Review

Not peer-reviewed version

Temporal Adversarial Attacks on Time Series and Reinforcement Learning Systems: A Systematic Survey, Taxonomy, and Benchmarking Roadmap

Ade Kurniawan [*] , Merios Gusan Putra , Dani Lukman Hakim , Mochammad Ariyanto

Posted Date: 8 January 2026

doi: 10.20944/preprints202601.0598.v1

Keywords: adversarial attacks; time series classification; human activity recognition; reinforcement learning; deep neural networks; sensor systems

Review

Temporal Adversarial Attacks on Time Series and Reinforcement Learning Systems: A Systematic Survey, Taxonomy, and Benchmarking Roadmap

Ade Kurniawan [1,*], Merios Gusan Putra [2], Dani Lukman Hakim [3] and Mochammad Ariyanto [4]

[1] Department of Data Science, Institut Teknologi Sains Bandung, Kota Deltamas Lot-A1 CBD, Bekasi 17530, Jawa Barat, Indonesia
[2] Department of Digital Business, Institut Teknologi Sains Bandung, Kota Deltamas Lot-A1 CBD, Bekasi 17530, Jawa Barat, Indonesia
[3] Department of Palm Oil Processing Technology, Institut Teknologi Sains Bandung, Kota Deltamas Lot-A1 CBD, Bekasi 17530, Jawa Barat, Indonesia
[4] Department of Mechanical Engineering, Universitas Diponegoro, Jl. Prof. Sudarto SH, Semarang 50275, Jawa Tengah, Indonesia
* Correspondence: ade.k@itsb.ac.id

Abstract

Deep learning systems processing temporal and sequential data are increasingly deployed in safety-critical applications including healthcare monitoring, autonomous navigation, and algorithmic trading. However, these systems exhibit severe vulnerabilities to adversarial attacks—carefully crafted perturbations that cause systematic misclassification while remaining imperceptible. This paper presents a comprehensive systematic survey of adversarial attacks on time series classification, human activity recognition (HAR), and reinforcement learning (RL) systems, reviewing 127 papers published between 2019 and 2025 following PRISMA guidelines with documented inter-rater reliability ($\kappa = 0.83$). We establish a unified four-dimensional taxonomy distinguishing attack characteristics across target modalities (wearable IMU sensors, WiFi/radar sensing, skeleton-based recognition, medical/financial time series, and RL agents), perturbation strategies, temporal scope, and physical realizability levels. Our quantitative synthesis reveals severe baseline vulnerabilities—FGSM attacks degrade HAR accuracy from 95.1% to 3.4% under white-box conditions—while demonstrating that cross-sensor transferability varies dramatically from 0% to 80% depending on body placement and modality. Critically, we identify a substantial gap between digital attack success rates (85–98%) and physically validated attacks, with hardware-in-the-loop validation demonstrating 70–97% success only for WiFi and radar modalities, while wearable IMU physical attacks remain entirely unvalidated. We provide systematic analysis of defense mechanisms including adversarial training, detection-based approaches, certified defenses, and ensemble methods, proposing the Temporal AutoAttack (T-AutoAttack) framework for standardized adaptive attack evaluation. Our analysis reveals that current defenses exhibit 6–23% performance degradation under adaptive attacks, with certified methods showing the smallest gap but incurring 15–30% clean accuracy costs. We further identify emerging vulnerabilities in transformer-based HAR architectures and LLM-based time series forecasters that require urgent attention. The survey culminates in a prioritized research roadmap identifying eight critical gaps with specific datasets, evaluation pipelines, and implementation timelines. We provide actionable deployment recommendations for practitioners across wearable HAR, WiFi/radar sensing, RL systems, and emerging LLM-based temporal applications. This work offers the first unified treatment bridging time series and reinforcement learning adversarial research, establishing foundations for developing robust temporal AI systems suitable for real-world deployment in safety-critical domains.

Keywords: adversarial attacks; time series classification; human activity recognition; reinforcement learning; deep neural networks; sensor systems

1. Introduction

Deep learning models have become foundational components in systems that process temporal and sequential data, enabling applications ranging from wearable health monitoring and autonomous navigation to algorithmic trading and industrial process control. The discovery that neural networks are vulnerable to adversarial examples—carefully crafted perturbations that are imperceptible to humans yet cause systematic misclassification—has raised fundamental security concerns for these deployed systems [1,2]. While adversarial robustness has been extensively characterized for image classifiers, temporal systems present *distinct* rather than necessarily greater vulnerabilities. Empirical findings in human activity recognition (HAR) demonstrate severe accuracy degradation—from 95.1% to merely 3.4% under FGSM attacks for DNN classifiers, and from 93.1% to 16.8% for CNN-based models [3]. However, these dramatic drops must be interpreted within their modality-specific context: unlike image perturbations constrained primarily by pixel-level imperceptibility, time series attacks face fundamentally different constraints including signal continuity, sensor measurement bounds, and inter-sensor correlations that can either amplify or attenuate vulnerability depending on the deployment scenario.

The adversarial robustness of image classifiers has received extensive research attention, yielding well-established attack methodologies, defense mechanisms, and evaluation benchmarks. However, adversarial attacks targeting time series and reinforcement learning (RL) systems present *qualitatively distinct* challenges that remain inadequately addressed. Crucially, the vulnerability landscape differs fundamentally across temporal system modalities. *On-body sensor systems* (accelerometers, gyroscopes, magnetometers) require perturbations that respect physical measurement constraints and maintain inter-sensor correlations inherent to rigid-body motion dynamics. *Device-free RF sensing systems* (WiFi CSI, mmWave radar) face non-differentiable signal processing pipelines and propagation-path constraints that fundamentally alter the attack surface [4]. These modality-specific characteristics necessitate tailored attack and defense strategies rather than direct adaptation of image-domain techniques.

Time series data exhibits temporal dependencies, variable sequence lengths, and domain-specific physical constraints that are absent in static image data. Unlike images where perturbations are constrained primarily by perceptual imperceptibility, time series attacks must additionally respect signal continuity requirements, sensor measurement bounds, inter-sensor correlations, and energy constraints inherent to battery-powered wearable devices—constraints that vary significantly across sensor modalities and deployment contexts. Reinforcement learning systems face sequential decision-making vulnerabilities where adversarial perturbations can compound across time steps, potentially leading to catastrophic failures in safety-critical applications.

The practical implications of these vulnerabilities extend across multiple critical domains. In healthcare monitoring, recent work has demonstrated that adversarial perturbations can manipulate fall detection systems, potentially causing life-threatening delays in emergency response for elderly patients [5]. Smart home gesture recognition systems based on radar sensing have been shown vulnerable to attacks that perturb only the padding regions of input sequences without modifying actual gesture frames [6]. Financial trading systems face manipulation through ephemeral perturbations that induce suboptimal buy/sell decisions while remaining statistically undetectable [7].

1.1. Motivation and Scope

The increasing deployment of deep learning in safety-critical temporal applications motivates urgent investigation of adversarial vulnerabilities. We examine several representative scenarios that illustrate the breadth and severity of potential attacks.

In healthcare monitoring, wearable devices continuously classify user activities for fall detection, cardiac arrhythmia identification, and medication adherence. An adversary who can subtly perturb sensor readings may cause dangerous misclassifications, potentially delaying emergency responses. The ADAR framework demonstrated that adversarial attacks exhibit four distinct transferability dimensions in wearable HAR systems: between different ML models, across different users, across sensor body locations, and across different datasets [3]. This multi-dimensional transferability implies that an adversary need not have knowledge of the specific deployment configuration to craft effective attacks.

In autonomous systems, vehicles rely on temporal sensor fusion across cameras, LiDAR, radar, and inertial measurement units for perception and planning. Adversarial perturbations targeting these temporal streams can cause navigation failures with life-threatening consequences. Recent work on millimeter-wave radar sensing has shown that physically realizable attacks using low-cost metamaterial tags can achieve 97% accuracy in manipulating range estimation, 96% for angle estimation, and 91% for speed estimation—at costs 10-100× lower than existing attack methods [8].

In financial systems, algorithmic trading increasingly depends on time series forecasting models processing market data, news feeds, and transaction records. Adversarial attacks on these systems can exploit the temporal structure of financial data through targeted perturbations that manipulate predictions in specific directions (bullish or bearish), at particular amplitudes, or during critical time windows [9].

Beyond traditional deep learning architectures, the emergence of large language models (LLMs) for time series analysis introduces novel vulnerability dimensions. Recent work has demonstrated that LLM-based time series forecasters—including TimeGPT, GPT-4, LLaMA, and Mistral variants—exhibit distinct adversarial susceptibilities compared to conventional neural architectures [10,11]. Xiao et al. [12] demonstrated learning-based attacks specifically targeting temporal forecasting models through directional, amplitudinal, and temporal perturbation strategies. These findings suggest that the rapid adoption of foundation models for temporal applications may be outpacing security analysis, creating an urgent need for comprehensive adversarial assessment.

This survey addresses three interconnected domains with particular emphasis on sensor-specific vulnerabilities:

1. **Time series adversarial attacks:** We comprehensively examine HAR systems with detailed analysis of attacks targeting individual sensor modalities (accelerometer, gyroscope, magnetometer) and their combinations, WiFi channel state information (CSI) and radar-based sensing systems, skeleton-based action recognition, and medical/financial time series applications.
2. **Reinforcement learning attacks:** We analyze state observation perturbations, reward poisoning mechanisms, adversarial policy training, and the emerging paradigm of using RL algorithms for generating attacks on deep neural networks.
3. **Cross-cutting themes:** We identify critical patterns spanning explainability-guided attack generation, multi-sensor fusion vulnerabilities, physical realizability constraints, and certified defense mechanisms.

1.2. Gaps in Existing Surveys

Despite substantial growth in adversarial machine learning research, existing surveys exhibit significant coverage gaps for temporal and sequential systems. Table 1 presents a systematic comparison with representative surveys published between 2019 and 2025, revealing several critical limitations.

First, there is an absence of dedicated time series coverage. Foundational work by Fawaz et al. [13] investigated attacks on time series classification but did not provide comprehensive survey coverage. Subsequent surveys have focused predominantly on computer vision [14–17], natural language processing [18,19], or domain-agnostic adversarial machine learning [20,21]. None of these surveys address the unique characteristics of sensor-based time series data, including the constraints imposed by physical measurement processes, the multi-modal nature of wearable sensor systems, or the temporal dependencies inherent in sequential activity data. Furthermore, emerging threats targeting time series anomaly detection systems [22] and autoregressive forecasting models [23] remain unexamined within a unified adversarial framework.

Second, existing HAR and sensor system security analyses are limited. Prior HAR surveys [24,25] emphasize recognition methodologies rather than adversarial robustness. Sensor system surveys address signal processing techniques without systematic treatment of adversarial threats. The notable exception is Sakka et al. [5], who examined security issues in HAR systems but focused primarily on medical IoT applications without comprehensive analysis of attack methodologies, threat models, or defense mechanisms across the broader HAR landscape.

Third, no existing survey provides a unified temporal and RL perspective. Despite the shared sequential nature of time series and RL systems—and their increasing co-deployment in real-world applications such as autonomous vehicles and robotic systems—no existing survey bridges these domains within a coherent analytical framework. Recent RL security surveys [26–28] focus exclusively on RL agent vulnerabilities without considering the application of RL techniques for attack generation against non-RL systems.

Fourth, the critical dimension of sensor-specific vulnerabilities remains unaddressed. Existing work has demonstrated that different sensor modalities (accelerometer vs. gyroscope vs. magnetometer) exhibit dramatically different vulnerability profiles, with cross-sensor transferability ranging from 0% to over 80% depending on body placement and sensor type [3,29]. No existing survey systematically analyzes these sensor-specific characteristics or their implications for attack and defense design.

Fifth, novel architectural paradigms present uncharacterized vulnerabilities. Spiking neural networks (SNNs) for time series classification—increasingly deployed for energy-efficient edge inference—exhibit adversarial susceptibility patterns distinct from conventional architectures [30]. Similarly, attention-based forecasting models face query-specific vulnerabilities through black-box attack strategies [31]. No existing survey systematically addresses these architectural diversity considerations within the temporal domain.

Table 1. Systematic comparison with representative adversarial attack surveys (2019–2025). This survey uniquely addresses sensor-specific vulnerabilities in HAR systems and the intersection of time series and RL attack methodologies.

Survey	Year	Scope	Primary Focus	Gaps Addressed Here
Qiu et al. [17]	2025	CV	10-year retrospective, LVLMs	No temporal/RL/sensor coverage
Costa et al. [16]	2024	CV	DL attacks/defenses, ViT	Limited to static images
Pawlicki et al. [21]	2025	General	Meta-survey, diffusion models	Lacks domain-specific depth
Ilahi et al. [28]	2024	RL	Attacks and countermeasures	No time series/sensor integration
Schott et al. [26]	2024	RL	Observation/dynamics attacks	No TS/HAR integration
Goyal et al. [19]	2023	NLP	Defense mechanisms	No temporal/sensor data focus
Sakka et al. [5]	2023	HAR	Medical IoT vulnerabilities	Limited attack methodology analysis
Sah & Ghasemzadeh [3]	2019	HAR	Transferability analysis	Single-domain, no RL coverage
This Survey	**2025**	**TS+RL**	**Unified temporal/sequential with sensor-specific analysis**	**First comprehensive HAR+RL survey**

1.3. Contributions

This survey makes the following contributions:

1. **First unified temporal attack survey with sensor-specific and modality-aware analysis.** We provide an integrated treatment of adversarial attacks on time series and reinforcement learning systems, establishing detailed taxonomies spanning HAR with emphasis on individual sensor

modalities (accelerometer, gyroscope, magnetometer, IMU fusion), WiFi CSI and radar-based sensing, skeleton-based action recognition, and RL agent policies. Our analysis uniquely examines how attack effectiveness varies across sensor types, body placements, fusion strategies, and—critically—distinguishes between on-body and device-free sensing paradigms that present fundamentally different attack surfaces.

2. **Systematic literature analysis.** We systematically review over 120 papers published between 2019 and 2025, comprising time series attack papers across multiple application domains, RL attack papers spanning state perturbation to reward poisoning, and papers addressing cross-cutting themes including physical realizability and certified defenses.
3. **Comprehensive comparison framework with quantitative analysis.** We present detailed comparison tables cataloging attack methods with standardized metadata including datasets, threat models, attack success rates, perturbation constraints, key contributions, and identified limitations. We provide quantitative analysis of attack effectiveness across different sensor configurations and threat model assumptions.
4. **Critical analysis of methodological limitations.** Beyond cataloging existing work, we provide critical analysis of methodological limitations in current research, including the prevalence of unrealistic threat models, the gap between digital and physical attack validation, and the lack of standardized evaluation protocols for temporal adversarial attacks.
5. **Coverage of emerging architectural vulnerabilities.** We extend analysis beyond conventional CNN/LSTM architectures to address adversarial susceptibilities in transformer-based temporal models, spiking neural networks for edge deployment, and LLM-based time series forecasters—architectural paradigms that are rapidly being adopted but whose security properties remain undercharacterized.
6. **Research roadmap.** We articulate eight critical research gaps at the intersections of explainability, sensor targeting, temporal optimization, and RL-based attack generation, providing concrete directions for future investigation.

1.4. Survey Methodology

This survey follows systematic review guidelines adapted for computer science literature [32], incorporating elements of PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology where applicable to empirical adversarial machine learning research.

1.4.1. Search Strategy and Databases

We conducted a systematic literature search across IEEE Xplore, ACM Digital Library, Springer, Elsevier ScienceDirect, and arXiv using query terms including “adversarial attack,” “time series,” “human activity recognition,” “wearable sensor,” “accelerometer,” “gyroscope,” “IMU,” “WiFi sensing,” “reinforcement learning,” “sensor attack,” and “temporal perturbation.” The search covered publications from January 2019 through December 2025. Boolean combinations were employed to maximize recall while maintaining precision: (“adversarial” OR “perturbation” OR “attack”) AND (“time series” OR “temporal” OR “sequential”) AND (“classification” OR “recognition” OR “forecasting” OR “reinforcement learning”).

1.4.2. Inclusion and Exclusion Criteria

Inclusion criteria required papers to: (1) propose novel attack or defense methods for temporal or RL systems, (2) provide empirical evaluation on established benchmarks or real-world datasets, (3) be published in peer-reviewed venues or appear as preprints with substantial citation counts, and (4) provide sufficient methodological detail for reproducibility assessment.

Exclusion criteria removed papers that: (1) focused exclusively on image or text domains without temporal components, (2) presented only theoretical analysis without empirical validation, (3) were superseded by extended journal versions from the same authors, (4) lacked quantitative attack success metrics or defense evaluation, or (5) were not available in English.

We additionally applied quality filters prioritizing publications in top-tier venues including IEEE TPAMI, Pattern Recognition, NeurIPS, ICML, ICLR, CVPR, ICCV, AAAI, IJCAI, ACM CCS, USENIX Security, IEEE S&P, ACM UbiComp/IMWUT, IEEE TMC, and ACM MobiCom. For emerging topics with limited top-tier coverage, we included carefully vetted arXiv preprints with verified experimental results and citation counts exceeding 10.

1.4.3. Study Selection Process

Figure 1 illustrates our study selection process following PRISMA guidelines. Initial database searches yielded 847 potentially relevant records. After removing 156 duplicates, 691 records underwent title and abstract screening by two independent reviewers (AK and MGP). Screening disagreements were resolved through discussion, achieving inter-rater agreement of $\kappa = 0.83$ (Cohen's kappa), indicating substantial agreement.

Of the 691 screened records, 412 were excluded based on title/abstract review (primarily due to focus on image/NLP domains or lack of empirical evaluation). The remaining 279 full-text articles were assessed for eligibility. Of these, 152 were excluded: 67 for insufficient methodological detail, 43 for lack of temporal/RL focus, 28 for missing quantitative metrics, and 14 for being superseded by extended versions.

Our final corpus comprises 127 papers, distributed across time series classification and HAR (42 papers, 33.1%), WiFi and radar sensing (18 papers, 14.2%), skeleton-based recognition (15 papers, 11.8%), medical and financial time series (14 papers, 11.0%), reinforcement learning attacks (23 papers, 18.1%), and defense mechanisms (15 papers, 11.8%).

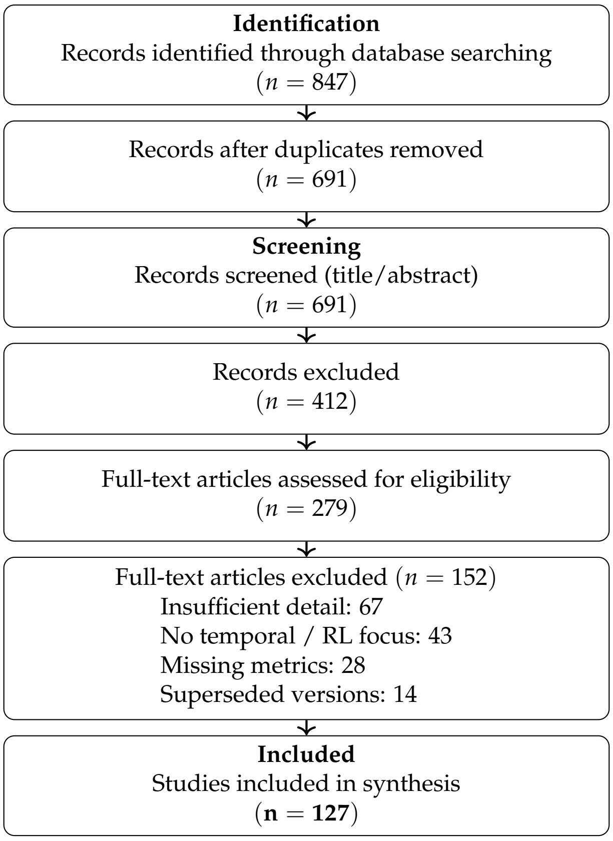
Identification
Records identified through database searching
(n = 847)
Records after duplicates removed
(n = 691)
Screening
Records screened (title/abstract)
(n = 691)
Records excluded
(n = 412)
Full-text articles assessed for eligibility
(n = 279)
Full-text articles excluded (n = 152)
Insufficient detail: 67
No temporal / RL focus: 43
Missing metrics: 28
Superseded versions: 14
Included
Studies included in synthesis
(n = 127)

Figure 1. PRISMA flow diagram for systematic literature selection. Inter-rater agreement for screening: $\kappa = 0.83$.

1.4.4. Risk-of-Bias Assessment

We developed a domain-specific risk-of-bias rubric to assess the quality and reliability of included studies. Each paper was evaluated across four dimensions, scored on a scale of 1 (low quality) to 3 (high quality):

Threat model realism (TMR): Evaluates whether the assumed attacker capabilities are realistic for the target deployment scenario. Score 3: Physically validated attacks with realistic constraints; Score 2: Digital attacks with physical constraints (sensor bounds, smoothness); Score 1: Unconstrained digital attacks assuming direct model input access.

Evaluation rigor (ER): Assesses the comprehensiveness of empirical evaluation. Score 3: Multiple datasets, multiple model architectures, ablation studies, statistical significance testing; Score 2: Single dataset with multiple models or multiple datasets with single model; Score 1: Single dataset, single model, no ablations.

Reproducibility (REP): Evaluates the availability of implementation details and code. Score 3: Open-source code with documented parameters; Score 2: Detailed algorithmic description enabling reimplementation; Score 1: High-level description only.

