Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence
Canadian AI 2026
25–29 May 2026,Vancouver,BC

Vulnerability of machine learning models for gender recognition in Virtual Reality

Emna Kraiem* and Alan Davoust*
Université du Québec en Outaouais
Gatineau, Québec, Canada

Abstract

Virtual Reality (VR) systems continuously capture fine-grained behavioral signals such as head motion, hand trajectories, and gaze dynamics. These spatio-temporal signals have been shown to contain distinctive patterns enabling accurate gender classification through machine learning models. While predictive performance under nominal conditions is often high, the robustness of such models to adversarial behavior remains largely unexplored.

In this paper, we present a systematic robustness analysis of VR-based gender classification models under a comprehensive catalog of realistic behavioral adversarial attacks. We evaluate multiple model families, including ensemble-based tabular classifiers and neural architectures, using statistical and dynamic motion features extracted from public VR datasets. More than a dozen perturbation scenarios targeting metric coherence, global motion style, multimodal synchronization, and latent behavioral structure are assessed using balanced accuracy, flip rate, and confidence stability metrics.

Our results reveal significant vulnerability to coordinated, structurally consistent attacks, particularly those affecting global motion properties or metric integrity, while localized noise-like perturbations exhibit limited impact. These findings demonstrate that high nominal accuracy does not guarantee robustness and highlight the necessity of robustness-aware evaluation frameworks for VR-based behavioral inference systems.

Keywords: Virtual Reality, Gender Classification, Behavioral Biometrics, Adversarial Attacks, Robustness, Machine Learning, Motion Analysis

1. Introduction

Platforms for virtual reality (VR) gather high-resolution behavioral data, such as gaze dynamics, hand trajectories, and head motion. Subtle motor patterns that represent unique traits are captured by these spatiotemporal signals. According to recent research, these VR tracking signals can be used to accurately predict sensitive user features, such as gender [1–3].

Before allowing access to particular social or limited features, gender inference may be employed as part of identification verification procedures in real-world immersive environments. For example, a system might use behavioral signals to try and automatically confirm that a user is a woman if they portray themselves as such in a virtual setting. This poses an important security query: *is it possible for a VR user to purposefully alter their behavior to trick a gender classification system?*

This question carries direct relevance for platform integrity: gender-based access controls and identity verification mechanisms are increasingly deployed in VR environments to prevent impersonation, enforce community guidelines, and ensure safe participation in gender-restricted spaces such as women-only support groups or secure corporate meetings. Understanding the fragility of such systems is therefore not merely an academic exercise, but a practical security concern with real-world consequences for both platform designers and end users. We emphasize that our goal is not to advocate for behavioral profiling, but to expose vulnerabilities in deployed inference pipelines so that more robust and privacy-aware alternatives can be developed.

* krae01@uqo.ca, alan.davoust@uqo.ca

While prior research has demonstrated the feasibility of gender prediction from VR motion data [1, 2], the robustness of such classifiers against voluntary behavioral manipulations remains largely unexplored. In other domains, adversarial machine learning has shown that structured input perturbations can significantly degrade classifier performance [4, 5]. However, most adversarial studies focus on domains such as image classification or spam detection, where inputs can be modified directly and arbitrarily as long as they continue to fulfill their functional purpose. In image recognition, pixel-level perturbations can be added without affecting human perception of the object. Similarly, in spam detection, textual content can be strategically altered by inserting benign words or obfuscating keywords while preserving the message's intent. In these settings, the adversary operates directly on the digital artifact.

VR behavioral data, on the other hand, are motor signals produced by embodied physical engagement rather than static artifacts intended for human interpretation. Any alteration must be effected by means of real adjustments to limb coordination, gaze strategy, posture, or gesture amplitude. The altered behavior must be physically produced in real time; the attacker cannot change feature values at will. This leads to a fundamental difference in the appropriate threat model, which is limited not only by imperceptibility but also by biomechanics, coordination, and task feasibility.

To the best of our knowledge, no previous research has systematically compared gender classification algorithms in immersive virtual reality environments under organized, behaviorally reasonable perturbations.

