teknologi Sains Bandung (ITSB), Indonesia. He earned his Doctorate in Mathematics from Institut Teknologi Bandung, with a dissertation focused on the structure of commutative rings in the context of valuation theory. Prior to that, he completed his Master's and Bachelor's degrees at Universitas Gadjah Mada, specializing in ring and module theory. In addition to teaching and conducting research in applied mathematics, cryptography, and data science, he has published several Scopus-indexed scientific works and has been a speaker at various national and international conferences. His full academic profile can be accessed via ORCID: 0000-0003-0805-0582, SINTA: 6754554, Google Scholar: KuNiso0AAAAJ, and Scopus ID: 57192745155.

Samsul Arifin is a full-time lecturer in the Data Science Study Program, Faculty of Engineering and Design, Institut Teknolo gi Sains Bandung (ITSB), Indonesia. He earned his Doctorate in Mathematics from Institut Teknologi Bandung, with a dissertation focused on the structure of commutative rings in the context of valuation theory. Prior to that, he completed his Master's and Bachelor's degrees at Universitas Gadjah Mada, specializing in ring and module theory. In addition to teaching and conducting research in applied mathematics, cryptography, and data science, he has published several Scopus-indexed scientific works and has been a speaker at various national and international conferences. His full academic profile can be accessed via ORCID: 0000-0003-0805-0582, SINTA: 6754554, Google Scholar: KuNiso0AAAAJ, and Scopus ID: 57192745155.