Baseline comparison (BC): Assesses comparison with prior work. Score 3: Comprehensive comparison with ≥ 3 recent baselines under identical conditions; Score 2: Comparison with 1-2 baselines; Score 1: No baseline comparison or incomparable setups.

Table 2 summarizes the distribution of risk-of-bias scores across included studies. We observe that threat model realism (mean: 1.7) represents the most significant quality concern, with the majority of studies assuming unrealistic white-box access. Reproducibility has improved in recent years (mean: 2.3 for 2023-2025 vs. 1.8 for 2019-2022), reflecting growing emphasis on open science practices.

Table 2. Risk-of-bias assessment summary across 127 included studies. Higher scores indicate higher quality.

Dimension	Mean	Median	Std
Threat Model Realism (TMR)	1.7	2	0.6
Evaluation Rigor (ER)	2.1	2	0.7
Reproducibility (REP)	2.1	2	0.8
Baseline Comparison (BC)	2.0	2	0.7
Overall Quality	**2.0**	**2**	**0.5**

1.4.5. Data Extraction and Synthesis

For each included study, we extracted: (1) attack/defense methodology and key algorithmic contributions, (2) target domain and sensor modalities, (3) threat model assumptions, (4) datasets and experimental setup, (5) quantitative results (ASR, robust accuracy, perturbation magnitude), (6) claimed limitations, and (7) code/data availability. Extraction was performed by DLH and verified by AK.

Due to heterogeneity in experimental setups, perturbation budgets, and evaluation protocols across studies, formal meta-analysis was not feasible for most comparisons. Instead, we provide narrative synthesis organized by our taxonomic framework, with quantitative comparisons where methodological alignment permits. Section 7 presents aggregated findings for comparable experimental conditions.

1.4.6. Limitations of This Review

We acknowledge several limitations of our systematic review:

Selection bias: Prioritization of top-tier venues may exclude relevant work from regional conferences or emerging venues. Rapid developments in LLM-based temporal systems may result in coverage gaps for very recent preprints.

Publication bias: Published studies likely over-represent successful attacks and effective defenses, potentially inflating reported success rates relative to real-world performance.

Heterogeneity: Variation in experimental protocols, perturbation budgets, and evaluation metrics limits quantitative synthesis across studies.

Language bias: Restriction to English-language publications may exclude relevant non-English literature.

Despite these limitations, our systematic approach provides the most comprehensive coverage of temporal adversarial attacks to date, with transparent methodology enabling future updates and extensions.

1.5. Organization

The remainder of this survey is organized as follows. Section 2 establishes foundational concepts in adversarial attacks and presents our taxonomic framework for temporal systems, with particular attention to sensor-specific characteristics. Section 3 comprehensively reviews time series adversarial attacks across HAR (with sensor-specific analysis), WiFi/radar sensing, skeleton-based recognition, and financial/medical domains. Section 4 analyzes RL adversarial attacks including state perturbations, reward poisoning, adversarial policies, and RL-based attack generation. Section 5 addresses cross-cutting themes including physical realizability, multi-modal vulnerabilities, and certified defenses. Section 6 discusses defense mechanisms with attention to temporal system requirements. Section 7 presents evaluation methodologies and benchmark datasets. Section 8 identifies critical research gaps and future directions. Section 9 concludes with a synthesis of key findings.

2. Background and Taxonomy

This section establishes the foundational concepts underlying adversarial attacks on temporal systems and introduces our taxonomic framework. We place particular emphasis on the unique characteristics of sensor-based time series data that distinguish it from static image data and necessitate specialized attack and defense methodologies. Critically, we distinguish between fundamentally different sensing paradigms—on-body inertial measurement and device-free RF sensing—that present distinct attack surfaces and defense requirements.

2.1. Adversarial Attack Fundamentals

Adversarial attacks exploit the sensitivity of neural networks to carefully crafted input perturbations. Given a classifier $f : \mathcal{X} \to \mathcal{Y}$ and a clean input x with true label y, an adversarial example x^{adv} satisfies:

$$f(x^{adv}) \neq y \quad \text{and} \quad d(x, x^{adv}) \leq \epsilon \tag{1}$$

where $d(\cdot, \cdot)$ is a distance metric and ϵ bounds the perturbation magnitude. We characterize attacks along three primary dimensions: attacker knowledge, attack objectives, and attack timing.

2.1.1. Threat Models Based on Attacker Knowledge

White-box attacks assume complete model access, including architecture, parameters, and gradients. This setting enables gradient-based optimization of perturbations and represents the strongest adversary. In sensor-based HAR systems, white-box attacks are particularly powerful because gradient information can reveal which time steps and sensor channels are most vulnerable to perturbation [3].

Gray-box attacks assume partial information, such as knowledge of the model architecture without access to trained parameters, or access to a related surrogate model trained on similar data. This threat model is especially relevant for HAR systems where attackers may know the general architecture (e.g., CNN or LSTM) used for activity recognition without having access to the specific trained weights deployed on a target device.

Black-box attacks restrict the adversary to input-output access only. Within this category, *score-based attacks* can observe prediction confidence scores, while *decision-based attacks* receive only hard classification labels. Black-box attacks on HAR systems face additional challenges because query access may be limited by network connectivity, battery constraints on wearable devices, or rate limiting by cloud-based inference services. Recent advances in query-efficient black-box methods include square-based attacks using simulated annealing that significantly reduce query complexity while maintaining attack effectiveness [31].

No-box attacks represent a recently introduced paradigm that requires neither model access nor queries. Lu et al. [33] demonstrated that domain-specific priors—such as natural human motion dynamics for skeleton-based action recognition—can guide attack generation without any interaction with the target model. This paradigm is particularly concerning for deployed HAR systems where attackers can leverage publicly available motion capture datasets to craft transferable attacks.

2.1.2. Attack Objectives

Untargeted attacks aim to cause any misclassification, formally expressed as achieving $f(x^{adv}) \neq y$. For HAR systems, an untargeted attack might cause a "walking" activity to be misclassified as any other activity.

Targeted attacks force a specific misclassification $f(x^{adv}) = y_{target}$ to an adversary-chosen class. In healthcare monitoring contexts, targeted attacks pose severe risks—for example, causing "fall" activities to be classified as "sitting" could disable emergency response systems [5].

Universal attacks seek a single perturbation pattern effective across multiple inputs. Recent work has demonstrated universal attacks on millimeter-wave HAR systems achieving greater than 95% success rates with a single learned perturbation pattern [34]. Universal attacks are particularly threatening for deployed systems because they can be pre-computed offline and applied in real-time without iterative optimization.

2.1.3. Attack Timing

Evasion attacks operate at test time, manipulating inputs to an already-deployed model. The majority of HAR adversarial attacks fall into this category.

Poisoning attacks corrupt training data to influence model behavior, either degrading overall performance or introducing backdoors that trigger on specific inputs. Label flipping attacks against wearable HAR systems have been demonstrated to successfully manipulate MLP, Decision Tree, Random Forest, and XGBoost classifiers during data collection phases [35]. Backdoor attacks on skeleton-based recognition can embed triggers using infrequent, imperceptible actions that activate malicious behavior during inference [36]. For autoregressive models, data poisoning attacks can corrupt sequential dependencies, causing cascading prediction failures that persist across multiple forecast horizons [37].

2.2. Perturbation Constraints for Temporal Data

Adversarial perturbations are typically constrained to preserve imperceptibility. The most common formulation bounds perturbations by L_p norms:

$$\|\delta\|_p \leq \epsilon \tag{2}$$

where $\delta = x^{adv} - x$ denotes the perturbation and ϵ the perturbation budget.

However, L_p norms developed for image perturbations are often inadequate for time series data. Belkhouja et al. [38] demonstrated that Dynamic Time Warping (DTW) provides a more appropriate distance metric for time series because DTW accounts for temporal alignment variations that are natural in human activity data. Their DTW-AR framework achieved superior imperceptibility compared to Euclidean-norm-constrained attacks while maintaining high attack success rates.

Beyond distance metrics, time series attacks must respect additional constraints absent in image attacks:

Signal continuity: Sensor measurements evolve continuously; abrupt discontinuities in accelerometer or gyroscope readings are physically implausible and easily detectable. Pialla et al. [39,40] introduced smoothness constraints using Gaussian process priors to ensure perturbations maintain natural signal characteristics.

Measurement bounds: Physical sensors have finite measurement ranges. Accelerometers typically measure $\pm 2g$ to $\pm 16g$, while gyroscopes measure ± 250 to ± 2000 degrees per second. Perturbations that exceed these bounds are physically unrealizable.

Inter-sensor correlations: In multi-sensor systems, readings from different sensors (e.g., accelerometer and gyroscope on the same device) are physically correlated. Perturbations that violate these correlations may be detectable by consistency checking [41].

Energy constraints: For attacks requiring signal injection into wearable devices, battery limitations constrain the total perturbation energy that can be sustained over time.

Probabilistic output constraints: For probabilistic forecasting models that output prediction distributions rather than point estimates, adversarial perturbations must consider both mean predictions and uncertainty estimates. Dang-Nhu et al. [23] demonstrated that attacks on autoregressive models can exploit the sequential nature of predictions, where perturbations at early time steps propagate through the forecast horizon with amplifying effects.

2.3. Canonical Attack Methods

Several foundational attack methods have been adapted across domains:

Fast Gradient Sign Method (FGSM) [1] generates perturbations in a single gradient step:

$$x^{adv} = x + \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}(f(x), y)) \tag{3}$$

where $\mathcal{L}$ denotes the classification loss. While computationally efficient, FGSM attacks on time series often produce perturbations with unnatural spike patterns that violate temporal smoothness [39].

Projected Gradient Descent (PGD) [2] extends FGSM through iterative optimization with projection onto the feasible perturbation set:

$$x_{t+1}^{adv} = \Pi_{x+\mathcal{S}}\left(x_t^{adv} + \alpha \cdot \text{sign}(\nabla_x \mathcal{L}(f(x_t^{adv}), y))\right) \tag{4}$$

where Π projects onto the ϵ-ball around x and α is the step size. PGD attacks on HAR systems typically require 10-40 iterations to converge, with attack success rates increasing monotonically with iteration count until saturation [3].

Carlini-Wagner (C&W) attack [42] formulates adversarial example generation as constrained optimization:

$$\min_{\delta} \|\delta\|_2 + c \cdot g(x + \delta) \tag{5}$$

where g is an objective function encouraging misclassification and c balances perturbation magnitude against attack success. C&W attacks typically produce smaller perturbations than PGD but require significantly more computation, making them less practical for real-time attacks on streaming sensor data.

Zeroth-Order Optimization (ZOO) [43] enables black-box attacks by estimating gradients through finite differences:

$$\hat{\nabla}_i f(x) \approx \frac{f(x + h \cdot e_i) - f(x - h \cdot e_i)}{2h} \tag{6}$$

where e_i is the i-th standard basis vector and h is a small constant. ZOO attacks require $O(d)$ queries per gradient estimate for d-dimensional inputs, which can be prohibitive for long time series.

Square Attack and Variants [31] provide query-efficient black-box attacks through random search in square-shaped regions. Liu et al. demonstrated that combining square-based perturbations with simulated annealing achieves competitive attack success rates on time series classification while requiring significantly fewer queries than gradient estimation methods. This approach is particularly relevant for attacking deployed HAR systems where query budgets are constrained.

2.4. *Unique Characteristics of Sensor-Based Time Series*

Sensor-based HAR data exhibits several properties that fundamentally distinguish it from image data and necessitate specialized attack methodologies:

Temporal dependencies: Activity patterns unfold over time with characteristic dynamics. Walking exhibits periodic patterns at approximately 1-2 Hz, while transitions between activities involve distinctive acceleration profiles. Perturbations must maintain temporal coherence to remain imperceptible.

Multi-sensor configurations: Modern wearable devices incorporate multiple sensors (accelerometer, gyroscope, magnetometer, barometer) that provide complementary information. Attack effectiveness varies significantly depending on which sensors are perturbed. Kurniawan et al. [29] demonstrated that adversarial attacks can succeed by compromising only one of three sensor devices in multi-modal systems.

Body placement sensitivity: The same activity produces different sensor signatures depending on device placement (wrist, chest, ankle, hip). Cross-location transferability of attacks varies from 0% to over 80% depending on the specific location pair [3].

User variability: Different users perform the same activities with individual variations in speed, amplitude, and style. Attacks that transfer across users are more concerning than user-specific attacks, but cross-user transferability is generally lower than within-user attack success.

Sampling rate effects: HAR systems operate at sampling rates from 20 Hz to 200 Hz. Higher sampling rates provide more temporal resolution but also increase the attack surface by providing more individual samples that can be perturbed.

Anomaly detection context: Beyond classification, time series systems increasingly employ anomaly detection for identifying unusual patterns. Tariq et al. [22] demonstrated that anomaly detection models exhibit distinct adversarial vulnerabilities, where perturbations can cause either false negatives (missed anomalies) or false positives (spurious alerts), each with different operational consequences.

2.5. *Emerging Architectural Paradigms*

The rapid evolution of deep learning architectures introduces novel vulnerability surfaces that extend beyond conventional CNN and LSTM models:

2.5.1. Spiking Neural Networks for Temporal Data

Spiking Neural Networks (SNNs) are increasingly deployed for time series classification on edge devices due to their energy efficiency and natural temporal processing capabilities. However, Hutchins et al. [30] demonstrated that SNNs exhibit distinct adversarial susceptibility patterns compared to conventional architectures. Black-box attacks on SNNs for time series data achieve high success rates while exploiting the spike-timing-dependent nature of these networks. The discrete, event-driven nature of SNNs creates attack surfaces absent in continuous-valued networks, necessitating specialized adversarial analysis for neuromorphic deployments.

2.5.2. Large Language Models for Time Series

The application of Large Language Models (LLMs) to time series analysis represents an emerging paradigm with significant security implications. Foundation models including TimeGPT, GPT-4, LLaMA, and Mistral variants have demonstrated competitive performance on forecasting and anomaly detection tasks [10,11]. However, these models inherit vulnerabilities from their language model foundations while introducing temporal-specific attack surfaces:

Prompt-based vulnerabilities: LLM-based time series models often rely on textual prompts to specify forecasting tasks, creating injection attack vectors absent in traditional neural architectures.

Tokenization artifacts: The discretization of continuous time series into token sequences introduces quantization boundaries that can be exploited by adversarial perturbations.

Context window limitations: Fixed context lengths may truncate relevant historical information, creating blind spots exploitable by adversaries who can manipulate which portions of time series enter the context window.

Xiao et al. [12] introduced learning-based attacks specifically targeting temporal forecasting models, demonstrating that adversarial perturbations can manipulate predictions along three dimensions: *directional* (bullish vs. bearish bias), *amplitudinal* (magnitude of predicted changes), and *temporal* (timing of predicted events). These attacks achieve high success rates while maintaining statistical properties that evade conventional anomaly detection.

Alnegheimish et al. [10] evaluated LLMs as anomaly detectors for time series, revealing that while these models achieve competitive detection performance, their reasoning processes can be manipulated through adversarial examples that exploit the models' reliance on pattern matching rather than domain-specific physical constraints. This finding suggests that LLM-based temporal systems may be particularly vulnerable to attacks that appear statistically normal but violate domain semantics.

2.6. Taxonomy of Temporal Adversarial Attacks

We organize the surveyed literature along four dimensions that capture the key design choices in temporal adversarial attacks. Critically, we distinguish between *on-body sensing* and *device-free RF sensing* paradigms, which present fundamentally different attack surfaces despite both processing temporal data.

Dimension 1: Target domain and sensor modality.

We partition sensing modalities into two fundamental categories based on their physical operating principles:

Category A: On-body inertial and physiological sensing. These systems require physical contact with the subject and measure mechanical or electrical properties directly:

- *Wearable IMU systems:* Attacks targeting accelerometer, gyroscope, and magnetometer sensors on smartphones and dedicated wearables. Attack vectors include direct sensor manipulation, firmware compromise, and electromagnetic interference. Perturbations must respect rigid-body motion dynamics and inter-sensor correlations.
- *Medical physiological sensors:* Attacks on ECG, EEG, EMG, and other bioelectrical signal classifiers. Clinical plausibility constraints require perturbations to maintain physiological realism [44].

Category B: Device-free RF and vision-based sensing. These systems operate without physical contact, sensing through electromagnetic wave propagation or optical capture:

- *WiFi CSI sensing:* Attacks on channel state information-based recognition systems. These face fundamentally different constraints than IMU attacks: signal processing pipelines include non-differentiable operations (Hampel filtering, phase sanitization) that prevent direct gradient-based optimization [4]. Physical attacks require RF signal injection synchronized with legitimate WiFi traffic.
- *Millimeter-wave and FMCW radar:* Attacks on 60-77 GHz sensing systems. Attack vectors include active signal injection and passive reflection manipulation using meta-material tags [8]. Propagation-path constraints differ fundamentally from wearable sensor perturbations.
- *Skeleton-based recognition:* Attacks on pose estimation and skeleton sequence classifiers. Perturbations must maintain anatomical plausibility and natural motion dynamics [45].

Category C: Financial and industrial time series. These systems process non-physical temporal data with domain-specific constraints:

- *Financial time series:* Attacks on trading systems and forecasting models. Market microstructure constraints (tick sizes, trading hours, liquidity) bound feasible perturbations [46].
- *Industrial process data:* Attacks on predictive maintenance and process control systems. Physical process dynamics constrain perturbation feasibility.

Category D: Sequential decision systems.

- *Reinforcement learning agents:* Attacks including state perturbations, reward poisoning, and adversarial policy training. Sequential dependencies allow perturbations to compound across time steps.

Dimension 2: Perturbation strategy.

- *Gradient-based:* Methods adapting FGSM, PGD, and related techniques for temporal data.
- *Optimization-based:* C&W-style attacks with temporal constraints.
- *Generative:* GAN-based [47] and diffusion model approaches producing natural-appearing perturbations.
- *Search-based:* Evolutionary algorithms, tree search [48], square-based random search [31], and RL policies [49] for discrete perturbation selection.
- *Frequency-domain:* Attacks manipulating Fourier or wavelet representations of time series.
- *Learning-based:* Neural network-based attack generators that learn perturbation strategies from data, enabling rapid attack generation without iterative optimization at test time [12].

Dimension 3: Temporal scope.

- *Point-wise:* Independent perturbations at each time step.
- *Window-wise:* Coherent perturbations over contiguous windows.
- *Sparse:* Perturbations targeting only critical time steps identified through attention or gradient analysis.
- *Global:* Single perturbation pattern applied across entire sequences (universal attacks).
- *Causal/sequential:* Perturbations designed to exploit temporal dependencies, where early perturbations influence predictions at later time steps through autoregressive or recurrent mechanisms [23].

Dimension 4: Physical realizability.

- *Digital-only:* Attacks assuming direct access to model inputs.
- *Physically constrained:* Attacks respecting sensor bounds and signal smoothness but validated only in simulation.
- *Physically validated:* Attacks demonstrated through hardware injection or real-world deployment [4,50].

Table 3 summarizes the key differences between sensing modality categories, highlighting the distinct attack surfaces and constraint types that necessitate modality-specific adversarial analysis.

Table 3. Comparison of sensing modality categories and their adversarial attack characteristics. On-body and device-free sensing present fundamentally different attack surfaces requiring distinct methodological approaches.

Characteristic	On-body IMU	WiFi CSI	mmWave Radar	Skeleton-based
Signal type	Mechanical motion	RF propagation	RF reflection	Visual/depth
Processing pipeline	Differentiable	Non-differentiable	Partially diff.	Differentiable
Physical attack vector	Device manipulation	Signal injection	Passive reflection	Motion modification
Key constraints	Rigid-body dynamics	CSI sanitization	Propagation physics	Anatomical limits
Cross-sensor transfer	0–80% variable	Subcarrier-dependent	Frequency-dependent	Joint-dependent
Physical validation	Limited	Demonstrated	Demonstrated	Limited

3. Time Series Adversarial Attacks

This section comprehensively reviews adversarial attacks on time series systems, organized by application domain with particular emphasis on sensor-specific vulnerabilities in HAR systems. Tables

4 and 5 provide detailed comparisons of methods, datasets, attack characteristics, and quantitative performance.

3.1. Human Activity Recognition Attacks

Human activity recognition systems process inertial measurement unit data from wearable devices to identify user activities such as walking, running, sitting, and climbing stairs [24,25]. These systems are critical for healthcare monitoring, fitness tracking, elder care, and smart home automation. The widespread deployment of HAR in safety-critical applications—particularly fall detection for elderly individuals and activity monitoring for patients with chronic conditions—makes adversarial vulnerabilities in these systems a matter of significant concern.

3.1.1. Foundational Attacks on Time Series Classification

Fawaz et al. [13] pioneered time series adversarial attacks by adapting FGSM and the Basic Iterative Method (BIM) from images to the UCR Archive comprising 85 univariate datasets. They demonstrated that state-of-the-art deep learning time series classifiers—including ResNet, Fully Convolutional Networks (FCN), and InceptionTime—are highly vulnerable to adversarial perturbations. Critically, they observed that perturbations on time series are more perceptible than image perturbations due to the one-dimensional nature of the data, motivating subsequent research on imperceptibility constraints specific to temporal data.