In this work, we study *behavioral attacks* that a virtual reality user can realistically perform in a black-box setting, without access to model parameters. Rather than crafting imperceptible gradient-based perturbations, we investigate whether physically plausible and voluntarily enacted behavioral changes can meaningfully alter model decisions.

We first conduct a comprehensive evaluation of multiple model families for VR-based gender prediction, including temporal neural architectures, multilayer perceptrons, and ensemble-based tabular classifiers.

Then, the robustness of these models is assessed under a diverse set of algorithmically generated, user-executable perturbations. Performance degradation is quantified using balanced accuracy, flip rate, and confidence calibration. We identify perturbation categories that severely destabilize predictions, and characterize robustness trade-offs in immersive behavioral inference.

Our results show that, although several architectures reach high nominal performance (up to 93% balanced accuracy), none remain stable under structured behavioral perturbations. Coherent global changes such as posture neutralization or multimodal desynchronization lead to pronounced performance drops, in some cases approaching chance level. This suggests that behavioral inference in VR environments remains fundamentally sensitive to intentional user manipulation.

These findings serve as a cautionary signal for the broader field: the deployment of behavioral inference systems in immersive environments must account for the inherent manipulability of motor signals. We call for a shift toward adversarially robust model design and for greater scrutiny of the assumptions underlying behavioral biometrics in VR.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the comparative evaluation of gender prediction models under nominal conditions. Section 4 introduces the robustness analysis under voluntary behavioral perturbations. Finally, Section 5 discusses the broader implications for VR-based gender inference systems and concludes the paper.

2. Related Work

2.1. Gender Prediction from Virtual Reality Signals

Public datasets such as Alyx [6] and FAST [7] have enabled systematic evaluation of gender identification and user recognition tasks in VR. These benchmarks provide multi-session motion tracking signals and support comparisons across models and representations.

Previous studies have demonstrated that eye-tracking data, head and hand tracking signals, and motion trajectories contain sufficiently distinctive patterns to infer sensitive user attributes, including gender [1, 2, 8, 9]. Although high predictive performance is frequently reported, substantial generalization gaps across domains and experimental setups highlight the contextual sensitivity of current models.

The discriminative capacity of behavioral signals is further supported by related work on gait and motion-based recognition [3], while eye-tracking dynamics have also been investigated as complementary cues for gender inference [10]. However, most existing studies evaluate prediction accuracy under nominal conditions. Robustness to intentional and physically executable behavioral modification remains largely unexplored.

2.2. Adversarial Attacks on Behavioral and Temporal Data

Adversarial machine learning has demonstrated that structured perturbations can significantly degrade classification systems. Early work showed that modifying a small subset of discriminative features may induce misclassification [11]. These ideas were later extended to black-box settings [12, 13] and to tree-based models vulnerable to coordinated feature manipulation [14]. In image classification, gradient-based attacks such as DeepFool [5] and PGD [4] illustrate how subtle pixel-level perturbations can reliably alter predictions.

Adversarial strategies have also been applied to time-series and motion data, where perturbations are constrained to preserve signal plausibility while altering temporal structure [15]. Similar approaches have been proposed for skeletal action recognition [16] and wearable sensor systems [17].

In immersive environments, latency and tracking instability have primarily been studied as technical constraints [18, 19]. Behavioral biometrics research further suggests that voluntary motor pattern modifications may suffice to evade classification systems without artificial numerical noise [20].

Most adversarial studies assume direct numerical manipulation of inputs. In VR, motion signals are produced through physical user movement and cannot be arbitrarily edited. The perturbations considered in this work are therefore restricted to behaviorally realizable transformations. Each attack applies bounded shifts, rescaling, or temporal offsets to interpretable feature groups (e.g., posture height, velocity statistics, variability measures, or limb synchronization). The magnitude of these transformations is constrained to reflect voluntary adjustments that a user could enact during normal interaction. The objective is not to generate abstract worst-case perturbations, but to evaluate whether structured behavioral modifications are sufficient to degrade model predictions.