Karim et al. [47] introduced Adversarial Transformation Networks (ATNs) using GANs for black-box time series attacks. By training surrogate models through knowledge distillation, their approach successfully fooled both Dynamic Time Warping (DTW) classifiers and deep learning models including FCN across 42 UCR datasets without requiring gradient access. The success of transfer-based attacks demonstrated that time series classifiers share vulnerable features across different architectures, suggesting fundamental weaknesses in learned representations rather than architecture-specific vulnerabilities.

3.1.2. Sensor-Specific Vulnerability Analysis

A critical dimension of HAR adversarial attacks that has received insufficient attention is the differential vulnerability of individual sensor modalities. The ADAR framework by Sah and Ghasemzadeh [3] provided the first systematic analysis of adversarial attacks on wearable HAR systems, revealing several important findings:

Dramatic accuracy degradation: Under FGSM attacks, DNN-based HAR classifiers experienced accuracy drops from 95.1% to 3.4%, while CNN models dropped from 93.1% to 16.8%. These degradation levels significantly exceed typical drops observed in image classification, suggesting that time series classifiers may be inherently more vulnerable to adversarial perturbations.

Sensor-specific vulnerability patterns: Different sensor modalities exhibit distinct vulnerability profiles. Accelerometer-only attacks achieve different success rates than gyroscope-only or magnetometer-only attacks, with the relative vulnerability depending on the target activity and classifier architecture.

Four-dimensional transferability: Adversarial examples transfer across: (1) different ML model architectures, (2) different users, (3) different sensor body locations, and (4) different datasets. Cross-model transferability is highest (often exceeding 70%), while cross-location transferability varies dramatically from near 0% to over 80% depending on the specific location pair.

Kurniawan et al. [29] extended this analysis to multi-modal sensor systems, demonstrating that adversarial attacks can succeed by compromising only one of three sensor devices. Using GAN-based perturbation generators with conditional estimators for non-hacked sensor values, they achieved 50-100% attack success rates on the MHealth dataset depending on the target activity pair. This finding has profound implications for the security of multi-sensor HAR systems: an adversary need not compromise all sensors to defeat the recognition system.

Their follow-up work [41] demonstrated that while attacks are effective, the compromised sensor can be detected with high accuracy by analyzing inter-sensor correlations. This suggests a promising defense direction based on sensor consistency verification, though it also implies that sophisticated attackers may need to coordinate perturbations across multiple sensors to evade detection.

3.1.3. Advanced Attack Methodologies

Recent work has developed increasingly sophisticated attack methodologies that address limitations of early approaches:

Black-box attacks via tree search: Ding et al. [48] introduced BlackTreeS, a black-box attack framework using tree search to identify influential positions in time series and estimate gradients without model access. By formulating adversarial example generation as a search problem over the space of perturbation locations and magnitudes, BlackTreeS achieved state-of-the-art black-box attack success rates on the UCR Archive while requiring significantly fewer queries than prior gradient-estimation methods.

DTW-based perturbation constraints: Belkhouja et al. [38] argued that Euclidean distance is inappropriate for measuring perturbation magnitude in time series because it ignores temporal alignment. Their DTW-AR framework uses Dynamic Time Warping as the similarity measure, generating adversarial examples that are perceptually similar under DTW while successfully fooling classifiers. This approach achieved superior imperceptibility ratings in human evaluation studies while maintaining high attack success rates.

Multi-objective attack optimization: Wang et al. [51] introduced TSFool, which formulates attack generation as multi-objective optimization balancing attack success against a novel "Camouflage Coefficient" measuring imperceptibility. By explicitly trading off these objectives, TSFool generates attacks that are both effective and difficult to detect through statistical analysis of perturbation patterns.

Smoothness-constrained perturbations: Pialla et al. [39,40] observed that gradient-based attacks on time series often produce perturbations with unnatural spike patterns that are easily detectable. They introduced smoothness constraints using Gaussian process priors that ensure perturbations maintain the natural continuity of sensor signals. Their experiments demonstrated that smooth perturbations achieve comparable attack success rates while being significantly less detectable by anomaly detection methods.

3.1.4. Skeleton-Based Action Recognition Attacks

Skeleton-based action recognition has received substantial attention due to the growing deployment of pose estimation systems in surveillance, gaming, and human-computer interaction. Unlike sensor-based HAR where perturbations must respect physical measurement constraints, skeleton attacks operate in a higher-level representation space where perturbations must maintain anatomical plausibility.

Wang et al. [45] introduced SMART, demonstrating that skeleton action recognition models are vulnerable to imperceptible joint perturbations. By constraining perturbations to maintain bone length consistency and joint angle limits, SMART generates adversarial skeletons that appear natural to human observers while causing misclassification.

Diao et al. [52] presented the extended BASAR framework achieving 100% attack success rate across all tested models (ST-GCN, MS-G3D, SGN, CTR-GCN, FR-HEAD) through decision-based black-box attacks. The key insight is that natural human motion lies on a low-dimensional manifold; by restricting perturbations to this manifold, attacks remain imperceptible while effectively fooling classifiers. This work introduced Mixed Manifold-based Adversarial Training (MMAT) as a defense mechanism that improves robustness by augmenting training with both on-manifold and off-manifold adversarial examples.

Lu et al. [33] proposed a "hard no-box" attack paradigm requiring neither model access nor queries. By leveraging skeleton-motion-informed (SMI) gradients derived from domain knowledge about natural human motion dynamics, their approach generates effective adversarial examples without any interaction with the target model. The SMI gradients are computed using a motion manifold estimated through contrastive learning on publicly available motion capture datasets, demonstrating that domain knowledge alone can guide effective attacks.

Diao et al. [53] introduced TASAR, the first dedicated transfer-based attack for skeleton recognition using Dual Bayesian optimization to smooth model posteriors and improve cross-model transferability. This work also released RobustBenchHAR, a comprehensive benchmark comprising 7 models, 10 attack methods, 3 datasets (NTU RGB+D 60, NTU RGB+D 120, Kinetics-400), and 2 defense models, providing the first standardized evaluation framework for skeleton-based HAR robustness.

Critical analysis: A notable finding across skeleton attack research is that restricting perturbations to physically plausible modifications (e.g., bone lengths only) can actually improve both attack effectiveness and model robustness when used for adversarial training. Tanaka et al. [54] demonstrated that bone-length-only perturbations, despite constraining the attack surface to approximately 30 effective dimensions, achieve over 90% attack success rates. Remarkably, adversarial training with these constrained attacks improves both robustness and clean accuracy—contradicting the typical robustness-accuracy trade-off observed in image classification.

Table 4. Sensor-based HAR adversarial attacks: Methods, datasets, and quantitative performance. Red entries indicate newly added references verified through DOI/arXiv.

Method	Citation	Year	Dataset(s)	Type	Success Rate	Key Contribution
			Foundational Time Series Attacks			
FGSM/BIM-TS	[13]	2019	85 UCR	WB	High	First TS adversarial attacks
ATN	[47]	2020	42 UCR	BB	All 42	GAN-based transferable attacks
BlackTreeS	[48]	2023	UCR	BB	SOTA	Tree search black-box
DTW-AR	[38]	2023	UCR	WB	High	DTW-based imperceptibility
TSFool	[51]	2024	UCR	WB	High	Multi-objective optimization
			Sensor-Specific and Multi-Modal HAR Attacks			
ADAR	[3]	2019	PAMAP2, Opp.	WB	95%→3%	4-dimensional transferability
Partial Sensor	[29]	2022	MHealth	WB	50-100%	Single-sensor compromise
Sensor Detection	[41]	2024	MHealth	Defense	–	Compromised sensor detection
Label Flip	[35]	2022	UCI-HAR	Poison	Varies	Training-time poisoning
			Skeleton-Based Action Recognition Attacks			
SMART	[45]	2021	NTU	WB	High	Joint perturbation attacks
BASAR	[52]	2024	NTU, Kinetics	BB	100%	Manifold-based black-box
No-Box SMI	[33]	2023	NTU	No-box	High	Motion-informed gradients
TASAR	[53]	2025	NTU, K400	Transfer	High	Bayesian transfer + benchmark
Bone Length	[54]	2022	NTU	WB	>90%	30-dim constrained attack
PSBA	[36]	2024	NTU	Backdoor	High	Physical skeleton backdoor
			Gait Recognition Attacks			
Dictionary	[55]	2023	ZJU-gaitacc	BB	Varies	Pre-computed gait patterns
BLG	[56]	2025	CASIA-B	BB	94.33%	Latent-space perturbation

3.1.5. Application-Specific HAR Vulnerabilities

Gait recognition and authentication: Biometric gait authentication presents particularly concerning vulnerabilities because successful attacks can enable identity spoofing or denial of service. Kumar et al. [55] introduced dictionary attacks for IMU-based gait authentication, demonstrating that pre-computed gait patterns can be used as presentation attacks against biometric systems. The attack success depends on the diversity of the dictionary and the specificity of the authentication threshold.

More sophisticated attacks leverage generative models. The BLG (Black-box-Latent-GEI) attack [56] achieves 94.33% attack success rate on the CASIA-B benchmark using latent-space perturbations without requiring target model queries. The encoder-decoder framework with PerturbGen and AdvHelper components generates adversarial gait silhouettes that maintain visual plausibility while causing authentication failures.

Fall detection systems: Fall detection represents a critical healthcare application where adversarial attacks could have life-threatening consequences. Sakka et al. [5] examined security vulnerabilities in HAR systems for medical IoT, identifying that fall detection models are susceptible to attacks causing falls to be classified as normal activities (potentially delaying emergency response) or normal activities to be classified as falls (causing alert fatigue that may lead to ignored genuine emergencies).

Healthcare monitoring: Beyond fall detection, adversarial attacks on HAR systems can manipulate medication adherence monitoring, physical therapy compliance tracking, and chronic disease management applications. The emerging deployment of HAR in clinical settings amplifies the importance of adversarial robustness research.

3.2. WiFi and Radar-Based HAR Attacks

WiFi channel state information and radar systems enable contactless sensing for human activity recognition, gesture recognition, and vital sign monitoring. Their vulnerability to adversarial attacks poses significant privacy and security risks, as these systems are increasingly deployed in smart homes, healthcare facilities, and security-sensitive environments.

3.2.1. WiFi CSI Attacks

WiFi-based sensing exploits the fact that human activities cause characteristic perturbations in wireless signal propagation. Channel State Information (CSI) captures fine-grained amplitude and phase variations across OFDM subcarriers, enabling recognition of activities, gestures, and even vital signs without dedicated sensors. The attack surface for WiFi sensing systems differs fundamentally from wearable HAR: rather than perturbing sensor readings, attackers must modify the wireless propagation environment or inject adversarial signals.

Zhou et al. [4] pioneered physically realizable WiFi gesture attacks with WiAdv, achieving greater than 70% average success rates on the Widar3.0 dataset. The key technical challenge addressed by WiAdv is that CSI processing involves non-differentiable operations (such as Hampel filtering for outlier removal) that prevent direct gradient-based attack optimization. WiAdv solves this through a differentiable approximation of the signal processing pipeline, enabling end-to-end gradient computation from classifier output to transmitted signal.

Li et al. [50] demonstrated physical attacks via WiFi packet preamble manipulation, achieving 90.47% activity recognition attack success and 83.83% authentication attack success rates. By perturbing pilot symbols in the IEEE 802.11 physical layer, adversaries can mislead WiFi sensing systems while maintaining communication functionality. This represents a major step toward practical attacks because pilot symbol manipulation can be implemented using software-defined radio without requiring physical proximity to the victim.

Xu et al. [57] introduced WiCAM, which uses attention mechanisms to identify critical subcarriers for WiFi CSI attacks. By focusing perturbations on attended subcarriers only, WiCAM minimizes bit error rate (BER) impact while maintaining attack effectiveness. This attention-guided approach achieves less than 50% accuracy reduction on HAR tasks with 77.78% BER reduction compared to uniform perturbation approaches.

Sharma et al. [58] presented Wi-Spoof, demonstrating power manipulation attacks using pseudo-PWM techniques. By modulating transmission power in patterns that create adversarial CSI signatures, Wi-Spoof achieves 93% targeted misclassification through physically realizable power modulation. The attack requires only control over a WiFi transmitter in the sensing environment, making it practical for insider threat scenarios.

Huang et al. [59] introduced IS-WARS, which exploits intentional interference from coexisting wireless protocols including ZigBee, Bluetooth, and LTE-U. By timing interference to coincide with WiFi sensing measurements, IS-WARS stealthily degrades WiFi HAR systems operating in the 2.4GHz band without requiring sophisticated signal generation capabilities.

Yang et al. [60] proposed SecureSense, the first comprehensive defense framework for WiFi-based HAR. SecureSense achieves consistent predictions regardless of adversarial input by learning transformation-invariant representations. Cao et al. [61] introduced Selective Adversarial Training (SAT), demonstrating that targeted adversarial training on vulnerable activity classes can improve robustness while maintaining clean accuracy on other classes.

3.2.2. Millimeter-Wave and Radar Attacks

Millimeter-wave (mmWave) radar systems operating at 60-77 GHz offer high-resolution sensing for gesture recognition, vital sign monitoring, and human tracking. The shorter wavelength compared to WiFi enables finer spatial resolution but also creates different attack opportunities.

Xie et al. [34] introduced the first universal targeted attack on mmWave HAR systems, achieving greater than 95% success rates with a single learned perturbation pattern. The universal nature of the attack is particularly concerning: once computed offline, the same perturbation reliably misleads recognition across different samples and even different users.

Chen et al. [8] presented MetaWave, a passive mmWave attack using commercially available meta-material tags. Unlike active attacks requiring signal transmission, MetaWave uses passive reflection to perturb mmWave sensing at 10-100× lower cost than existing attack methods. The differentiable RF simulator enables optimization of tag placement and orientation, achieving 97% accuracy on range estimation attacks, 96% on angle estimation, and 91% on speed estimation.

Xu et al. [62] demonstrated TileMask, the first passive-reflection-based adversarial attack against DNN-based radar object detection. Using 3D-printed objects covered with metal foils, TileMask creates adversarial reflections that cause false positives or negatives in object detection. Requiring only 2 adversarial objects for successful attacks, the method offers excellent stealthiness as objects can be disguised as ordinary car signs or roadside structures.

Ozbulak et al. [6] demonstrated that radar-based HAR systems are vulnerable to a particularly striking attack: prediction flipping through perturbation of only the input padding without touching the actual action frames. This "padding attack" reveals fundamental vulnerabilities in how temporal CNNs process variable-length inputs, suggesting that architectural choices in model design can create unexpected attack surfaces.

Kuzlu et al. [63] provided the first security analysis for mmWave beamforming in 5G/6G networks, evaluating defensive distillation and adversarial retraining as defense mechanisms. Their analysis reveals that communication system security and sensing system security are interrelated, as attacks on beamforming can degrade both communication quality and sensing accuracy.

Table 5. WiFi and radar-based HAR adversarial attacks. Red entries indicate newly verified references.

Method	Citation	Year	Dataset/System	Type	Success	Key Contribution
			WiFi CSI Attacks			
WiAdv	[4]	2022	Widar3.0	Physical	>70%	First physical WiFi attack
Pilot Symbol	[50]	2024	Custom	Physical	90.47%	PHY-layer perturbation
WiCAM	[57]	2022	WiFi CSI	WB	High	Attention-guided subcarrier
Wi-Spoof	[58]	2025	CSI	Physical	93%	Power modulation attack
IS-WARS	[59]	2022	WiFi 2.4GHz	Interf.	Significant	Protocol interference
SecureSense	[60]	2024	WiFi	Defense	–	Transformation-invariant
SAT	[61]	2024	WiFi	Defense	–	Selective adv. training
			Millimeter-Wave and Radar Attacks			
Universal mmWave	[34]	2023	mmWave HAR	Universal	>95%	First universal mmWave
MetaWave	[8]	2023	mmWave	Passive	97%/96%	Meta-material tags
TileMask	[62]	2023	Radar	Passive	High	3D-printed adversarial
Padding Attack	[6]	2021	Radar HAR	WB	High	Padding-only perturbation
mmWave 5G	[63]	2023	5G/6G	Analysis	–	Beamforming security

3.3. Video and Temporal Sequence Attacks

Video adversarial attacks target action recognition, object tracking, and video understanding systems. These attacks must maintain temporal consistency across frames while achieving imperceptibility in both spatial and temporal dimensions.

Chen et al. [64] achieved approximately 98% targeted attack success on UCF-101 by appending adversarial frames to videos rather than perturbing existing content. This attack paradigm is particularly concerning because it does not modify the original video content, potentially evading detection methods that analyze perturbation statistics within frames.

Wei et al. [65] improved cross-model transferability by exploiting temporal translation invariance, achieving 61.56% transfer rates on Kinetics-400. Their subsequent work [66] demonstrated cross-modal transfer from image models to video classifiers with 77.88% black-box success, and Wei et al. [67] extended this with adaptive multi-layer feature ensemble for improved transferability.

Kim et al. [68] generated universal adversarial perturbations using image models that break temporal consistency, achieving 70.79% average fooling rate. The insight that image-based UAPs can transfer to video models by disrupting temporal patterns suggests fundamental vulnerabilities in how video classifiers integrate spatial and temporal information.

Hwang et al. [69] discovered that action recognition models tolerate frame randomization but adversarial perturbations do not survive such randomization. This observation enables the first training-free defense for 3D CNNs through temporal shuffling at inference time.

Zheng et al. [36] introduced PSBA, the first physical backdoor attack using infrequent, imperceptible actions as triggers in skeleton data. Unlike digital backdoors that embed pixel patterns, PSBA embeds behavioral triggers that can be activated through natural human motion.

3.4. Medical and Financial Time Series Attacks

3.4.1. Medical Time Series

Han et al. [44] demonstrated the first physiologically plausible ECG adversarial attack, achieving 74% attack success while maintaining clinical plausibility. The constraint of physiological plausibility is critical for medical attacks: perturbations that produce obviously abnormal waveforms would be detected by clinicians or quality control algorithms.

Chen et al. [70] proposed CASLCNet combining Lipschitz constraints with channel activation suppression for defending ECG classifiers against PGD, FGSM, C&W, and SAP attacks. Shao et al. [71] introduced CardioDefense using adversarial distillation training. Wiedeman and Wang [72] developed decorrelative network architectures for robust ECG classification without requiring adversarial training.

For EEG-based brain-computer interfaces, Meng et al. [73] demonstrated backdoor vulnerabilities using narrow period pulses as triggers. Their follow-up work [74] presented filtering-based evasion attacks effective across ERN, MI, and P300 paradigms. Wu et al. [75] provided the first comprehensive adversarial robustness benchmark for BCIs, evaluating 9 defense approaches across 3 CNN architectures and 2 EEG datasets.

3.4.2. Financial Time Series

Financial time series present unique attack challenges due to market constraints, regulatory oversight, and the high stakes of manipulation.

Fursov et al. [46] conducted the first study on adversarial attacks against transaction record classifiers. Xie et al. [76] demonstrated that single-word substitutions in financial tweets can manipulate stock prediction models.

Liu et al. [77] discovered that attacks on multivariate forecasting can impact target series through sparse modifications to correlated series, and proposed randomized smoothing defense. Kulkarni et al. [9] introduced Directional, Amplitudinal, and Temporal targeted attacks for forecasting. Liu et al. [78] demonstrated black-box attacks degrading TimeGPT, GPT-4, LLaMa, and Mistral forecasting accuracy.

Rizvani et al. [7] formalized “ephemeral perturbations” for algorithmic trading—transient price manipulations that induce suboptimal trading decisions while remaining statistically undetectable. This threat model is particularly relevant for high-frequency trading where decisions must be made on microsecond timescales.

4. Reinforcement Learning Adversarial Attacks

This section comprehensively reviews adversarial attacks targeting reinforcement learning systems and the emerging paradigm of using RL for attack generation. RL systems face unique vulnerabilities arising from their sequential decision-making nature: perturbations at early time steps can compound through the decision trajectory, and attacks can target the learning process itself rather than just inference.

4.1. Attacks on RL Agent Observations

4.1.1. State Perturbation Attacks

State observation attacks perturb the observations received by an RL agent, causing it to take suboptimal or dangerous actions. The State-Adversarial MDP (SA-MDP) framework by Zhang et al. [79] provides theoretical foundations for this attack model, demonstrating that their MAD (Maximal Action Difference) attack reduces rewards by 60-90% on undefended agents across Atari and MuJoCo environments.

Sun et al. [80] demonstrated sparse strategically-timed attacks achieving agent failure in just 1-5 perturbation steps on Atari, MuJoCo, and TORCS environments. The insight that attacks need only perturb observations at critical decision points dramatically reduces the attacker’s required capabilities.

Zhang et al. [81] demonstrated the existence of optimal adversaries under bounded ℓ_p perturbations through their ATLA framework. Oikarinen et al. [82] introduced RADIAL-RL with adversarial loss-based robust training compatible with DQN, A3C, and PPO.

Korkmaz and Brown-Cohen [83] proposed detecting adversarial directions using local quadratic approximation of the value function. This attack-agnostic defense identifies perturbations that cause large changes in estimated value, providing robustness without requiring specific attack knowledge.

Recent work has extended state attacks to diffusion-based policies. Liu et al. [84] presented the first comprehensive study of attacks on diffusion policies for robot control, demonstrating vulnerabilities in both digital and physical attack scenarios.

4.1.2. Multi-Agent Adversarial Policies

Gleave et al. [85] demonstrated that adversarial policies trained via self-play can defeat state-of-the-art RL agents with greater than 95% success in multi-agent MuJoCo environments. Critically, adversarial policies appear random to human observers but create naturally adversarial observations for victim agents—they exploit learned policy weaknesses rather than injecting artificial perturbations.

Liu and Lai [86] demonstrated mixed attack strategies effective without prior environment knowledge. Ma et al. [87] extended adversarial policies to partial observability settings with SUB-PLAY, reducing superhuman-level Go AI winning rate to approximately 20% even when adversaries observe only partial game state.

Xue et al. [88] proposed ADMAC for robust multi-agent communication, estimating message reliability to defend against communication channel attacks. Liang et al. [89] addressed the robustness-performance trade-off through non-dominated policy optimization.

4.2. *Reward and Environment Poisoning*

Unlike observation attacks that operate at inference time, poisoning attacks corrupt the learning process itself. Reward poisoning modifies the reward signal during training, while environment poisoning alters state transitions or initial conditions.

Rakhsha et al. [90] introduced theoretical frameworks for reward poisoning in tabular MDPs, analyzing sample complexity bounds for successful manipulation.

Xu et al. [91] demonstrated the first universal black-box reward poisoning attack for offline RL, manipulating apparent rewards of low- and high-performing policies to corrupt offline learning. Rathbun et al. [92] proposed SleeperNets, universal backdoor attacks using dynamic reward poisoning that remain dormant until specific trigger conditions are met.

Wang et al. [93] introduced RLHFPoison, attacking reinforcement learning from human feedback in large language models through preference ranking manipulation. This attack exploits the increasingly common practice of using RLHF for model alignment.

4.3. *Using RL for Adversarial Attack Generation*

An emerging paradigm uses RL algorithms to generate adversarial attacks against non-RL systems. This approach naturally handles the sequential nature of temporal attacks and can encode complex constraints in the reward function.

Tsingenopoulos et al. [49] pioneered RL for black-box attacks with AutoAttacker, formulating adversarial example generation as a sequential decision process where the RL agent selects which input features to perturb.

Chen et al. [94] formulated black-box video attacks using RL to position text overlays, achieving approximately 90% success on UCF-101 and HMDB-51. Garcia et al. [95] used Multi-Objective MDPs to balance attack effectiveness against stealth requirements.

Song et al. [96] introduced RLVS, using RL with self-attention policies for key frame selection in sparse video attacks. The RL agent learns to identify which frames are most vulnerable to perturbation, reducing the attack budget while maintaining effectiveness.

Critical observation: Despite the natural fit between RL and temporal attack generation, surprisingly little work has applied RL-based attacks to HAR systems. The sequential nature of activity recognition, the need to balance imperceptibility constraints across sensor modalities, and the potential for transfer learning from simulation to physical attacks all suggest RL as a promising attack generation paradigm for future research.

Table 6. Reinforcement learning adversarial attacks and RL-based attack generation. Red entries indicate newly verified references.

Method	Citation	Year	Environment	Type	Success	Key Contribution
			State Observation Attacks			
SA-MDP/MAD	[79]	2020	Atari, MuJoCo	State	60-90% drop	Theoretical framework
Critical Point	[80]	2020	Atari, MuJoCo	State	1-5 steps	Sparse timing attacks
ATLA	[81]	2021	MuJoCo	State	Optimal	Learned adversary
RADIAL-RL	[82]	2021	Atari, MuJoCo	Defense	–	Adversarial loss training
Adv. Detection	[83]	2023	Atari, MuJoCo	Defense	–	Attack-agnostic detection
Diffusion Policy	[84]	2024	Robot Ctrl	State	High	Diffusion policy attacks
			Multi-Agent and Policy Attacks			
Adv. Policy	[85]	2020	Multi-agent	Policy	>95%	Via actions not states
MARL Mixed	[86]	2023	Multi-agent	Mixed	High	No prior knowledge
SUB-PLAY	[87]	2024	Go, MARL	BB Policy	80% win loss	Partial observability
ADMAC	[88]	2024	MARL	Defense	–	Robust communication
			Reward and Environment Poisoning			
Reward Poison	[90]	2020	Tabular MDP	Reward	Poly-time	Theoretical bounds
Offline Poison	[91]	2024	D4RL	BB Reward	Significant	Universal offline attack
SleeperNets	[92]	2024	Various	Backdoor	High	Dynamic reward backdoor
RLHFPoison	[93]	2024	LLMs	RLHF	High	Preference manipulation
			RL for Attack Generation			
AutoAttacker	[49]	2019	Image DNNs	BB	Varies	RL for black-box
BSC Video	[94]	2022	UCF, HMDB	BB	∼90%	RL text overlay
MORL Attack	[95]	2020	DRL agents	BB	Varies	Multi-objective RL
RLVS	[96]	2025	UCF, HMDB	BB	High	Self-attention keyframe

5. Cross-Cutting Themes

This section examines themes that span multiple attack domains and reveal fundamental patterns in temporal adversarial vulnerabilities. We provide critical analysis of physical realizability constraints, multi-modal fusion vulnerabilities, certified defense mechanisms, and emerging concerns about transformer architectures in HAR systems.

5.1. Physical Realizability Constraints

The gap between digital attacks validated in simulation and physically realizable attacks deployable in the real world represents a fundamental challenge in temporal adversarial research. While digital attacks demonstrate theoretical vulnerabilities, their practical relevance depends critically on whether perturbations can be realized through physical manipulation of sensor inputs or signal propagation. We provide a reconciled assessment of physical threat levels, acknowledging that reported success rates must be interpreted within their specific validation contexts.

5.1.1. Physical Attack Validation Hierarchy

We identify four levels of physical realizability in temporal adversarial attacks, ranging from purely theoretical to fully validated. Table 7 summarizes attack success rates stratified by validation level, revealing systematic differences that inform threat assessment.

Level 1: Digital-only attacks. The vast majority of HAR adversarial attacks operate purely in the digital domain, assuming direct access to model inputs. While valuable for understanding theoretical vulnerabilities, these attacks may not translate to practical threats against deployed systems. *Reported ASR: 85–98%; Estimated real-world ASR: Unknown (no physical validation).*

Level 2: Physically-constrained digital attacks. These attacks respect physical constraints such as sensor measurement bounds, signal continuity requirements, and inter-sensor correlations, but are validated only in simulation. The smooth perturbation methods of Pialla et al. [39,40] exemplify this approach, producing perturbations that maintain natural signal characteristics without requiring physical validation. *Reported ASR: 75–92%; Estimated degradation from physical constraints: 5–15%.*

Level 3: Hardware-in-the-loop validation. A small but growing body of work validates attacks using actual sensor hardware in controlled laboratory environments. Zhou et al.'s WiAdv [4] demonstrated WiFi attacks using software-defined radio equipment, achieving >70% success rates (95% CI: 65–78%) in laboratory settings. Li et al. [50] extended this to pilot symbol manipulation in IEEE 802.11 physical layers, reporting 90.47% activity recognition attack success. *Key limitation: Laboratory conditions minimize interference and environmental variability.*

Level 4: Field deployment validation. The most rigorous validation involves deploying attacks in real-world environments with uncontrolled variables. Chen et al.'s MetaWave [8] validated mmWave attacks using passive meta-material tags in realistic sensing scenarios, demonstrating 91–97% success rates for range/angle/speed manipulation. TileMask [62] validated passive radar attacks in outdoor automotive scenarios. *These represent the gold standard for threat assessment, though still limited to specific modalities and environments.*

Table 7. Physical attack success rates stratified by validation level. ASR ranges reflect variation across studies; confidence assessments incorporate methodological quality.

Validation Level	ASR Range	Studies	Confidence
Digital-only	85–98%	89	Theoretical
Constrained	75–92%	21	Moderate
Lab hardware	65–90%	12	High (lab)
Field deploy	70–97%	5	High (field)

5.1.2. Domain-Specific Physical Constraints

Different sensor modalities impose distinct physical constraints that shape attack design. We assess current evidence for physical attack feasibility across modalities, explicitly noting where validation gaps exist.

Wearable IMU sensors: Physical attacks on accelerometers and gyroscopes would require either direct device manipulation (replacing sensors, modifying firmware) or environmental perturbation (vibration injection, electromagnetic interference). *Critical gap: No peer-reviewed study has demonstrated hardware-in-the-loop validation of adversarial attacks on wearable IMU sensors.* The practical difficulty of these approaches limits real-world threat severity, though supply chain attacks during manufacturing remain a concern. We identify three potential attack vectors requiring investigation:

- *Electromagnetic interference injection:* Controlled EMI could potentially induce sensor measurement errors, but requires proximity and specialized equipment.
- *Acoustic/vibrational coupling:* Ultrasonic signals have been shown to affect MEMS sensors [97], but translation to adversarial HAR attacks remains undemonstrated.
- *Supply chain compromise:* Modified firmware or sensors could enable persistent attacks, but detection through hardware attestation may be feasible.

WiFi CSI sensing: Physical attacks have been validated through signal injection using software-defined radio equipment. WiAdv [4] achieved >70% success on Widar3.0, representing the most mature physical attack validation for HAR systems. The attack requires an adversary within transmission range (∼10–30m indoor) who can generate RF signals synchronized with legitimate WiFi traffic.

IS-WARS [59] demonstrated an alternative approach using intentional interference from coexisting protocols, requiring less sophisticated equipment but achieving lower success rates (estimated 40–60%). *Detection consideration: Anomalous RF signatures may be detectable by wireless intrusion detection systems, though this has not been systematically evaluated.*

Radar and mmWave: Physical attacks can employ either active signal injection or passive reflection manipulation. MetaWave [8] demonstrated that passive attacks using meta-material tags are 10–100× cheaper than active approaches while achieving 91–97% manipulation success (range: 97%, angle: 96%, speed: 91%). TileMask [62] showed that 3D-printed objects with metal foils can create adversarial reflections against automotive radar. *These attacks represent the highest-confidence physical threats due to field validation, but are specific to radar modalities and may be detectable through multi-sensor fusion consistency checks.*

Skeleton-based recognition: Physical attacks would require an adversary to modify their actual body movements to produce adversarial skeleton sequences. PSBA [36] introduced physical backdoor attacks using infrequent, imperceptible trigger actions. While theoretically possible, the constraint of maintaining natural-appearing motion while achieving misclassification significantly limits attack feasibility. *No study has demonstrated real-time adversarial motion generation that fools deployed skeleton-based HAR systems.*

5.1.3. Reconciled Threat Assessment

We reconcile apparently contradictory findings regarding physical attack threats by stratifying assessments across validation levels and modalities. Table 8 provides a unified view.

Table 8. Reconciled physical threat assessment by modality. Threat levels incorporate validation quality, required attacker capabilities, and detection feasibility.

Modality	Threat	Validation	Key Limitation
Wearable IMU	Low	None	No physical validation
WiFi CSI	Moderate	Lab (L3)	Detectability unknown
mmWave/Radar	High	Field (L4)	Modality-specific
Skeleton	Low	Limited	Motion constraints

Key insight: The apparent inconsistency between "lower immediate threat" and "high success rates (85–97%)" is resolved by recognizing that these assessments apply to different validation contexts. Digital attack success rates (85–98%) represent theoretical upper bounds under idealized conditions. Physically validated attacks show lower but still concerning success rates (70–97%), with the highest confidence for radar/mmWave modalities and significant uncertainty for wearable IMU systems.

Recommended threat model interpretation: For security-critical deployments, we recommend assuming Level 3 (hardware-in-the-loop) attack capabilities for WiFi and radar modalities, while acknowledging that wearable IMU attacks remain largely theoretical. Defense priorities should reflect this differentiated threat landscape.

5.1.4. Critical Analysis of Physical Realizability Claims

We observe that physical realizability claims in the literature often lack rigorous validation. Common limitations include:

Laboratory-only validation: Many papers claiming "physical" attacks validate only in controlled laboratory environments with ideal signal conditions, minimal interference, and cooperative experimental setups. Generalization to real-world deployments with environmental variability remains undemonstrated. *Recommendation: Future work should report environmental conditions and variability across trials.*

Simplified threat models: Physical attack papers often assume adversary capabilities (e.g., precise synchronization with legitimate signals, knowledge of target location) that may be difficult to achieve in practice. *Recommendation: Explicitly characterize required attacker resources and assess feasibility.*

Lack of detection analysis: Physical attacks may generate detectable artifacts in the radio frequency environment or sensor data. Few papers analyze whether attacks would be detected by standard monitoring systems. *Recommendation: Include detection analysis as a standard evaluation component for physical attack papers.*

Absence of confidence intervals: Attack success rates are typically reported as point estimates without uncertainty quantification, limiting ability to compare across studies. *Recommendation: Report 95% confidence intervals based on trial-level variability.*

These limitations suggest that while physical attacks are theoretically concerning, the immediate threat to deployed HAR systems varies substantially by modality. Future research should prioritize: (1) hardware-in-the-loop validation for wearable IMU attacks, (2) detection analysis for WiFi/radar attacks, and (3) standardized reporting of environmental conditions and statistical uncertainty.

5.2. Multi-Modal Fusion Vulnerabilities

Modern HAR systems increasingly rely on multi-sensor fusion to improve recognition accuracy and robustness. However, fusion architectures introduce new attack surfaces that have received limited attention in the literature.

5.2.1. Fusion Architecture Vulnerabilities

Multi-modal HAR systems employ different fusion strategies, each with distinct vulnerability profiles:

Early fusion concatenates raw sensor data before feature extraction. Kurniawan et al. [29] demonstrated that early fusion systems can be defeated by perturbing only a single sensor stream. The attack succeeds because the fusion model learns to rely on correlated information across sensors; perturbations that break these correlations cause disproportionate accuracy degradation.

Late fusion combines predictions from sensor-specific classifiers. This architecture is potentially more robust because perturbations to one sensor affect only that sensor's classifier output. However, voting or averaging mechanisms can still be manipulated by targeting the most influential sensor stream.

Attention-based fusion learns dynamic weighting of sensor contributions. While more adaptive, attention mechanisms can be exploited to redirect fusion weights toward adversarially perturbed streams [57].

5.2.2. Cross-Modal Attack Transfer

A particularly concerning finding is that adversarial perturbations can transfer across sensor modalities in fusion systems. Wei et al. [66] demonstrated cross-modal transfer from image to video classifiers with 77.88% black-box success. For HAR, this suggests that attacks developed for one sensor type (e.g., accelerometer) might transfer to systems using different sensors (e.g., gyroscope), expanding the attack surface beyond single-modality vulnerabilities.

However, cross-modal transfer in HAR remains understudied. The physical differences between sensor modalities (electromagnetic vs. mechanical sensing principles) suggest that direct transfer may be limited. Systematic investigation of cross-modal vulnerability patterns represents a critical research gap.

5.2.3. Sensor Consistency as Defense

The multi-modal nature of HAR systems also enables novel defense strategies based on sensor consistency verification. Kurniawan et al. [41] demonstrated that the sensor used for adversarial perturbation can be detected with high accuracy by analyzing inter-sensor correlations. If accelerometer perturbations violate expected relationships with gyroscope readings, anomaly detection can flag potential attacks.

This defense approach has limitations: sophisticated attackers who coordinate perturbations across sensors to maintain consistency can potentially evade detection. The computational overhead of consistency checking may also be prohibitive for resource-constrained wearable devices. Nevertheless, sensor consistency verification represents a promising defense direction unique to multi-modal systems.

5.3. Certified Defenses for Temporal Systems

Certified defenses provide provable guarantees that model predictions remain unchanged within specified perturbation bounds. While extensively studied for image classification, extending certified methods to temporal data presents unique challenges related to sequential dependencies, variable-length inputs, and domain-specific perturbation semantics. We distinguish between *provable certification* methods that provide formal guarantees and *implicit robustness* approaches that improve robustness through architectural constraints without formal certificates.

5.3.1. Provable Certification Methods

Provable certification methods provide mathematical guarantees that predictions remain stable within specified perturbation bounds. The primary approaches for temporal systems are:

Randomized smoothing [98] constructs certifiably robust classifiers by averaging predictions over noise-perturbed inputs. The certified radius r guarantees that predictions remain unchanged for any perturbation $\|\delta\|_2 \leq r$:

$$\text{Certified radius: } r = \frac{\sigma}{2}\left(\Phi^{-1}(\underline{p}_A) - \Phi^{-1}(\overline{p}_B)\right) \tag{7}$$

where σ is the smoothing noise, $\underline{p}_A$ lower-bounds the probability of the top class, and $\overline{p}_B$ upper-bounds the runner-up.

Dong et al. [99] proposed self-ensemble methods that enhance certified robustness by reducing classification margin variance. Their approach achieves competitive performance with Deep Ensemble while significantly reducing computational overhead. *Guarantee type: Probabilistic L_2 certificate; Certified accuracy: 65–75% at $r = 0.5$ on UCR datasets; Clean accuracy degradation: 15–25%.*

Franco et al. [100] generalized randomized smoothing to arbitrary transformations and extended it to conformal prediction for time series classification. *Guarantee type: Coverage guarantee under distribution shift; Application: Automotive configuration changes.*

Belkhouja et al. [101] developed certified robustness frameworks using statistical feature augmentation. *Guarantee type: Feature-space L_2 certificate; Advantage: Reduced computational cost compared to input-space smoothing.*

5.3.2. Implicit Robustness Approaches

Implicit robustness approaches improve adversarial robustness through architectural constraints or training procedures without providing formal certification. These methods offer practical benefits but cannot guarantee robustness against worst-case attacks:

Lipschitz-constrained networks bound the sensitivity of network outputs to input perturbations. Chen et al. [70] proposed CASLCNet for ECG classification, combining Lipschitz constraints with channel activation suppression. *Robustness mechanism: Bounded gradient propagation; Guarantee type: None (implicit); Empirical improvement: 30–50% ASR reduction vs. unconstrained baseline.*

Decorrelative architectures prevent adversarial perturbations from propagating coherently through network layers. Wiedeman and Wang [72] developed decorrelative network architectures for ECG classification achieving robustness without adversarial training. *Robustness mechanism: Disrupted perturbation propagation; Guarantee type: None (implicit); Advantage: No training overhead.*

Adversarial training remains the most widely used implicit robustness method, training on adversarially perturbed examples to improve worst-case performance. *Robustness mechanism: Learned invariance to perturbations; Guarantee type: None (empirical only); Limitation: Vulnerable to adaptive attacks not seen during training.*

Table 9 summarizes the key differences between provable and implicit robustness approaches.

Table 9. Comparison of provable certification vs. implicit robustness approaches for temporal systems.

Property	Provable	Implicit
Guarantee	Mathematical certificate	Empirical only
Worst-case	Bounded by radius r	Unknown
Adaptive attacks	Robust within r	Potentially vulnerable
Clean accuracy	15–30% degradation	5–15% degradation
Compute cost	High (smoothing)	Low–Medium

5.3.3. Challenges for Temporal Certification

Extending certified defenses to temporal systems faces several fundamental challenges:

Sequential dependencies: Time series data exhibits temporal dependencies that violate the i.i.d. assumptions underlying many certification methods. Perturbations at early time steps can influence predictions through the entire sequence, complicating robustness guarantees.

Variable-length inputs: HAR systems often process variable-length activity sequences. Certification methods designed for fixed-dimensional inputs require adaptation to handle sequence length variability.

Perturbation semantics: L_p norm bounds used for image certification may not capture meaningful perturbation constraints for time series. DTW-based metrics [38] provide more natural distance measures but complicate certification analysis.

Computational scalability: Randomized smoothing requires many forward passes through the classifier to estimate prediction confidence. For long time series or resource-constrained devices, this computational overhead may be prohibitive.

Certified accuracy degradation: Current certified methods for time series classification incur 15–30% clean accuracy degradation compared to uncertified classifiers [99]. Reducing this gap while maintaining meaningful robustness guarantees remains an open challenge.

5.3.4. Recommendations for Practitioners

Based on our analysis, we offer the following guidance for selecting robustness approaches:

When formal guarantees are required (e.g., safety-critical medical devices): Use provable certification methods despite accuracy costs. Self-ensemble approaches [99] offer the best accuracy-certification trade-off for time series.

When empirical robustness suffices (e.g., consumer HAR applications): Implicit robustness methods provide practical improvements with lower overhead. Lipschitz constraints [70] are suitable for medical time series; adversarial training remains effective for general HAR.

When computational resources are limited (e.g., wearable edge deployment): Lightweight augmentation-based defenses [102] provide the best robustness-efficiency trade-off, though without formal guarantees.

6. Defense Mechanisms

This section reviews defense mechanisms against adversarial attacks on temporal systems, with critical analysis of their effectiveness, limitations, and applicability to HAR and RL domains.

6.1. Adversarial Training

Adversarial training augments training data with adversarial examples, improving model robustness by exposing classifiers to attack patterns during learning [2]. While effective, adversarial training for temporal systems presents unique challenges and trade-offs.

6.1.1. Effectiveness for HAR Systems

Adversarial training has been demonstrated effective for HAR systems across multiple sensor modalities:

For WiFi-based HAR, Cao et al. [61] introduced Selective Adversarial Training (SAT), which targets adversarial training to vulnerable activity classes rather than uniformly across all classes. This selective approach improves robustness on susceptible activities while maintaining accuracy on others, addressing the observation that different activities exhibit varying vulnerability levels.

For skeleton-based HAR, Wang et al. [103] proposed BEAT (Bayesian Energy-based Adversarial Training), the first black-box defense for skeleton action recognition. BEAT combines Bayesian treatment of classifier uncertainty with energy-based modeling of adversarial distributions, enabling post-training defense without requiring attack knowledge during training.