3. Comparative Evaluation of Gender Prediction Models

3.1. Datasets and Data Structure

We conduct our experiments using two public VR tracking datasets, *FAST* [21] and *Who Is Alyx?* [22]. The latter contains 71 unique VR users recorded across two separate gameplay sessions each, resulting in 142 session recordings and more than 110 hours of multimodal tracking data. Raw recordings are organized hierarchically by `player_id` and `session_id`. Each session includes synchronized streams of head-mounted display (HMD) and controller

positions/rotations, button states, and eye-tracking measurements (gaze trajectories, fixation points, pupil dynamics). Sessions typically last approximately 45 minutes and are sampled at high frequency ($\approx$ 90Hz), yielding tens of thousands of time steps per session.

The *FAST* dataset contains 108 participants performing two full-scale assembly tasks in immersive VR, resulting in at least 216 multimodal session recordings. Similar to Alyx, the dataset provides high-frequency tracking of head and hand kinematics, enabling the extraction of statistical and dynamic behavioral descriptors. Although session durations and motion intensity vary depending on task complexity, the acquisition protocol and sensor modalities remain consistent across users, ensuring a homogeneous structure for sample construction. Both datasets include some demographic information, and the binary target label `gender` is used for classification[1].

VR behavioral signals exhibit strong within-subject consistency. If segments from the same user appear across training and evaluation sets, models may implicitly learn identity-specific motion signatures rather than generalizable gender-related cues. To prevent such identity leakage, all Train/Validation/Test splits are performed **strictly at the player level**: sessions belonging to a given `player_id` never appear in more than one split. This player-independent protocol follows methodological recommendations for VR behavioral modeling and evaluation [2, 23] and ensures that reported performance reflects generalization to unseen users.

3.2. Feature Representation and Selection

VR tracking datasets are made up of numerous synchronized sensor streams and lengthy multivariate time series with thousands of time steps each session. Because traditional machine learning algorithms require fixed-dimensional input vectors, they are unable to directly use these raw recordings, which vary in duration between sessions.

Thus, two modeling approaches are taken into account. First, each session must be summarized into a single feature vector for tabular models (such as Random Forest, Gradient Boosting, SVM, and MLP). Second, although they require organized fixed-length input segments, temporal neural architectures (such as LSTM, CNN–LSTM, ResNet1D, and TimesNet) work directly on sequential data. Therefore, feature engineering is required to control dimensionality, convert unstructured multimodal motion signals into structured representations, and allow for equitable comparison across model families.

Continuous signals are used to extract session-level statistical characteristics for tabular models. First- and second-order statistics (mean, standard deviation, minimum, maximum), robust dispersion measures (range and IQR), and distributional shape descriptors (skewness and kurtosis) are calculated for every temporal variable. Derivative-based features including velocity, acceleration, and global signal energy are used to capture motion dynamics. To represent fine-grained motor properties, additional behavioral descriptors are included, such as entropy, smoothness indices, and straightness ratio [3, 24]. One fixed-dimensional vector is created for each session by aggregating all features at the session level (`player_id`/`session_id`) [25].

In the high-dimensional domain ($p \gg n$), supervised feature selection is incorporated into the learning pipeline to reduce overfitting and enhance generalization. With `SelectKBest` [26, 27], we use the ANOVA F-test, treating the amount of retained features k as a hyperparameter that is optimized by nested cross-validation. A nested evaluation technique is used [28, 29] to prevent optimistic bias during joint adjustment of model hyperparameters and feature dimensionality. While the outer loop offers objective performance estimates, the inner

[1] it is unclear from the dataset description whether the subjects' biological sex sometimes differs from their gender identity, and it is possible that physiological signals more accurately predict biological sex rather than gender. However, *gender* is the attribute found in the data and it is therefore the term we use here.

`GroupKFold` loop (grouped by `player_id`) tunes the model and feature selection. A regulated and equitable comparison across tabular and temporal model families under equal experimental settings is ensured by this unified evaluation approach.