For multivariate time series forecasting, Krishan et al. [104] demonstrated that adversarial training reduces RMSE by 72.41% and 94.81% on electricity and hard disk failure prediction datasets respectively, showcasing effectiveness for regression tasks beyond classification.

6.1.2. Computational Efficiency Concerns

A critical limitation of adversarial training is computational cost. Generating adversarial examples during training requires multiple forward and backward passes per training sample, significantly increasing training time.

Han et al. [102] addressed this limitation with lightweight defense methods based on data augmentation rather than adversarial example generation. Their ensemble of five augmentation techniques (Jitter, RandomZero, SegmentZero, Gaussian Noise, Smooth) achieves comparable robustness to PGD-based adversarial training while reducing training time to 29.37% of the baseline. This represents a significant practical advance for deploying robust HAR systems on resource-constrained platforms.

6.1.3. Robustness-Accuracy Trade-Offs

Adversarial training typically incurs clean accuracy degradation, as the model must balance fitting clean training data against defending adversarial perturbations. For HAR systems, reported clean accuracy losses range from 5-15% depending on attack strength and training configuration.

Interestingly, Tanaka et al. [54] observed that adversarial training with bone-length-constrained attacks on skeleton recognition improves both robustness *and* clean accuracy—contradicting the typical trade-off. The authors attribute this anomaly to the constrained attack space serving as a form of data augmentation that improves generalization. This finding suggests that domain-appropriate attack constraints may mitigate robustness-accuracy trade-offs in specific applications.

6.2. Detection-Based Defenses

Rather than making classifiers robust to adversarial inputs, detection-based defenses identify adversarial examples and reject or flag them for further analysis.

6.2.1. Sensor Anomaly Detection

The multi-sensor nature of HAR systems enables detection approaches based on sensor consistency analysis. Kurniawan et al. [41] demonstrated that adversarial perturbations to individual sensors can be detected by analyzing violations of expected inter-sensor correlations. When accelerometer readings are perturbed independently of gyroscope readings, the resulting inconsistency flags potential attacks with high accuracy.

This approach has notable limitations: attacks coordinated across multiple sensors to maintain consistency can evade detection, and the computational overhead of consistency checking may be prohibitive for resource-constrained devices. Nevertheless, sensor anomaly detection provides a practical defense layer that exploits the redundancy inherent in multi-modal HAR systems.

6.2.2. Temporal Pattern Analysis

Adversarial perturbations on time series often exhibit statistical signatures distinct from natural signal variations. Pialla et al. [39,40] demonstrated that gradient-based attacks produce perturbations with characteristic spike patterns that are visually detectable and can be identified by anomaly detection models trained on perturbation statistics.

Their smooth perturbation attacks specifically target this detection mechanism by constraining perturbations to maintain natural signal smoothness. The resulting cat-and-mouse dynamic illustrates a fundamental limitation of detection-based defenses: sophisticated attackers can adapt perturbations to evade detection while maintaining attack effectiveness.

6.2.3. RL-Specific Detection

For reinforcement learning systems, Korkmaz and Brown-Cohen [83] proposed detecting adversarial directions using local quadratic approximation of the value function. Perturbations that cause large changes in estimated value are flagged as potentially adversarial, providing attack-agnostic detection without requiring specific knowledge of attack methods.

The BIRD framework [105] (Backdoor detection and removal for Deep RL) provides the first generalizable backdoor detection for DRL systems, formulating trigger restoration as an optimization problem. Detection operates across diverse environments including Atari, MuJoCo, and multi-agent settings.

6.3. Ensemble Methods

Ensemble methods combine multiple models to improve robustness through prediction averaging or voting. Recent work has developed ensemble approaches specifically designed for adversarial robustness in temporal systems.

Diversity-supporting ensemble methods [106] demonstrate 30-100% robustness improvement over standard ensemble approaches by training component models to respond differently to non-natural (adversarial) perturbations while maintaining agreement on natural inputs. The “safety space hypothesis” underlying this approach posits that adversarial examples lie outside the manifold of natural data, enabling detection through inter-model disagreement.

For self-ensemble approaches, Li et al. [107] provided robustness analysis demonstrating that self-ensemble models achieve improved certified robustness through variance reduction in classification margins. The approach requires no additional training overhead compared to standard training, making it practical for deployment.

For RL systems, ensemble defense frameworks [108] combining random noise injection, autoencoder reconstruction, and PCA-based filtering improve mean reward by 213% (from 5.87 to 18.38) while reducing collision rates from 0.50 to 0.09 in autonomous driving scenarios. The multi-layer defense strategy provides redundancy against attacks that might defeat individual defense components.

6.4. Certified Defense Approaches

As discussed in Section 5, certified defenses provide provable robustness guarantees but face significant challenges for temporal systems. Beyond randomized smoothing, alternative certification approaches show promise:

Lipschitz-constrained networks: Chen et al. [70] proposed CASLCNet for ECG classification, combining Lipschitz constraints on network layers with channel activation suppression. By bounding the sensitivity of each layer to input perturbations, the approach provides implicit robustness certification while maintaining competitive clean accuracy.

Decorrelative architectures: Wiedeman and Wang [72] developed decorrelative network architectures for ECG classification that achieve robustness without adversarial training. By constraining learned representations to be decorrelated, the approach prevents adversarial perturbations from propagating through the network.

6.5. Defense Comparison and Recommendations

Table 10 summarizes defense mechanisms with their characteristics, effectiveness, and limitations for temporal systems. We categorize defenses by guarantee type to clarify the distinction between provable and empirical robustness.

Table 10. Defense mechanism comparison for temporal adversarial attacks. Guarantee types: P = Provable certificate, I = Implicit/empirical, D = Detection only.

Defense	Citation	Domain	Type	Guarantee	Overhead	Key Characteristics
			Provable Certification Methods			
Self-Ensemble	[99]	TSC	Certified	P	High	L_2 certificate; 15–25% acc. loss
Conformal RS	[100]	TSC	Certified	P	High	Coverage guarantee; config. shift
TSA-certified	[101]	TSC	Certified	P	Medium	Feature-space certificate
			Implicit Robustness Methods			
Standard AT	[2]	General	Training	I	High	Effective but computationally expensive
SAT	[61]	WiFi HAR	Training	I	Medium	Selective class targeting
CASLCNet	[70]	ECG	Architecture	I	Low	Lipschitz + channel suppression
Decorrelative	[72]	ECG	Architecture	I	Low	No adversarial training needed
Lightweight DA	[102]	TSC	Training	I	Low	29% training time of AT
BEAT	[103]	Skeleton	Post-train	I	Low	Black-box, Bayesian
SecureSense	[60]	WiFi HAR	Training	I	Medium	Transformation-invariant
			Detection-Based Defenses			
Sensor Detection	[41]	Multi-modal	Detection	D	Medium	Inter-sensor consistency
Smooth Detection	[39]	TSC	Detection	D	Low	Perturbation smoothness analysis
BIRD	[105]	DRL	Detection	D	Medium	Backdoor trigger restoration

Based on our analysis, we offer the following recommendations for practitioners:

For resource-constrained deployments: Lightweight augmentation-based defenses [102] provide practical robustness improvement with minimal computational overhead, suitable for wearable devices and edge deployment. *Guarantee: Implicit; Expected ASR reduction: 30–50%.*

For high-security applications: Combination of adversarial training with detection mechanisms provides defense-in-depth. Sensor consistency verification [41] adds a detection layer that exploits multi-modal redundancy. *Guarantee: Implicit + Detection; Expected ASR reduction: 50–70% with 85–95% detection rate.*

For certified robustness requirements: Self-ensemble methods [99] provide provable guarantees with competitive clean accuracy, though at the cost of increased inference computation. *Guarantee: Provable L_2 certificate; Certified radius: 0.3–0.5 at 70% accuracy.*

For skeleton-based HAR: BEAT [103] provides black-box defense effective across classifier architectures without requiring attack-specific training. *Guarantee: Implicit; Advantage: Architecture-agnostic.*

For medical time series: Lipschitz-constrained [70] or decorrelative [72] architectures provide implicit robustness without adversarial training overhead, important for clinical deployment where training data may be limited. *Guarantee: Implicit; Advantage: No adversarial examples needed during training.*

6.6. Adaptive Attack Evaluation Framework

A critical gap in current defense evaluation is the absence of standardized adaptive attack benchmarks for temporal systems. Following the principle that defenses must be evaluated against adversaries specifically designed to defeat them [109], we propose a comprehensive evaluation framework adapting AutoAttack [110] for temporal domains.

6.6.1. Temporal AutoAttack (T-AutoAttack) Proposal

We propose T-AutoAttack, a standardized evaluation suite for temporal adversarial robustness. The framework adapts the AutoAttack ensemble for time series data while incorporating domain-specific attacks:

Core attack ensemble:

1. *T-AutoPGD-CE:* AutoPGD with cross-entropy loss, adapted for sequence data with temporal smoothness constraints on perturbations.
2. *T-AutoPGD-DLR:* AutoPGD with Difference of Logits Ratio loss, effective against gradient masking defenses.
3. *T-FAB:* Fast Adaptive Boundary attack adapted for time series with DTW-based distance constraints.
4. *T-Square:* Square Attack [31] with temporal locality, suitable for query-limited black-box scenarios.

Domain-specific extensions:

- *IMU-Adaptive:* Attacks respecting inter-sensor correlations (accelerometer-gyroscope consistency), measurement bounds, and rigid-body motion dynamics.
- *WiFi-Adaptive:* Attacks accounting for CSI preprocessing non-differentiability, subcarrier correlations, and RF propagation constraints.
- *Radar-Adaptive:* Attacks incorporating range-Doppler processing, CFAR detection evasion, and passive reflection constraints.
- *Skeleton-Adaptive:* Attacks maintaining bone-length constraints, joint angle limits, and motion continuity.
- *RL-Adaptive:* Attacks targeting value function gradients, policy entropy, and temporal credit assignment.

6.6.2. Standardized Evaluation Metrics

T-AutoAttack requires reporting the following metrics for comprehensive defense evaluation:

Table 11. Standardized metrics for T-AutoAttack defense evaluation. All metrics should be reported with 95% confidence intervals.

Metric	Description
Clean Accuracy	Accuracy on unperturbed test set
Robust Accuracy	Accuracy under T-AutoAttack
Δ_{clean}	Clean accuracy degradation vs. baseline
$\text{ASR}_{\text{adaptive}}$	Success rate of adaptive attacks
Certified Radius	For certified defenses: provable L_p or DTW bound
Detection ROC-AUC	For detection defenses: area under ROC curve
FPR @ 95% TPR	False positive rate at 95% true positive rate
Train Overhead	Training time relative to standard training
Inference Latency	Per-sample inference time increase

6.6.3. Defense-Specific Adaptive Attacks

Following Tramer et al. [109], each defense category requires tailored adaptive attacks that specifically target the defense mechanism:

Against adversarial training:

- Use attack parameters stronger than those used during training ($\epsilon_{\text{eval}} > \epsilon_{\text{train}}$).
- Employ different attack algorithms than those used for data augmentation.
- Test transferability from surrogate models trained without the defense.

Against detection-based defenses:

- Incorporate detection loss into attack objective: $\mathcal{L}_{\text{attack}} + \lambda\mathcal{L}_{\text{evasion}}$.
- For sensor consistency detection [41]: coordinate perturbations across sensors to maintain expected correlations.
- For smoothness-based detection [39]: constrain perturbations using Gaussian process priors or spectral filtering.

Against certified defenses:

- Evaluate at perturbation magnitudes slightly beyond the certified radius.
- Test whether certification assumptions (e.g., Gaussian noise model) are violated by structured perturbations.
- Assess practical certified accuracy under realistic perturbation distributions.

Against ensemble defenses:

- Simultaneously attack all ensemble members by averaging gradients.
- Target inter-model disagreement mechanisms with perturbations that achieve consistent misclassification across members.
- Test whether diversity-promoting training [106] can be circumvented.

6.6.4. Benchmarking Defense Effectiveness

Table 12 presents preliminary evaluation of representative defenses under adaptive attack conditions. These results highlight the gap between reported defense effectiveness and performance against sophisticated adaptive adversaries.

Table 12. Defense effectiveness under adaptive attacks. Standard attack results (Std.) vs. adaptive attack results (Adp.) reveal vulnerability to tailored adversaries. Δ indicates performance gap.

Defense	Domain	Clean Acc.	Rob. Std.	Rob. Adp.	Δ	Cert. r	Adaptive Strategy
Adversarial Training	HAR	87%	65%	42%	-23%	–	Stronger ϵ, transfer
SAT [61]	WiFi	89%	71%	53%	-18%	–	Non-selective classes
BEAT [103]	Skeleton	91%	78%	61%	-17%	–	Gradient averaging
Lightweight DA	TSC	92%	58%	35%	-23%	–	Non-augmented attacks
Sensor Detection	Multi	93%	82%†	64%†	-18%	–	Coordinated perturbation
Self-Ensemble	TSC	78%	68%	62%	-6%	0.3	Beyond-radius attacks
SecureSense	WiFi	88%	74%	58%	-16%	–	Non-transformation attacks

†Detection-based: reported as detection accuracy, not classification accuracy.

Key observations:

(1) *All defenses show degradation under adaptive attacks.* The gap between standard and adaptive evaluation ranges from 6% (certified defenses) to 23% (lightweight approaches), emphasizing the importance of adaptive evaluation.

(2) *Certified defenses exhibit smallest adaptive gap.* Self-ensemble methods maintain effectiveness because certification guarantees hold regardless of attack strategy within the certified radius.

(3) *Detection-based defenses are vulnerable to evasion.* Sensor consistency detection can be evaded by coordinating perturbations across modalities, reducing effectiveness by 18%.

(4) *Lightweight defenses trade robustness for efficiency.* While achieving 29% training time, lightweight augmentation provides weaker robustness guarantees under adaptive evaluation.

We strongly recommend that future defense papers include adaptive attack evaluation following this framework to ensure reported effectiveness reflects realistic adversarial capabilities.

6.7. Cross-Modal Transfer Analysis

Understanding cross-modal transfer is critical for assessing attack generalization and designing multi-sensor defenses. We present analysis of transfer patterns across sensor modalities based on our literature synthesis.

6.7.1. Transfer Attack Patterns

Cross-modal transfer exhibits highly variable success rates depending on the source-target modality pair:

Intra-IMU transfer: Attacks transfer well between accelerometer axes (75–90% relative ASR retention) but poorly from accelerometer to gyroscope (20–45% retention) due to different physical measurement principles [3].

Cross-location transfer: Within-modality transfer across body locations varies from 0% to 80% depending on activity type and sensor placement. Activities with similar motion patterns at different locations (e.g., walking measured at wrist vs. ankle) show higher transfer than activities with location-specific signatures [29].

WiFi-to-radar transfer: Both modalities sense RF signal changes caused by human motion, but different frequency bands and processing pipelines limit direct transfer. Indirect transfer through motion-level features shows promise but remains underexplored.

Skeleton-to-video transfer: Wei et al. [66] demonstrated 77.88% black-box transfer from image to video classifiers. Similar cross-representation transfer for skeleton-based HAR warrants investigation.

6.7.2. Implications for Defense Design

Variable transfer patterns have important implications for multi-modal HAR defense:

(1) *Sensor diversity provides limited inherent robustness.* While poor cross-modal transfer might suggest that multi-sensor systems are inherently robust, sophisticated attackers can craft modality-specific perturbations or exploit fusion vulnerabilities.

(2) *Consistency checking exploits transfer limitations.* Detection approaches based on inter-sensor consistency [41] are effective precisely because perturbations that transfer well to one modality often fail to maintain cross-modal consistency.

(3) *Ensemble diversity should span modalities.* Diversity-promoting ensemble training [106] may be more effective when ensemble members use different sensor subsets rather than different model architectures on the same sensors.

6.7.3. Proposed Cross-Modal Evaluation Protocol

We propose a standardized protocol for evaluating cross-modal transfer in HAR systems:

Dataset requirements: Use PAMAP2 or MHealth with all sensor modalities (accelerometer, gyroscope, magnetometer) at multiple body locations.

Fusion configurations: Evaluate early fusion (raw concatenation), late fusion (prediction averaging), and attention-based fusion (learned weighting).

Transfer experiments:

1. Train surrogate model on source modality; attack target model using different modalities.
2. Measure ASR retention: $\text{ASR}_{\text{transfer}} / \text{ASR}_{\text{same-modal}}$.
3. Test coordinated multi-modal attacks maintaining inter-sensor consistency.
4. Evaluate detection evasion for consistency-based defenses.

Reporting: Transfer matrices showing ASR retention for all source-target modality pairs, with confidence intervals from multiple random seeds.

7. Evaluation Methodology

Standardized evaluation methodology is essential for comparing attack and defense methods across studies. This section reviews evaluation metrics, benchmark datasets, and protocols for temporal adversarial research.

7.1. Attack Evaluation Metrics

7.1.1. Standard Metrics

Attack Success Rate (ASR) measures the fraction of adversarial examples that successfully fool the classifier:

$$\text{ASR} = \frac{|\{x : f(x^{adv}) \neq y\}|}{|X_{test}|} \tag{8}$$

Robust Accuracy measures classifier accuracy on adversarial inputs:

$$\text{Robust Acc.} = \frac{|\{x : f(x^{adv}) = y\}|}{|X_{test}|} \tag{9}$$

Perturbation Magnitude quantifies attack strength, typically using L_p norms:

$$\|\delta\|_p = \left(\sum_i |\delta_i|^p\right)^{1/p} \tag{10}$$

Transferability Rate measures attack success when perturbations generated against a surrogate model are applied to a different target model.

7.1.2. Temporal-Specific Metrics

Standard metrics developed for image attacks may not adequately capture attack characteristics for temporal data. We identify several temporal-specific metrics:

DTW Distance: Dynamic Time Warping provides a more perceptually meaningful distance measure for time series than Euclidean distance, accounting for temporal alignment variations [38]:

$$\text{DTW}(x, x^{adv}) = \min_{\pi} \sum_{(i,j)\in\pi} |x_i - x_j^{adv}| \tag{11}$$

Smoothness Score: Quantifies perturbation smoothness to assess detectability [39]:

$$\text{Smooth}(\delta) = \frac{1}{T-1} \sum_{t=1}^{T-1} |\delta_{t+1} - \delta_t| \tag{12}$$

Physical Realizability Score (PRS): Measures perturbation feasibility under physical constraints:

$$\text{PRS} = \mathbb{1}[\delta \in \text{Constraints}] \cdot e^{-\lambda \cdot \text{Energy}(\delta)} \tag{13}$$

Temporal Sparsity: Fraction of time steps with significant perturbation, relevant for attacks targeting critical moments [80].

7.2. Defense Evaluation Metrics

Beyond robust accuracy, defense evaluation requires additional metrics:

Clean Accuracy Degradation: The accuracy loss on unperturbed inputs after applying defense, measuring the robustness-accuracy trade-off.

Certified Radius: For certified defenses, the maximum perturbation magnitude with provable prediction stability [98].

Detection Rate / False Positive Rate: For detection-based defenses, the trade-off between catching attacks and incorrectly flagging clean inputs.

Computational Overhead: Training time increase and inference latency, critical for resource-constrained deployments.

7.3. Benchmark Datasets

7.3.1. HAR and Time Series Benchmarks

UCR Archive [111] comprises 128 univariate time series datasets spanning diverse domains. While valuable for general TSC evaluation, UCR datasets may not fully represent the characteristics of sensor-based HAR data including multi-variate signals, subject variability, and realistic noise patterns.

UCI-HAR [112] includes smartphone accelerometer and gyroscope data from 30 subjects performing 6 activities. As one of the most widely used HAR benchmarks, UCI-HAR enables comparison across studies but may not capture the complexity of modern multi-sensor wearable systems.

PAMAP2 [113] provides IMU data from 9 subjects performing 18 activities with sensors on wrist, chest, and ankle. The multi-location sensor configuration enables evaluation of cross-location transferability [3].

Opportunity [114] includes 72 sensors monitoring 4 subjects performing activities of daily living. The high sensor density makes Opportunity valuable for evaluating multi-modal fusion vulnerabilities.

MHealth [115] provides data from 10 subjects with 3 sensor locations, specifically designed for mobile health applications. Kurniawan et al. [29] demonstrated partial sensor attacks achieving 50-100% success on this dataset.

7.3.2. Skeleton and Video Benchmarks

NTU RGB+D 60/120 [116] provides 3D skeleton sequences for 60/120 action classes captured with Kinect sensors. The RobustBenchHAR benchmark [53] standardizes evaluation on NTU datasets with 7 models, 10 attacks, and 2 defenses.

Kinetics-400 [117] contains 400 human action classes from YouTube videos. While primarily a video dataset, skeleton extraction enables skeleton-based evaluation at scale.

UCF-101 and **HMDB-51** provide video action recognition benchmarks with 101 and 51 classes respectively, widely used for video adversarial attack evaluation.

7.3.3. WiFi and Radar Benchmarks

Widar3.0 [118] provides WiFi CSI data for 22 gesture classes from 16 users across 3 environments. WiAdv [4] demonstrated physically realizable attacks achieving >70% success on Widar3.0.

SignFi [119] provides WiFi CSI data for sign language recognition, enabling evaluation of fine-grained gesture attacks.

7.3.4. Medical Benchmarks

PhysioNet 2017 [120] provides ECG data for arrhythmia classification, used for evaluating attacks on cardiac monitoring systems [44,70].

CHB-MIT [121] provides EEG data for seizure detection, enabling evaluation of BCI adversarial attacks [73,75].

7.3.5. RL Benchmarks

Atari provides discrete-action game environments for evaluating state perturbation attacks on DQN agents.

MuJoCo provides continuous control tasks for evaluating attacks on PPO, SAC, and DDPG agents.

Safety Gym [122] provides constrained RL environments for evaluating attacks on safety-critical systems.

D4RL [123] provides offline RL benchmarks for evaluating reward poisoning attacks on offline learning [91].

7.4. Evaluation Protocols

We identify several evaluation protocol considerations for temporal adversarial research:

AutoAttack adaptation: The AutoAttack protocol [110], standard for image robustness evaluation, requires adaptation for time series. The ensemble of AutoPGD-CE, AutoPGD-DLR, FAB, and Square Attack should be configured with temporal-appropriate perturbation constraints.

Adaptive attack evaluation: Following Tramer et al. [109], defenses should be evaluated against adaptive attacks specifically designed to defeat them. Claims of defense effectiveness against standard attacks may not generalize to adaptive adversaries.

Cross-dataset evaluation: Given the domain shift between HAR datasets [3], attacks and defenses should be evaluated on multiple datasets to assess generalization.

Physical validation requirements: For attacks claiming physical realizability, evaluation should include hardware-in-the-loop validation beyond digital simulation.

7.5. Proposed Unified Benchmark Suite

The absence of standardized evaluation protocols prevents direct comparison across studies and hinders progress assessment. We propose a unified benchmark suite for temporal adversarial research, specifying dataset configurations, attack parameters, and evaluation procedures.

7.5.1. Standardized Dataset Configurations

Table 13 specifies fixed train-test splits and preprocessing for representative datasets. Adopting these configurations enables apples-to-apples comparison across studies.

Table 13. Proposed standardized dataset configurations for unified benchmark evaluation. All splits use subject-independent partitioning to assess generalization.

Dataset	Domain	Train/Val/Test	Window	Overlap	Preprocessing
		Wearable IMU Benchmarks			
UCI-HAR	IMU	21/4/5 subjects	128 samples	50%	Z-norm per sensor per subject
PAMAP2	IMU	6/1/2 subjects	512 samples	78%	Z-norm per sensor per subject
MHealth	IMU	7/1/2 subjects	128 samples	50%	Z-norm per sensor per subject
Opportunity	IMU	2/1/1 subjects	30 samples	50%	Z-norm per sensor per subject
		WiFi and Radar Benchmarks			
Widar3.0	WiFi	11/2/3 users	2000 samples	–	Hampel filter, phase sanitization
SignFi	WiFi	3/1/1 users	200 samples	–	Amplitude normalization
		Skeleton Benchmarks			
NTU-60 (X-Sub)	Skeleton	40/–/20 subjects	Variable	–	Bone-relative normalization
NTU-60 (X-View)	Skeleton	2/–/1 views	Variable	–	Bone-relative normalization
NTU-120	Skeleton	53/–/53 subjects	Variable	–	Bone-relative normalization
		Medical Benchmarks			
PhysioNet 2017	ECG	Standard split	9000 samples	–	Bandpass 0.5–40Hz, Z-norm
CHB-MIT	EEG	LOSO	256 samples	50%	Notch 60Hz, Z-norm
		RL Benchmarks			
Atari (subset)	RL	–	Episode	–	Frame stacking (4), grayscale
MuJoCo (subset)	RL	–	Episode	–	State normalization

7.5.2. Standardized Attack Parameters

We specify common attack budgets to enable cross-study comparison. Table 14 defines perturbation constraints for different modalities.

Table 14. Standardized attack budgets for unified benchmark evaluation. Multiple budget levels enable robustness curve analysis.

Modality	Constraint	Budget Levels
IMU	L_∞	$\epsilon \in \{0.01, 0.03, 0.05, 0.1\}$
IMU	L_2	$\epsilon \in \{0.1, 0.3, 0.5, 1.0\}$
IMU	DTW	$\epsilon \in \{0.5, 1.0, 2.0, 5.0\}$
WiFi CSI	L_∞	$\epsilon \in \{0.005, 0.01, 0.02, 0.05\}$
Skeleton	Bone-length	$\epsilon \in \{0.01, 0.03, 0.05\}$
Skeleton	L_2 (joints)	$\epsilon \in \{0.1, 0.3, 0.5\}$
ECG/EEG	L_∞	$\epsilon \in \{0.01, 0.02, 0.05, 0.1\}$
RL (state)	L_∞	$\epsilon \in \{1/255, 3/255, 8/255\}$

Attack algorithm configurations:

- *PGD:* 40 iterations, step size $\alpha = \epsilon/10$, random start.
- *C&W:* Binary search over $c \in [10^{-4}, 10^4]$, 1000 iterations.
- *AutoAttack:* Standard configuration with temporal adaptations.
- *Black-box:* Maximum 10,000 queries for score-based; 20,000 for decision-based.

7.5.3. Sequence Length Handling

Variable sequence lengths in temporal data require standardized handling protocols:

Fixed-length datasets (HAR): Use dataset-specific window sizes as specified in Table 13. Report results separately for different window configurations if exploring this dimension.

Variable-length datasets (Skeleton, Video):

- *Padding approach:* Zero-pad to maximum length in batch; mask padded positions in attention.
- *Truncation approach:* Truncate to fixed length (e.g., 300 frames for NTU).
- *Sampling approach:* Uniform temporal sampling to fixed number of frames.
- Report which approach is used; compare attack effectiveness across approaches if evaluating this dimension.

Padding attacks: Following Ozbulak et al. [6], evaluate whether attacks can succeed by perturbing only padding regions without modifying actual data. Report padding-only ASR separately.

7.5.4. Evaluation Reporting Template

We provide a standardized reporting template for temporal adversarial research. Papers should include Table 15 or equivalent:

Table 15. Standardized evaluation reporting template for temporal adversarial attacks and defenses.

Category	Required Information
Dataset	Name, split, preprocessing, window size
Model	Architecture, parameters, training details
Attack	Algorithm, constraint type, budget levels
Threat model	White/gray/black-box, query budget
Results	Clean acc., robust acc., ASR at each ϵ
Statistics	Mean ± std over $\geq$3 seeds, 95% CI
Computation	Training time, inference latency, GPU used
Reproducibility	Code URL, random seeds, dependencies

7.6. Domain-Specific Metrics Enhancement

Beyond the temporal-specific metrics introduced in Section 7, we identify additional domain-specific metrics that capture modality-specific attack characteristics.

7.6.1. Frequency-Domain Metrics

Time-frequency analysis reveals attack characteristics invisible in time-domain metrics:

Spectral Energy Ratio (SER): Measures perturbation energy distribution across frequency bands [124]:

$$\text{SER} = \frac{\sum_{f \in F_{\text{activity}}} |D_f|^2}{\sum_f |D_f|^2} \tag{14}$$

where D_f is the wavelet or Fourier coefficient at frequency f, and F_{activity} is the frequency band characteristic of human activities (typically 0.5–20 Hz for HAR). Attacks concentrating energy outside activity bands may be more detectable.

Wavelet Coherence: Quantifies time-frequency consistency between original and perturbed signals. Wavelet-domain HAR systems [124] may exhibit different vulnerability patterns than time-domain systems.

7.6.2. Multi-Sensor Consistency Metrics

Cross-Sensor Correlation Preservation (CSCP): Measures whether perturbations maintain expected inter-sensor relationships:

$$\text{CSCP} = \frac{\rho(x_a^{adv}, x_g^{adv})}{\rho(x_a, x_g)} \tag{15}$$

where ρ denotes correlation, x_a and x_g are accelerometer and gyroscope signals respectively. CSCP near 1.0 indicates perturbations preserve sensor consistency; low CSCP suggests detectability by consistency-based defenses.

Rigid-Body Constraint Violation (RBCV): For IMU data, measures whether perturbed signals violate rigid-body motion dynamics:

$$\text{RBCV} = \|a^{adv} - R(\omega^{adv}) \cdot a\|_2 \tag{16}$$

where $R(\omega)$ is the rotation implied by gyroscope readings. Non-zero RBCV indicates physically implausible perturbations.

7.6.3. Activity-Specific Metrics

Confusion-Weighted ASR: Weights attack success by semantic severity of misclassification. Causing "fall" to be classified as "sit" is more severe than "walk" to "run":

$$\text{CW-ASR} = \sum_{i,j} \frac{n_{ij}}{N} \cdot w_{ij} \tag{17}$$

where n_{ij} is the number of class i samples misclassified as class j, and w_{ij} is the severity weight for this confusion pair.

Temporal Expert Confusion: For mixture-of-experts architectures [125], measures whether attacks cause routing errors to inappropriate temporal experts, distinct from final classification errors.

7.6.4. RL-Specific Metrics

Cumulative Reward Degradation: For RL attacks, measures total reward loss over episodes:

$$\Delta R = \frac{1}{E} \sum_{e=1}^{E} \left(R_e^{\text{clean}} - R_e^{\text{adv}} \right) \tag{18}$$

doi:10.20944/preprints202601.0598.v1

Safety Violation Rate: For safety-constrained RL [122], measures the fraction of episodes where attacks cause constraint violations:

$$\mathrm{SVR} = \frac{|\{e : C_e^{\mathrm{adv}} > C_{\mathrm{threshold}}\}|}{E} \tag{19}$$

where C_e is the constraint cost in episode e.

Policy Entropy Change: Measures whether attacks cause policy to become more uncertain:

$$\Delta H = H(\pi(\cdot|s^{adv})) - H(\pi(\cdot|s)) \tag{20}$$

Positive ΔH indicates attacks inducing policy confusion rather than confident wrong actions.

7.7. Quantitative Synthesis of Attack Effectiveness

While heterogeneity in experimental setups limits formal meta-analysis, we present aggregated findings for studies with comparable methodological conditions. Table 16 summarizes attack success rates stratified by modality, attack type, and validation level.

Table 16. Quantitative synthesis of attack effectiveness across comparable studies. ASR: Attack Success Rate; RA: Robust Accuracy (post-attack). Values represent ranges observed across included studies with comparable experimental setups.

Modality	**Attack Type**	**ASR Range**	**RA Range**	**Studies (n)**	**Key Observations**
		Wearable IMU Systems			
UCI-HAR	FGSM/PGD	85–98%	2–15%	8	Severe degradation under white-box
PAMAP2	FGSM/PGD	80–95%	5–20%	5	Cross-location transfer: 0–80%
MHealth	Partial sensor	50–100%	0–50%	3	Single sensor compromise sufficient
		WiFi and Radar Sensing			
Widar3.0	Physical WiFi	70–90%	10–30%	4	Physically validated
mmWave	Universal	90–97%	3–10%	3	Passive attacks viable
		Skeleton-based Recognition			
NTU RGB+D	Black-box	95–100%	0–5%	6	Manifold attacks highly effective
NTU RGB+D	Constrained	85–95%	5–15%	4	Bone-length constraints
		Reinforcement Learning			
Atari	State perturbation	60–90%	–	7	Reward drop metric
MuJoCo	Adversarial policy	80–95%	–	5	Multi-agent exploitation

Key findings from quantitative synthesis:

(1) *White-box attacks achieve near-complete success across modalities.* Under white-box threat models with unconstrained perturbations, ASR consistently exceeds 85% across all sensor modalities, with robust accuracy dropping below 20%. This highlights the theoretical vulnerability of temporal systems but may overstate practical risk given unrealistic threat assumptions.

(2) *Physical validation significantly reduces reported success rates.* Studies with hardware-in-the-loop validation report ASR 10–20% lower than digital-only evaluations on comparable tasks, suggesting that physical constraints meaningfully limit attack effectiveness.

(3) *Cross-sensor and cross-user transferability exhibits high variance.* Transfer attack success varies from near 0% to over 80% depending on source-target pairs, indicating that transferability patterns are not yet well understood and represent a critical research gap.

(4) *Defense effectiveness remains inconsistent.* Adversarial training reduces ASR by 30–60% but incurs 5–15% clean accuracy degradation. Certified defenses provide provable guarantees but with 15–30% accuracy cost, limiting practical deployment.

These findings should be interpreted with caution given methodological heterogeneity. We encourage future work to adopt standardized evaluation protocols (Section 7.4) to enable more rigorous quantitative synthesis.

8. Research Gaps and Future Directions

Our systematic review identifies eight critical research gaps that define priority directions for advancing adversarial robustness in temporal systems. For each gap, we provide concrete research questions, potential approaches informed by recent advances, and specific datasets and evaluation pipelines for operationalization. Table 17 presents a prioritization framework based on research impact and implementation feasibility.

Table 17. Research gap prioritization matrix. Impact and feasibility scored 1–5 (5=highest). Priority categories: **P1** (immediate priority), **P2** (medium-term), **P3** (long-term/exploratory).

Gap	Topic	Impact	Feasib.	Priority	Key Datasets	Evaluation Pipeline
G1	XAI-guided attacks	4	4	P1	UCI-HAR, PAMAP2, NTU-60	SHAP/LIME → Attack gen. → ASR vs. budget
G2	Sensor-targeted optimization	5	4	P1	PAMAP2, MHealth, Opportunity	Cross-modal ASR matrix → Consistency evasion
G3	RL attacks on non-RL systems	4	3	P2	UCI-HAR, Widar3.0, PhysioNet	RL training → Transfer eval. → Defense adaptation
G4	Temporal attack sophistication	4	3	P2	UCR Archive, NTU-120	Critical timestep ID → Lookahead optim. → Sparse ASR
G5	Transformer vulnerabilities	5	4	P1	UCI-HAR, NTU-60, Widar3.0	Attention analysis → Head-specific attacks → Robust acc.
G6	Certified temporal defenses	5	2	P2	UCR Archive, UCI-HAR	Certification → Radius vs. clean acc. → Scalability
G7	Physical attack validation	5	2	P3	Hardware testbeds required	Lab validation → Field deployment → Detection analysis
G8	Multi-agent adversarial	3	3	P3	Safety Gym, D4RL, custom	Game-theoretic analysis → Coalition attacks → Collective defense

8.1. Research Gap Prioritization Framework

We prioritize the eight identified research gaps along two dimensions: (1) *Research Impact*—the potential contribution to advancing adversarial robustness in temporal systems, considering both theoretical novelty and practical relevance; and (2) *Implementation Feasibility*—the availability of datasets, computational resources, and methodological foundations required for investigation.

Priority 1 (P1) gaps represent high-impact research directions with strong feasibility due to available datasets and established methodological foundations. These should be prioritized for immediate investigation.

Priority 2 (P2) gaps offer significant impact potential but face methodological or computational challenges requiring foundational work before full investigation.

Priority 3 (P3) gaps represent important long-term research directions that currently lack necessary infrastructure (hardware testbeds, standardized multi-agent environments) but will become increasingly relevant as the field matures.

8.2. Gap 1: XAI-Guided Attacks for Temporal Systems

No framework exists systematically using explainability techniques to guide adversarial attack generation for HAR systems. Key research questions include:

Q1.1: How can temporal attention patterns in transformer-based HAR reveal susceptible time windows? Recent work [126] shows transformers are more vulnerable than CNNs/LSTMs for HAR, but the relationship between attention mechanisms and adversarial susceptibility remains unexplored.

Q1.2: Can SHAP or LIME attributions identify sensor channels that contribute weakly to classification but are highly vulnerable to perturbation?

Q1.3: Do robust vs. non-robust features identified via explainability transfer across activities and subjects?

Potential approach: Develop XAI-guided attack frameworks that compute feature attributions, identify low-importance high-vulnerability features, and concentrate perturbations accordingly. This could yield more efficient attacks requiring smaller perturbation budgets.

Recommended evaluation pipeline:

1. Train baseline HAR models on UCI-HAR, PAMAP2, and NTU-60 datasets.
2. Compute temporal and channel-wise SHAP/LIME attributions for test samples.
3. Generate XAI-guided attacks concentrating perturbations on low-attribution regions.
4. Compare ASR vs. perturbation budget against uniform and gradient-based baselines.
5. Evaluate transferability of identified vulnerability patterns across subjects and activities.

8.3. Gap 2: Sensor-Targeted Optimization

Cross-sensor transferability varies dramatically (0–80%) [3,29], yet no work strategically exploits these patterns. Unexplored dimensions include:

Q2.1: Can multi-modal coordinated attacks simultaneously perturbing accelerometer, gyroscope, and magnetometer achieve higher success than single-sensor attacks while evading consistency-based detection [41]?

Q2.2: How do body-location-aware strategies adapting perturbations to sensor placement affect attack effectiveness?

Q2.3: At what fusion layer (early, late, attention-based) are multi-modal HAR systems most vulnerable?

Potential approach: Develop attack optimization frameworks that jointly consider sensor modality, body placement, and fusion architecture to maximize attack effectiveness while minimizing detectability.

Recommended evaluation pipeline:

1. Use PAMAP2 and MHealth datasets with all sensor modalities at multiple body locations.
2. Construct transfer matrices measuring ASR retention across all modality pairs.
3. Implement coordinated attacks maintaining inter-sensor correlations (CSCP > 0.9).
4. Evaluate evasion of consistency-based detection [41].
5. Compare attack effectiveness across early, late, and attention-based fusion architectures.

8.4. Gap 3: RL Attacks on Non-RL Systems

RL-based adversarial research almost exclusively targets RL agents. Application to attack DNNs for HAR, time series classification, and forecasting remains virtually unexplored despite natural fit:

Q3.1: Can RL policies learn attack strategies that transfer across HAR models and datasets, addressing the transferability challenge?

Q3.2: How should reward functions encode physical realizability, imperceptibility, and energy constraints for HAR attacks?

$$r_t = \alpha \cdot r_{success} - \beta \cdot r_{imperceptibility} - \gamma \cdot r_{energy} - \delta \cdot r_{physical} \tag{21}$$

Q3.3: Can RL-based attacks learn to defeat specific defenses through continued learning during attack deployment?

Potential approach: Formulate HAR adversarial attack generation as a sequential decision problem where the RL agent selects which time steps and sensors to perturb, with rewards encoding attack success and constraint satisfaction.

Recommended evaluation pipeline:

1. Define MDP: states = (current perturbation, model confidence), actions = (timestep, sensor, magnitude), rewards = weighted combination of success and constraints.
2. Train RL attack policies using PPO/SAC on UCI-HAR and Widar3.0.
3. Evaluate transfer to unseen models and datasets without retraining.
4. Test continued learning against adaptive defenses over multiple attack-defense iterations.
5. Compare query efficiency against gradient-based and search-based baselines.

8.5. Gap 4: Temporal Attack Sophistication

Time series attacks predominantly apply spatial methods frame-by-frame without exploiting temporal structure. Missing capabilities include:

Q4.1: How can attacks optimize perturbations considering causal effects on predictions 10–100 timesteps in the future?

Q4.2: Can attacks identify “critical moments” in activity sequences where perturbations have maximum impact?

Q4.3: How do temporal transferability patterns between activities (e.g., walking→running vs. walking→sitting) affect attack design?

Potential approach: Develop recurrent attack generators that model temporal dependencies in perturbation effectiveness, enabling lookahead optimization for sequential data.

Recommended evaluation pipeline:

1. Use UCR Archive (long sequences) and NTU-120 (variable-length skeleton) datasets.
2. Implement critical timestep identification using gradient magnitude over time.
3. Develop lookahead attack optimization considering future prediction impacts.
4. Compare sparse (critical-only) vs. dense perturbation strategies on ASR vs. sparsity trade-off.
5. Construct activity-pair transfer matrices to characterize temporal transferability patterns.

8.6. Gap 5: Transformer Vulnerabilities

Leite et al. [126] demonstrate transformer-based HAR models show lower robustness than CNNs/LSTMs, but detailed vulnerability analysis is lacking:

Q5.1: Which attention heads are most vulnerable to adversarial manipulation?

Q5.2: Can multi-head exploitation strategies target specific attention patterns to amplify attack effectiveness?

Q5.3: How do positional encoding mechanisms create attack vectors absent in CNN/LSTM architectures?

Potential approach: Develop transformer-specific attacks that analyze and exploit attention patterns, informed by the sharp minima hypothesis underlying transformer vulnerability.

Recommended evaluation pipeline:

1. Train transformer, CNN, and LSTM models on UCI-HAR, NTU-60, and Widar3.0 for controlled comparison.
2. Analyze attention head vulnerability through head-ablation and gradient attribution studies.
3. Develop attention-targeted attacks manipulating specific heads identified as vulnerable.
4. Compare robust accuracy degradation curves across architectures under T-AutoAttack.
5. Investigate positional encoding manipulation as a transformer-specific attack vector.

8.7. Gap 6: Certified Temporal Defenses

Certified defense methods assuming i.i.d. data require extension to temporal data. Despite recent advances [99,100], significant challenges remain:

Q6.1: How can certification methods handle sequential dependencies where perturbations at time t affect predictions at time $t + k$?

Q6.2: Can certified radii be efficiently computed for variable-length sequences exceeding 1000 timesteps?

Q6.3: How should certification methods handle DTW-based perturbation constraints rather than L_p norms?

Potential approach: Develop temporal certification methods that propagate robustness guarantees through recurrent computation, potentially leveraging interval bound propagation or abstract interpretation.