3.3. **Classical and Neural Tabular Models**

We evaluate several supervised classifiers trained on the session-level tabular representation, including Random Forest, Gradient Boosting, XGBoost, LightGBM, linear and RBF-kernel SVM, and two multilayer perceptrons (MLP).The first MLP corresponds to the shallow implementation provided by *scikit-learn*, consisting of a single hidden layer of size 100 with ReLU activation. The second MLP is a deeper fully-connected neural network implemented in *PyTorch*. It contains two hidden layers: the first with 128 neurons and the second with 64 neurons, each followed by ReLU activation and a dropout layer (dropout rate = 0.3). The network outputs a 2-dimensional logit vector for binary classification. The distinction between the two models is therefore architectural (depth and representational capacity), rather than library-related.

3.3.1. **Training and Evaluation Protocol**

All experiments follow a strictly player-independent protocol. Data are split at the `player_id` level to prevent identity leakage, ensuring that sessions from a given user appear in only one partition. The dataset is divided into 75% training players and 25% held-out test players.

Nested cross-validation is performed exclusively on the 75% training portion. The outer 5-fold *GroupKFold* (protected by `player_id`) is used to estimate validation performance, while the inner 5-fold *GroupKFold* is used for joint hyperparameter tuning and feature selection. Feature selection (`SelectKBest`) is re-fitted at each inner fold to prevent data leakage.

Balanced accuracy is used as the primary evaluation metric. After nested cross-validation, the best hyperparameter configuration is retrained on the full training set and evaluated once on the held-out 25% test set to obtain the final unbiased performance.

All models are trained using the player-independent 75% / 25% split described in Section 3.3. Hyperparameters and the number of selected features k are tuned using nested cross-validation [28, 29]. Performance is measured using balanced accuracy, with nominal results summarized in Table 1.

3.3.2. **Results**

Table 1. Tabular model performance (player-independent test set).

Model	k	**CV BalancedAcc**	**Test BalancedAcc**
Random Forest	418	0.83 ± 0.04	**0.93**
Gradient Boosting	205	0.83 ± 0.06	0.91
SVM (RBF)	33	0.82 ± 0.06	0.89
MLP (sklearn)	18	0.81 ± 0.08	0.88
XGBoost	121	0.83 ± 0.08	**0.93**
LightGBM	115	0.82 ± 0.08	0.89
MLP (deep, PyTorch)	418	0.84 ± 0.05	0.91

Ensemble models (Random Forest and XGBoost) achieve the best generalization performance (BalancedAcc $\approx$ 0.93), indicating strong robustness to inter-user variability. The RBF-SVM remains competitive while relying on a compact feature subset, suggesting that discriminative structure is concentrated in a limited number of descriptors.

The shallow *scikit-learn* MLP achieves approximately 88% balanced accuracy, while the deeper PyTorch MLP reaches about 91%. This improvement suggests that increased representational capacity can capture additional nonlinear feature interactions, although the gain remains moderate given the dataset size.

Overall, the consistency between nested CV estimates and held-out test performance indicates limited overfitting and validates the player-independent evaluation protocol.

3.4. **Sequence Models**

In contrast to tabular aggregation, sequence models operate directly on frame-level multivariate time series, preserving temporal ordering and multimodal synchronization. Such architectures are widely used for motion modeling and time-series classification tasks [30, 31]. Each session is segmented into fixed-length windows (`SEQ_LEN` = 100, stride = 25), generating overlapping temporal segments while preserving strict player-level separation between training and test sets. After windowing, feature-wise normalization is applied by flattening the data from (N, T, F) to $(N \cdot T, F)$ to stabilize optimization. Such normalization strategies are known to improve convergence and robustness in sequential deep learning pipelines [32, 33]. We evaluate four temporal architectures: a two-layer LSTM with Layer Normalization, dropout, gradient clipping, learning-rate scheduling, and early stopping; a CNN–LSTM architecture combining local convolutional feature extraction with recurrent temporal modeling; a ResNet1D model for deep residual learning on time-series data [30]; and TimesNet, designed to capture multi-scale temporal patterns [31]. Recurrent architectures such as LSTM are well-established for modeling medium- and long-range temporal dependencies in sequential data [34]. Table 2 reports the nominal performance.