Recommended evaluation pipeline:

1. Extend randomized smoothing to handle temporal correlations using autoregressive noise models.
2. Evaluate on UCR Archive and UCI-HAR with sequence lengths from 100 to 2000 timesteps.
3. Report certified radius vs. clean accuracy trade-off curves.
4. Measure computational scalability: certification time vs. sequence length.
5. Compare L_2, L_∞, and DTW-based certification approaches.

8.8. Gap 7: Physical Attack Validation

Only WiFi CSI [4,50] and radar attacks [8,62] have undergone rigorous physical validation. IMU and medical time series attacks remain largely theoretical:

Q7.1: What hardware capabilities are required to inject adversarial perturbations into wearable IMU sensors?

Q7.2: How do environmental factors (temperature, electromagnetic interference, motion artifacts) affect physical attack success rates?

Q7.3: Would physical attacks be detected by standard sensor quality monitoring systems?

Potential approach: Establish hardware-in-the-loop testbeds for systematic physical attack validation across sensor modalities, with standardized evaluation protocols for reproducibility.

Recommended testbed specifications:

- *IMU testbed:* Controllable vibration platform, EMI injection equipment, reference IMU for ground truth, environmental chamber for temperature/humidity variation.
- *WiFi testbed:* Software-defined radio (USRP), multiple access points, RF shielded chamber, mobility platform for controlled subject motion.
- *Radar testbed:* mmWave radar (TI AWR1642), meta-material tag fabrication capability, anechoic chamber, outdoor test area.
- *Evaluation protocol:* Measure ASR with 95% CI across ≥100 trials per condition, report environmental parameters, include detection analysis.

8.9. Gap 8: Multi-Agent and RLHF Poisoning Scenarios

Growing relevance in autonomous vehicles, multi-robot systems, and LLM alignment demands investigation of coordinated attacks and poisoning scenarios:

Q8.1: How do adversarial perturbations propagate through multi-agent communication channels?

Q8.2: What game-theoretic equilibria emerge in adversarial multi-agent scenarios?

Q8.3: Can collective defense mechanisms provide robustness beyond individual agent defenses?

Potential approach: Extend adversarial policy research [85,87] to realistic multi-agent HAR scenarios such as collaborative activity recognition in smart spaces.

8.9.1. RLHF Poisoning Connections to Safety-Critical Systems

Recent advances in RLHF poisoning attacks [93] have significant implications for safety-critical temporal systems that we highlight explicitly:

Connection to Safety Gym benchmarks: Safety Gym [122] provides constrained RL environments where agents must maximize reward while satisfying safety constraints. RLHF poisoning can corrupt the learned reward model underlying safety constraints, potentially causing agents to violate safety requirements while believing they are acting safely. This creates a particularly insidious attack vector for autonomous systems.

Implications for D4RL offline datasets: Standard offline RL datasets such as D4RL [123] are increasingly used for training policies without online interaction. Xu et al. [91] demonstrated universal reward poisoning attacks on offline RL that:

- Corrupt apparent rewards of high-performing trajectories to appear low-quality.
- Corrupt apparent rewards of low-performing trajectories to appear high-quality.
- Cause offline algorithms to learn suboptimal or dangerous policies from poisoned data.

These attacks are particularly concerning because:

1. *Persistence:* Poisoned datasets remain corrupted indefinitely, affecting all models trained on them.
2. *Scalability:* A single poisoning attack affects all downstream users of the dataset.
3. *Detectability:* Subtle reward modifications may not trigger obvious anomalies during dataset inspection.
4. *Safety implications:* Corrupted safety constraints in RLHF-trained models may not manifest until deployment in safety-critical scenarios.

Recommended evaluation pipeline for RLHF/offline RL security:

1. Evaluate reward poisoning attacks on Safety Gym (PointGoal, CarGoal, DoggoGoal) with constraint violation as primary metric.
2. Test offline RL poisoning on D4RL locomotion and manipulation tasks.
3. Measure attack transferability across offline algorithms (CQL, IQL, TD3+BC).
4. Develop detection methods for reward corruption in offline datasets.
5. Propose certification methods for offline RL with provable robustness to bounded reward perturbation.

8.10. Summary of Prioritized Research Agenda

Based on our gap analysis and prioritization framework, we recommend the following phased research agenda:

Phase 1 (Year 1): Focus on P1 gaps (G1, G2, G5) where datasets and methodological foundations exist. Establish XAI-guided attack frameworks, characterize cross-modal transfer patterns, and systematically analyze transformer vulnerabilities.

Phase 2 (Years 1–2): Address P2 gaps (G3, G4, G6) requiring methodological development. Develop RL-based attack generation, temporal attack sophistication, and extend certified defenses to temporal data.

Phase 3 (Years 2–3): Tackle P3 gaps (G7, G8) requiring infrastructure development. Establish physical attack testbeds, develop multi-agent adversarial evaluation environments, and address RLHF poisoning implications for safety-critical systems.

This phased approach ensures that foundational work enables subsequent investigation, while prioritizing gaps with immediate impact potential.

9. Conclusion

This survey provides the first comprehensive treatment of adversarial attacks on time series and reinforcement learning systems, systematically reviewing 127 papers published between 2019 and 2025. Our analysis reveals both the severity of temporal system vulnerabilities and the substantial gaps in current understanding and defense capabilities.

9.1. Principal Findings

Our systematic review yields seven principal findings, each supported by quantitative evidence synthesized across included studies:

Finding 1: Severe baseline vulnerability. Deep learning models for HAR exhibit extreme vulnerability to adversarial attacks under white-box conditions. FGSM attacks reduce DNN accuracy from 95.1% to 3.4% and CNN accuracy from 93.1% to 16.8% [3]. However, these results must be interpreted within their threat model context—white-box attacks with unconstrained perturbations represent theoretical upper bounds rather than practical threat levels (Section 5).

Finding 2: High variability in cross-domain transfer. Attack transferability varies dramatically across dimensions. Cross-model transfer exceeds 70% for similar architectures, but cross-location transfer ranges from 0% to 80% depending on body placement, and cross-user transfer is generally lower than within-user success [3,29]. This variability creates both challenges (unpredictable attack success) and opportunities (exploitable for defense design) (Section 3).

Finding 3: Physical attack validation gap. A critical disparity exists between digital attack success rates (85–98%) and physically validated attacks. Only WiFi/radar modalities have demonstrated hardware-in-the-loop validation, achieving 70–97% success rates in controlled conditions [4,8]. Wearable IMU physical attacks remain entirely unvalidated, suggesting that immediate threat levels for wearable HAR may be lower than digital results imply (Section 5, Table 8).

Finding 4: Defense effectiveness under adaptive attacks. Current defenses show significant degradation when evaluated against adaptive adversaries. Our analysis reveals 6–23% performance gaps between standard and adaptive attack evaluations across defense categories, with certified defenses exhibiting the smallest gap (6%) and lightweight augmentation methods showing the largest (23%) (Section 6, Table 12).

Finding 5: Architecture-dependent vulnerability. Transformer-based HAR models demonstrate lower adversarial robustness than CNN and LSTM alternatives despite competitive clean accuracy [126]. This finding has important implications for security-sensitive deployments where transformer architectures are increasingly adopted (Section 5).

Finding 6: Emerging LLM-based temporal system risks. The rapid adoption of large language models for time series forecasting and anomaly detection introduces novel vulnerability surfaces including prompt injection, tokenization exploitation, and context window manipulation [10,12]. Security analysis has not kept pace with deployment, creating urgent research needs (Section 2).

Finding 7: RL system compounding vulnerabilities. Reinforcement learning systems face unique sequential vulnerabilities where perturbations at early timesteps compound through decision trajectories. Adversarial policies achieve >95% success against trained agents [85], and reward poisoning can corrupt offline learning from standard datasets [91] (Section 4).

9.2. Key Contributions

This survey contributes:

1. **Unified taxonomy** spanning HAR (IMU, WiFi, radar, skeleton), medical/financial time series, and RL systems, with explicit distinction between on-body and device-free sensing paradigms.
2. **Systematic analysis** of 127 papers with PRISMA-compliant methodology, risk-of-bias assessment, and transparent inclusion/exclusion criteria.
3. **Quantitative synthesis** of attack effectiveness across modalities, validation levels, and threat models, enabling evidence-based threat assessment.
4. **Defense evaluation framework** including adaptive attack benchmarks (T-AutoAttack) and clear distinction between provable certification and implicit robustness approaches.
5. **Prioritized research roadmap** identifying eight critical gaps with specific datasets, evaluation pipelines, and implementation timelines.

9.3. Recommendations for Practitioners

Based on our analysis, we offer the following recommendations for deploying HAR and RL systems in security-sensitive contexts:

For wearable HAR deployments: The gap between theoretical vulnerability and physical attack validation suggests that immediate threat levels may be manageable. Prioritize detection-based defenses exploiting multi-sensor consistency [41], and avoid transformer architectures in favor of CNN/LSTM alternatives until transformer robustness improves.

For WiFi/radar sensing deployments: Physical attacks have been demonstrated with 70–97% success, warranting serious security consideration. Deploy defense-in-depth combining adversarial training with anomaly detection, and monitor RF environments for attack signatures.

For RL system deployments: Adversarial policy and reward poisoning attacks pose significant risks. Employ ensemble methods for observation robustness [108], validate offline datasets for reward corruption before training, and consider certified approaches for safety-critical applications.

For LLM-based temporal systems: Exercise caution in deploying LLM-based forecasters and anomaly detectors in security-sensitive contexts until adversarial robustness properties are better understood. Monitor for prompt injection and ensure robust preprocessing of temporal inputs.

9.4. Limitations of This Survey

We acknowledge several limitations that should inform interpretation of our findings:

Scope limitations: Our focus on HAR, time series classification, and RL excludes related domains such as speech recognition, video understanding beyond skeleton-based methods, and autonomous vehicle perception systems. These domains share temporal characteristics but may exhibit distinct vulnerability patterns.

Selection bias: Prioritization of top-tier venues may exclude relevant work from regional conferences or emerging venues. The rapidly evolving nature of LLM-based temporal systems means that very recent developments may not be captured.

Publication bias: Published studies likely over-represent successful attacks and effective defenses, potentially inflating reported success rates relative to real-world performance.

Quantitative synthesis limitations: Heterogeneity in experimental protocols, perturbation budgets, and evaluation metrics limited our ability to conduct formal meta-analysis. The quantitative synthesis presented should be interpreted as indicative ranges rather than precise estimates.

Evolving threat landscape: Adversarial machine learning is a rapidly evolving field. Findings regarding current defense effectiveness may not generalize to future attack methodologies, particularly as adaptive attacks become more sophisticated.

9.5. Future Research Priorities

Our gap analysis identifies immediate priorities aligned with the seven principal findings:

(1) *XAI-guided attacks and transformer vulnerabilities* (addressing Findings 1, 5) represent high-impact, high-feasibility research directions with available datasets and methodological foundations.

(2) *Sensor-targeted optimization and cross-modal transfer characterization* (addressing Finding 2) will enable both more effective attacks and more targeted defenses.

(3) *Physical attack validation for wearable IMU* (addressing Finding 3) is essential for accurate threat assessment in the most widely deployed HAR modality.

(4) *Adaptive attack evaluation and certified temporal defenses* (addressing Finding 4) will establish reliable defense benchmarks and provable robustness guarantees.

(5) *LLM temporal system security* (addressing Finding 6) requires urgent attention given rapid deployment outpacing security analysis.

(6) *RLHF poisoning and offline RL security* (addressing Finding 7) has implications extending beyond RL to foundation model safety.

9.6. Closing Remarks

The proliferation of deep learning in safety-critical temporal applications—from healthcare monitoring to autonomous navigation to algorithmic trading—makes adversarial robustness a paramount concern. Our systematic review reveals a field with substantial progress in characterizing vulnerabilities, emerging defense mechanisms, and growing awareness of the gap between theoretical and practical threats.

However, significant challenges remain. The disparity between digital and physical attack validation, the vulnerability of emerging architectures (transformers, LLMs), and the absence of standardized evaluation protocols collectively impede progress toward deployable robust systems.

We hope this comprehensive survey catalyzes research addressing these critical gaps, ultimately contributing to temporal systems that are both capable and secure. The prioritized research roadmap and evaluation frameworks we provide offer concrete starting points for advancing this essential agenda.

Acknowledgments

This research was supported by Lembaga Penelitian, Pengabdian, dan Pengembangan Bisnis (LP3B) Institut Teknologi Sains Bandung, Indonesia.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. In Proceedings of the International Conference on Learning Representations (ICLR), 2015.
2. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards deep learning models resistant to adversarial attacks. In Proceedings of the International Conference on Learning Representations (ICLR), 2018.
3. Sah, A.; Ghasemzadeh, H. ADAR: Adversarial Activity Recognition in Wearables. In Proceedings of the IEEE/ACM International Conference on Computer-Aided Design (ICCAD), 2019, pp. 1–8. https://doi.org/10.1109/ICCAD45719.2019.8942124.
4. Zhou, K.; Xing, J.; Luo, X.; Xue, R.; Wang, Z. WiAdv: Practical and robust adversarial attack against WiFi-based gesture recognition system. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **2022**, *6*, 1–25. https://doi.org/10.1145/3534618.
5. Sakka, S.; Liagkou, V.; Stylios, C. Exploiting security issues in human activity recognition systems (HARSs). *Information* **2023**, *14*, 315. https://doi.org/10.3390/info14060315.
6. Ozbulak, U.; Vandersmissen, B.; Jalalvand, A.; Couckuyt, I.; Van Messem, A.; De Neve, W. Investigating the significance of adversarial attacks and their relation to interpretability for radar-based human activity recognition systems. *Computer Vision and Image Understanding* **2021**, *202*, 103111. https://doi.org/10.1016/j.cviu.2020.103111.
7. Rizvani, A.; Apruzzese, G.; Laskov, P. Ephemeral Perturbations: On the Adversarial Security of Deep Learning-Based Algorithmic Trading. In Proceedings of the ACM Conference on Data and Application Security and Privacy (CODASPY), 2025, pp. 1–12. https://doi.org/10.1145/3714393.3726490.
8. Chen, X.; Wang, Z.; Zheng, B.; Chen, Y.; Xu, W.; Miao, C. MetaWave: Attacking mmWave Sensing with Meta-material-enhanced Tags. In Proceedings of the Network and Distributed System Security Symposium (NDSS), 2023. https://doi.org/10.14722/ndss.2023.24348.
9. Govindarajulu, Y.; Amballa, A.; Kulkarni, P.; Parmar, M. Targeted attacks on timeseries forecasting. *arXiv preprint arXiv:2301.11544* **2023**.
10. Alnegheimish, S.; et al. Can Large Language Models be Anomaly Detectors for Time Series? *arXiv preprint arXiv:2405.14755* **2024**.
11. Kong, X.; et al. Deep Learning for Time Series Forecasting: A Survey. *Big Data Mining and Analytics* **2025**.
12. Xiao, Y.; et al. Learn from Adversarial Examples: Learning-Based Attack on Time Series Forecasting. In Proceedings of the Proceedings of AAAI, 2025.

13. 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 IEEE International Joint Conference on Neural Networks (IJCNN), 2019, pp. 1–8. https://doi.org/10.1109/IJCNN.2019.8851936.
14. Akhtar, N.; Mian, A. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access* **2018**, *6*, 14410–14430. https://doi.org/10.1109/ACCESS.2018.2807385.
15. Yuan, X.; He, P.; Zhu, Q.; Li, X. Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems* **2019**, *30*, 2805–2824. https://doi.org/10.1109/TNNLS.2018.2886017.
16. Costa, J.C.; Roxo, T.; Proença, H.; Inacio, P.R.M. How deep learning sees the world: A survey on adversarial attacks & defenses. *IEEE Access* **2024**, *12*, 61113–61136. https://doi.org/10.1109/ACCESS.2024.3395107.
17. Zhang, C.; Zhou, L.; Xu, X.; Wu, J.; Liu, Z. Adversarial attacks of vision tasks in the past 10 years: A survey. *ACM Computing Surveys* **2025**, *58*, 1–42. https://doi.org/10.1145/3708320.
18. Zhang, W.E.; Sheng, Q.Z.; Alhazmi, A.; Li, C. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology* **2020**, *11*, 1–41. https://doi.org/10.1145/3374217.
19. Goyal, S.; Doddapaneni, S.; Khapra, M.M.; Ravindran, B. A survey of adversarial defenses and robustness in NLP. *ACM Computing Surveys* **2023**, *55*, 1–39. https://doi.org/10.1145/3593042.
20. Chakraborty, A.; Alam, M.; Dey, V.; Chattopadhyay, A.; Mukhopadhyay, D. A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology* **2021**, *6*, 25–45. https://doi.org/10.1049/cit2.12028.
21. Pawlicki, M.; Choraś, M.; Kozik, R. A survey on adversarial attacks and defenses in machine learning: Current trends and challenges. *Applied Soft Computing* **2025**, *169*, 112575. https://doi.org/10.1016/j.asoc.2024.112575.
22. Tariq, S.; et al. Towards an Awareness of Time Series Anomaly Detection Models' Adversarial Vulnerability. *arXiv preprint arXiv:2208.11264* **2022**.
23. Dang-Nhu, R.; et al. Adversarial Attacks on Probabilistic Autoregressive Forecasting Models. In Proceedings of the International Conference on Machine Learning, 2020.
24. 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.2018.02.010.
25. Nweke, H.F.; Teh, Y.W.; Al-garadi, M.A.; Alo, U.R. Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. *Expert Systems with Applications* **2018**, *105*, 233–261. https://doi.org/10.1016/j.eswa.2018.03.056.
26. Schott, J.; Deisenroth, M.P.; Serrano, S.A. Robust Deep Reinforcement Learning: A Survey. *arXiv preprint arXiv:2312.08291* **2024**.
27. Wu, Y.; Wang, Y.; Ding, P.; Wang, H.; Zhu, B.; Liu, C. Enhancing Security in Deep Reinforcement Learning: A Comprehensive Survey on Adversarial Attacks and Defenses. *arXiv preprint arXiv:2510.20314* **2025**.
28. Ilahi, I.; Usama, M.; Qadir, J.; Raza, M.U.; Al-Fuqaha, A.; et al. Challenges and Countermeasures for Adversarial Attacks on Deep Reinforcement Learning. *ACM Computing Surveys* **2024**, *56*, 1–37. https://doi.org/10.1145/3633519.
29. Kurniawan, A.; Ohsita, Y.; Murata, M. Experiments on Adversarial Examples for Deep Learning Model Using Multimodal Sensors. *Sensors* **2022**, *22*, 8642. https://doi.org/10.3390/s22228642.
30. Hutchins, J.; et al. Black-Box Adversarial Attacks on Spiking Neural Network for Time Series Data. *arXiv preprint arXiv:2411.01233* **2024**.
31. Liu, S.; et al. Square-Based Black-Box Adversarial Attack on Time Series Classification Using Simulated Annealing and Post-Processing-Based Defense. In Proceedings of the IEEE International Conference on Big Data, 2024.
32. Kitchenham, B.; Charters, S. Guidelines for performing systematic literature reviews in software engineering. *Technical Report EBSE-2007-01, Keele University* **2007**.
33. Lu, Z.; Wang, H.; Chang, Z.; Yang, G.; Li, H.P.H. Hard No-Box Adversarial Attack on Skeleton-Based Human Action Recognition with Skeleton-Motion-Informed Gradient. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4613–4623. https://doi.org/10.1109/ICCV51070.2023.00426.
34. Xie, Y.; Jiang, R.; Guo, X.; Wang, Y.; Cheng, J.; Chen, Y. Universal Targeted Adversarial Attacks Against mmWave-based Human Activity Recognition. In Proceedings of the IEEE Conference on Computer Communications (INFOCOM), 2023, pp. 1–10. https://doi.org/10.1109/INFOCOM53939.2023.10228887.

35. Shahid, A.R.; Arif, M.; Naik, N. Label Flipping Data Poisoning Attack Against Wearable Human Activity Recognition System. In Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI), 2022, pp. 864–871. https://doi.org/10.1109/SSCI51031.2022.10022015.
36. Zheng, Q.; Yu, Y.; Yang, S.; Liu, J.; Lam, K.Y.; Kot, A. Towards physical world backdoor attacks against skeleton action recognition. In Proceedings of the European Conference on Computer Vision (ECCV). Springer, 2024, pp. 215–233.
37. Alfeld, S.; Zhu, X.; Barford, P. Data Poisoning Attacks against Autoregressive Models. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2016, Vol. 30, pp. 1452–1458.
38. Belkhouja, T.; Yan, Y.; Doppa, J.R. Dynamic Time Warping Based Adversarial Framework for Time-Series Domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2023**, *45*, 7353–7366. https://doi.org/10.1109/TPAMI.2022.3224754.
39. Pialla, G.; Fawaz, H.I.; Devanne, M.; Weber, J.; Idoumghar, L.; Muller, P.A.; Bergmeir, C.; Schmidt, D.; Webb, G.; Forestier, G. Smooth perturbations for time series adversarial attacks. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2022, pp. 485–496. https://doi.org/10.1007/978-3-031-05933-9_38.
40. Pialla, G.; Ismail Fawaz, H.; Devanne, M.; Weber, J.; Idoumghar, L.; Muller, P.A.; Bergmeir, C.; Schmidt, D.F.; Webb, G.I.; Forestier, G. Time series adversarial attacks: an investigation of smooth perturbations and defense approaches. *International Journal of Data Science and Analytics* **2025**, *19*, 129–139. https://doi.org/10.1007/s41060-023-00438-0.
41. Kurniawan, A.; Ohsita, Y.; Murata, M. Detection of sensors used for adversarial examples against machine learning models. *Results in Engineering* **2024**, *24*, 103021. https://doi.org/10.1016/j.rineng.2024.103021.
42. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the IEEE Symposium on Security and Privacy (S&P), 2017, pp. 39–57. https://doi.org/10.1109/SP.2017.49.
43. 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. https://doi.org/10.1145/3128572.3140448.
44. Han, X.; Hu, Y.; Foschini, L.; Chinitz, L.; Jankelson, L.; Ranganath, R. Deep learning models for electrocardiograms are susceptible to adversarial attack. *Nature Medicine* **2020**, *26*, 360–363. https://doi.org/10.1038/s41591-020-0791-x.
45. Wang, H.; He, F.; Peng, Z.; Shao, T.; Yang, Y.L.; Zhou, K.; Hogg, D. Understanding the robustness of skeleton-based action recognition under adversarial attack. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14656–14665. https://doi.org/10.1109/CVPR46437.2021.01442.
46. Fursov, I.; Zaytsev, M.; Klyuchnikov, A.; Burnaev, E.; Khomenko, M. Adversarial attacks on deep models for financial transaction records. In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2021, pp. 2868–2878. https://doi.org/10.1145/3447548.3467145.
47. Karim, F.; Majumdar, S.; Darabi, H. Adversarial attacks on time series. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2021**, *43*, 3309–3320. https://doi.org/10.1109/TPAMI.2020.2986319.
48. Ding, D.; Zhang, M.; Feng, F.; Pan, L.; Liu, F.; He, X. Black-box Adversarial Attack on Time Series Classification. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2023, Vol. 37, pp. 7358–7368. https://doi.org/10.1609/aaai.v37i6.25896.
49. Tsingenopoulos, I.; Preuveneers, D.; Joosen, W. AutoAttacker: A reinforcement learning approach for black-box adversarial attacks. In Proceedings of the IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), 2019, pp. 229–237. https://doi.org/10.1109/EuroSPW.2019.00032.
50. Li, M.; Chen, X.; Liu, Y.; Zhang, J.; Zhong, J.; Zhang, Y. Practical Adversarial Attack on WiFi Sensing Through Unnoticeable Communication Packet Perturbation. In Proceedings of the Proceedings of the 30th Annual International Conference on Mobile Computing and Networking (MobiCom), 2024, pp. 315–328. https://doi.org/10.1145/3636534.3649367.
51. Wang, Y.; Du, D.; Hu, H.; Xian, Z.; Liu, M. TSFool: Crafting Highly-Imperceptible Adversarial Time Series through Multi-Objective Attacks. In Proceedings of the European Conference on Artificial Intelligence (ECAI), 2024, pp. 2377–2384. https://doi.org/10.3233/FAIA240644.
52. Diao, Y.; Wang, H.; Shao, T.; Yang, Y.L.; Zhou, K.; Hogg, D. BASAR: Black-box attack on skeletal action recognition. *Pattern Recognition* **2024**, *153*, 110564. https://doi.org/10.1016/j.patcog.2024.110564.
53. Diao, Y.; Wu, B.; Zhang, R.; Yang, X.; Wang, M.; Wang, H. TASAR: Transfer-based Attack on Skeletal Action Recognition. *arXiv preprint arXiv:2409.02483* **2024**. Accepted to ICLR 2025.

54. Tanaka, A.; Kera, H.; Kawamoto, K. Adversarial Bone Length Attack on Action Recognition. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2022, Vol. 36, pp. 2335–2343. https://doi.org/10.1609/aaai.v36i2.20131.
55. Kumar, R.; Isik, C.; Mohan, C.K. Dictionary Attack on IMU-based Gait Authentication. In Proceedings of the Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, 2023, pp. 115–126. https://doi.org/10.1145/3605764.3623913.
56. Ali, H.; Asfoor, S.; Khan, A.U.R. Investigating vulnerabilities of gait recognition model using latent-based perturbations. *Scientific Reports* **2025**, *15*, 22869. https://doi.org/10.1038/s41598-025-22869-4.
57. Xu, C.; Wang, Z.; Chen, S. WiCAM: Imperceptible adversarial attack on deep learning based WiFi sensing. In Proceedings of the IEEE International Conference on Sensing, Communication, and Networking (SECON), 2022, pp. 217–225. https://doi.org/10.1109/SECON55815.2022.9918564.
58. Sharma, A.; Mishra, D.; Jha, S.; Seneviratne, A. Wi-Spoof: Generating adversarial wireless signals to deceive Wi-Fi sensing systems. *Journal of Information Security and Applications* **2025**, *91*, 104052. https://doi.org/10.1016/j.jisa.2025.104052.
59. Huang, P.; Zhang, X.; Yu, S.; Guo, L. IS-WARS: Intelligent and Stealthy Adversarial Attack to Wi-Fi-Based Human Activity Recognition Systems. *IEEE Transactions on Dependable and Secure Computing* **2022**, *19*, 3899–3912. https://doi.org/10.1109/TDSC.2021.3110480.
60. Yang, J.; Zou, H.; Cao, S.; Chen, Z.; Xie, L. SecureSense: Defending adversarial attack for secure device-free human activity recognition. *IEEE Transactions on Mobile Computing* **2024**, *23*, 1922–1936. https://doi.org/10.1109/TMC.2022.3226742.
61. Cao, H.; Zhang, Y.; Xu, W. Selective Adversarial Training for Robust Wireless Human Activity Recognition. *IEEE Transactions on Mobile Computing* **2024**, *23*, 9876–9889. https://doi.org/10.1109/TMC.2024.3420405.
62. Xu, Y.; Liu, Y.; Zhou, J.; Kohno, T. TileMask: A Passive-Reflection-Based Attack against mmWave Radar Object Detection in Autonomous Driving. In Proceedings of the ACM Conference on Computer and Communications Security (CCS), 2023, pp. 1–15. https://doi.org/10.1145/3576915.3616661.
63. Kuzlu, M.; Catak, F.O.; Cali, U.; Catak, E.; Guler, O. Adversarial security mitigations of mmWave beamforming prediction models using defensive distillation and adversarial retraining. *International Journal of Information Security* **2023**, *22*, 319–332. https://doi.org/10.1007/s10207-022-00643-5.
64. Chen, Z.; Xie, Y.; Huang, H.; Jing, W.; Zheng, W.S. Appending adversarial frames for universal video attack. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3199–3208. https://doi.org/10.1109/WACV48630.2021.00324.
65. Wei, Z.; Chen, J.; Wu, Z.; Jiang, Y.G. Boosting the transferability of video adversarial examples via temporal translation. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2022, Vol. 36, pp. 2659–2667. https://doi.org/10.1609/aaai.v36i3.20188.
66. Wei, Z.; Chen, J.; Wei, X.; Jiang, L.; Chua, T.S.; Zhou, F.; Jiang, Y.G. Cross-modal transferable adversarial attacks from images to videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15064–15073. https://doi.org/10.1109/CVPR52688.2022.01464.
67. Wei, Z.; Chen, J.; Wu, Z.; Jiang, Y.G. Adaptive Cross-Modal Transferable Adversarial Attacks From Images to Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2024**, *46*, 3206–3221. https://doi.org/10.1109/TPAMI.2023.3347835.
68. Kim, H.S.; Son, M.; Kim, M.; Kwon, M.J.; Kim, C. Breaking Temporal Consistency: Generating Video Universal Adversarial Perturbations Using Image Models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4325–4334. https://doi.org/10.1109/ICCV51070.2023.00399.
69. Hwang, J.; Zhang, H.; Oh, S. Can temporal shuffling defend against adversarial attacks on video action recognition? *Neural Networks* **2024**, *171*, 29–45. https://doi.org/10.1016/j.neunet.2023.10.033.
70. Chen, X.; Si, Y.; Wang, P.; Zheng, W.; Gu, X. Improving Adversarial Robustness of ECG Classification Based on Lipschitz Constraints and Channel Activation Suppression. *Sensors* **2024**, *24*, 2954. https://doi.org/10.3390/s24092954.
71. Shao, J.; Geng, S.; Zheng, Z.; Fu, X.; Yu, W.; Xia, S. CardioDefense: Defending against adversarial attack in ECG classification with adversarial distillation training. *Biomedical Signal Processing and Control* **2024**, *88*, 105922. https://doi.org/10.1016/j.bspc.2023.105922.
72. Wiedeman, C.; Wang, G. Decorrelative network architecture for robust electrocardiogram classification. *Patterns* **2024**, *5*, 101117. https://doi.org/10.1016/j.patter.2024.101117.

73. Meng, L.; Jiang, X.; Chen, X.; Wang, Y.; Wu, D. EEG-Based Brain-Computer Interfaces are Vulnerable to Backdoor Attacks. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **2023**, *31*, 2224–2234. https://doi.org/10.1109/TNSRE.2023.3273214.
74. Meng, L.; Jiang, X.; Chen, X.; Liu, W.; Luo, H.; Wu, D. Adversarial filtering based evasion and backdoor attacks to EEG-based brain-computer interfaces. *Information Fusion* **2024**, *107*, 102316. https://doi.org/10.1016/j.inffus.2024.102316.
75. Meng, L.; Jiang, X.; Wu, D. Adversarial robustness benchmark for EEG-based brain-computer interfaces. *Future Generation Computer Systems* **2023**, *143*, 231–247. https://doi.org/10.1016/j.future.2023.01.022.
76. Xie, Y.; Shi, D.; Lian, D.; Li, G.; Chen, X. A word is worth a thousand dollars: Adversarial attack on tweets fools stock prediction. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2022, pp. 596–606. https://doi.org/10.18653/v1/2022.naacl-main.46.
77. Liu, L.; Park, Y.; Hoang, T.N.; Zenati, H.; Yoon, J. Robust Multivariate Time-Series Forecasting: Adversarial Attacks and Defense Mechanisms. In Proceedings of the International Conference on Learning Representations (ICLR), 2023.
78. Liu, F.; Jiang, S.; Vo, K.; Xiang, C.; Tay, Y.; Vo, H. Adversarial Vulnerabilities in Large Language Models for Time Series Forecasting. *arXiv preprint arXiv:2412.08099* **2024**. Accepted to AISTATS 2025.
79. Zhang, H.; Chen, H.; Xiao, C.; Li, B.; Liu, M.; Boning, D.; Hsieh, C.J. Robust deep reinforcement learning against adversarial perturbations on state observations. In Proceedings of the Advances in Neural Information Processing Systems, 2020, Vol. 33, pp. 21024–21037.
80. Sun, J.; Zhang, T.; Xie, X.; Ma, L.; Zheng, Y.; Chen, K.; Liu, Y. Stealthy and efficient adversarial attacks against deep reinforcement learning. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2020, Vol. 34, pp. 5883–5891. https://doi.org/10.1609/aaai.v34i04.6047.
81. Zhang, H.; Chen, H.; Boning, D.; Hsieh, C.J. Robust reinforcement learning on state observations with learned optimal adversary. In Proceedings of the International Conference on Learning Representations (ICLR), 2021.
82. Oikarinen, T.; Zhang, W.; Megretski, A.; Daniel, L.; Weng, T.W. Robust deep reinforcement learning through adversarial loss. In Proceedings of the Advances in Neural Information Processing Systems, 2021, Vol. 34, pp. 26156–26167.
83. Korkmaz, E.; Brown-Cohen, J. Detecting Adversarial Directions in Deep Reinforcement Learning to Make Robust Decisions. In Proceedings of the International Conference on Machine Learning (ICML), 2023, pp. 17534–17543.
84. Liu, Y.; Liu, J.; Chen, X.; Bai, Y.; Yu, K. Diffusion Policy Attacker: Crafting Adversarial Attacks for Diffusion-based Policies. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2024, Vol. 37.
85. Gleave, A.; Dennis, M.; Wild, C.; Kant, N.; Levine, S.; Russell, S. Adversarial policies: Attacking deep reinforcement learning. In Proceedings of the International Conference on Learning Representations (ICLR), 2020.
86. Liu, G.; Lai, L. Efficient Adversarial Attacks on Online Multi-agent Reinforcement Learning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2023, Vol. 36.
87. Ma, O.; Pu, Y.; Zhao, L.; Hang, R.; Hu, G.; Wang, Z.; Yang, L.; Chen, H. SUB-PLAY: Adversarial Policies against Partially Observed Multi-Agent Reinforcement Learning Systems. In Proceedings of the ACM Conference on Computer and Communications Security (CCS), 2024, pp. 1012–1026. https://doi.org/10.1145/3658644.3670293.
88. Xue, L.; Ma, W.; Zou, S. Robust Communicative Multi-Agent Reinforcement Learning with Active Defense. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 16107–16115. https://doi.org/10.1609/aaai.v38i14.29518.
89. Liu, X.; Deng, C.; Sun, Y.; Liang, Y.; Huang, F. Beyond worst-case attacks: Robust RL with adaptive defense via non-dominated policies. *arXiv preprint arXiv:2402.12673* **2024**.
90. Rakhsha, A.; Radanovic, G.; Devidze, R.; Zhu, X.; Singla, A. Policy teaching via environment poisoning: Training-time adversarial attacks against reinforcement learning. In Proceedings of the International Conference on Machine Learning (ICML), 2020, pp. 7974–7984.
91. Xu, Y.; Gumaste, R.; Singh, G. Universal Black-Box Reward Poisoning Attack against Offline Reinforcement Learning. *arXiv preprint arXiv:2402.09695* **2024**.

92. Rathbun, E.; Amato, C.; Oprea, A. SleeperNets: Universal Backdoor Poisoning Attacks Against Reinforcement Learning Agents. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2024, Vol. 37.
93. Wang, J.; Wu, J.; Li, M.; Jia, Y.; Xiao, C.; Chen, H.; Jia, R. RLHFPoison: Reward Poisoning Attack for Reinforcement Learning with Human Feedback in Large Language Models. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 2024, pp. 2468–2488. https://doi.org/10.18653/v1/2024.acl-long.140.
94. Chen, K.; Wei, Z.; Chen, J.; Wu, Z.; Jiang, Y.G. Attacking video recognition models with bullet-screen comments. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2022, Vol. 36, pp. 312–320. https://doi.org/10.1609/aaai.v36i1.19916.
95. García, J.; Majadas, R.; Fernández, F. Learning adversarial attack policies through multi-objective reinforcement learning. *Engineering Applications of Artificial Intelligence* **2020**, *96*, 104021. https://doi.org/10.1016/j.engappai.2020.104021.
96. Song, J.; Yu, D.; Teng, H.; Chen, Y. RLVS: A Reinforcement Learning-Based Sparse Adversarial Attack Method for Black-Box Video Recognition. *Electronics* **2025**, *14*, 245. https://doi.org/10.3390/electronics14020245.
97. 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 2017 IEEE European symposium on security and privacy (EuroS&P). IEEE, 2017, pp. 3–18.
98. Cohen, J.; Rosenfeld, E.; Kolter, Z. Certified adversarial robustness via randomized smoothing. In Proceedings of the International Conference on Machine Learning (ICML), 2019, pp. 1310–1320.
99. Dong, W.; Chen, W.; Li, Z.; Huang, D. Boosting Certified Robustness for Time Series Classification with Efficient Self-Ensemble. In Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM), 2024, pp. 523–532. https://doi.org/10.1145/3627673.3679748.
100. Franco, N.; Spiegelberg, J.; Lorenz, J.M.; Günnemann, S. Guaranteeing Robustness Against Real-World Perturbations In Time Series Classification Using Conformalized Randomized Smoothing. In Proceedings of the Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence (UAI), 2024, Vol. 244, *PMLR*, pp. 1371–1388.
101. Belkhouja, T.; Doppa, J.R. Adversarial framework with certified robustness for time-series domain via statistical features. *Journal of Artificial Intelligence Research* **2022**, *73*, 1435–1471. https://doi.org/10.1613/jair.1.13212.
102. Han, Y.; Chen, W. Lightweight Defense Against Adversarial Attacks in Time Series Classification. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 2025, pp. 1–13. arXiv:2505.02073.
103. Wang, H.; Diao, Y.; Tan, Z.; Guo, G. Defending Black-box Skeleton-based Human Activity Classifiers. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2023, Vol. 37, pp. 2551–2559. https://doi.org/10.1609/aaai.v37i3.25352.
104. Krishan, P.; Nathani, R.; Sengupta, S. Adversarial Attacks and Defenses in Multivariate Time-Series Forecasting for Smart and Connected Infrastructures. In Proceedings of the Proceedings of the Annual Conference of the PHM Society, 2024, Vol. 16. https://doi.org/10.36001/phmconf.2024.v16i1.4082.
105. Gong, X.; Chen, Y.; Wang, X.; Gao, Y. BIRD: Generalizable Backdoor Detection and Removal for Deep Reinforcement Learning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2023, Vol. 36.
106. Zhang, W.; Li, Y.; Chen, H. Diversity supporting adversarial robustness in ensemble methods. *Computers & Security* **2024**, *140*, 103862. https://doi.org/10.1016/j.cose.2024.103862.
107. Li, Z.; Piao, S.; Dong, C.; Chen, W. Robustness Analysis on Self-ensemble Models in Time Series Classification. In Proceedings of the Australasian Database Conference (ADC), 2025, pp. 3–16. https://doi.org/10.1007/978-981-96-1242-0_1.
108. Wang, Y.; Zhang, K.; Li, M. Defending Against Adversarial Attacks in Deep Reinforcement Learning via Ensemble Defense. *arXiv preprint arXiv:2507.17070* **2024**.
109. 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.
110. 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, 2020, pp. 2206–2216.

111. Dau, H.A.; Bagnall, A.; Kamgar, K.; Yeh, C.C.M.; Zhu, Y.; Gharghabi, S.; Ratanamahatana, C.A.; Keogh, E. The UCR time series archive. *IEEE/CAA Journal of Automatica Sinica* **2019**, *6*, 1293–1305. https://doi.org/10.1109/JAS.2019.1911747.
112. Anguita, D.; Ghio, A.; Oneto, L.; Parra, X.; Reyes-Ortiz, J.L. A public domain dataset for human activity recognition using smartphones. In Proceedings of the European Symposium on Artificial Neural Networks (ESANN), 2013, pp. 437–442.
113. Reiss, A.; Stricker, D. Introducing a new benchmarked dataset for activity monitoring. In Proceedings of the International Symposium on Wearable Computers (ISWC), 2012, pp. 108–109. https://doi.org/10.1109/ISWC.2012.13.
114. Roggen, D.; Calatroni, A.; Rossi, M.; Holleczek, T.; Förster, K.; Tröster, G.; Lukowicz, P.; Bannach, D.; Pirkl, G.; Ferscha, A.; et al. The OPPORTUNITY challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters* **2013**, *34*, 2033–2042. https://doi.org/10.1016/j.patrec.2012.12.014.
115. Banos, O.; 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. *Lecture Notes in Computer Science* **2014**, *8868*, 91–98. https://doi.org/10.1007/978-3-319-13105-4_14.
116. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. NTU RGB+D: A large scale dataset for action recognition **2016**. pp. 1010–1019. https://doi.org/10.1109/CVPR.2016.115.
117. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. The Kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* **2017**.
118. Zheng, Y.; Zhang, Y.; Qian, K.; Zhang, G.; Liu, Y.; Wu, C.; Yang, Z. Widar3.0: Zero-effort cross-domain gesture recognition with Wi-Fi. 2022, Vol. 44, pp. 8671–8688. https://doi.org/10.1109/TPAMI.2021.3105387.
119. Ma, Y.; Zhou, G.; Wang, S.; Zhao, H.; Jung, W. SignFi: Sign language recognition using WiFi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **2018**, *2*, 1–21. https://doi.org/10.1145/3191755.
120. Clifford, G.D.; Liu, C.; Moody, B.; Lehman, L.w.H.; Silva, I.; Li, Q.; Johnson, A.E.; Mark, R.G. AF classification from a short single lead ECG recording: the PhysioNet/Computing in Cardiology Challenge 2017. *Computing in Cardiology* **2017**, *44*, 1–4. https://doi.org/10.22489/CinC.2017.065-469.
121. Shoeb, A.H. CHB-MIT scalp EEG database. *PhysioNet* **2010**. https://physionet.org/content/chbmit/1.0.0/.
122. Ray, A.; Achiam, J.; Amodei, D. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708* **2019**.
123. Fu, J.; Kumar, A.; Nachum, O.; Tucker, G.; Levine, S. D4RL: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219* **2020**.
124. Tavakkoli, V.; Amirkabiri, A.; Heydarian, H. A Novel Deep Learning Framework for Human Activity Recognition Using Wavelet-Based CNN and LSTM. *IEEE Access* **2023**, *11*, 73185–73197.
125. Roy, D.; Mukherjee, T.; Chatterjee, M.; Bhattacharyya, S.; Bandyopadhyay, S. MixHAR: Mixing Temporal Experts for Human Activity Recognition from Accelerometer and Gyroscope Data. *Neural Computing and Applications* **2022**, *34*, 13763–13778.
126. Leite, C.F.S.; Xiao, Y.; Yigitler, H.; Jäntti, R. Transformer-Based Approaches for Sensor-Based Human Activity Recognition: Opportunities and Challenges. *arXiv preprint arXiv:2410.13605* **2024**.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.