Table 2. Sequence model performance (held-out test set).

Model	**Test Accuracy**
LSTM (2-layer)	0.89
CNN–LSTM	0.83
ResNet1D	0.85
TimesNet	0.80

Temporal models achieve slightly lower performance than tabular ensembles, suggesting that session-level aggregation captures strong global discriminative cues. The LSTM reaches approximately 89% accuracy, indicating that medium-range temporal dependencies (50–100 frames) provide useful gender-related information. Architectures emphasizing local convolutional structure (CNN–LSTM, ResNet1D) show moderate performance, while TimesNet achieves around 80% accuracy, possibly reflecting high inter-subject variability and limited periodic structure in VR motion.

3.5. **Ensemble Tabular Classifier**

We combine many tabular models, including Random Forest, XGBoost, Gradient Boosting, and SVM-RBF, to create a soft *VotingClassifier*. These models promote error diversity and enhance stability through probabilistic aggregation by utilizing complimentary learning techniques (bagging, boosting, and margin maximization). The voting ensemble successfully classifies 22/23 samples for class 0 and 19/21 samples for class 1, achieving around 93% test accuracy. With an AUC of 0.988, the ROC curve shows good discriminative performance. The stabilizing impact of soft voting, in which model-specific errors are partially offset by probability aggregation, is reflected in this improvement.

3.6. **Nominal Performance Comparison**

Under nominal (non-adversarial) conditions, performance remains consistently high across model families. Ensemble-based tabular models (Random Forest and XGBoost) reach approximately 93% balanced accuracy on the player-independent test set. Gradient Boosting achieves around 91%, while LightGBM, SVM (RBF), and the scikit-learn MLP perform between 88% and 89%. The PyTorch MLP attains approximately 91%, confirming strong expressive capacity. Temporal models yield slightly lower but comparable performance, with the two-layer LSTM approaching 89%, and CNN–LSTM, ResNet1D, and TimesNet ranging between 80% and 85%. Overall, no architectural family overwhelmingly dominates under clean conditions, as different modeling choices lead to similar generalization levels.

These results confirm that VR motion signals contain sufficient behavioral information to enable relatively accurate gender verification. However, none of the evaluated models achieves near-perfect reliability, motivating further analysis under structured behavioral perturbations.

4. **Robustness Analysis**

4.1. **Methodology**

The central question of this work is whether a VR user can voluntarily influence a gender verification system using only behavioral motion signals. We therefore evaluate model robustness under structured behavioral transformations constrained to reflect physically executable movement adjustments. Unlike white-box adversarial attacks that rely on gradient access and imperceptible numerical perturbations, we consider model-agnostic transformations applied to interpretable movement features (e.g., posture, dynamics, timing, and coordination), approximating voluntary adaptations that can be enacted without knowledge of model internals.

We evaluate model robustness under four realistic behavioral scenarios that reflect plausible modifications a VR user may enact during interaction. All perturbations are applied exclusively to the held-out test set. After modification, the full preprocessing pipeline (`StandardScaler` $\rightarrow$ `VarianceThreshold` $\rightarrow$ `SelectKBest`) is re-applied using parameters learned from the training data, ensuring a leakage-free and deployment-consistent evaluation.

Robustness is quantified using balanced accuracy on clean data ($BAcc_{clean}$), balanced accuracy under perturbation ($BAcc_{adv}$), performance drop ($\Delta BAcc$), flip rate (proportion of predictions that change), and mean prediction confidence.

The four following scenarios are used to group the considered behavioral perturbations. Each scenario targets a distinct aspect of VR motor behavior: postural alignment, movement dynamics, gaze strategy, and multimodal coordination, reflecting the range of voluntary adjustments a user might realistically attempt during an immersive session. This decomposition allows us to isolate which behavioral dimensions contribute most to classifier instability, and to assess whether models are robust across all modalities or selectively vulnerable to specific types of manipulation.

4.1.1. **Scenario 1: Voluntary Behavioral Neutralization (Self-Masking) :**

This scenario models deliberate reduction of motion expressivity. It is instantiated by the following perturbations :

- `total_neutralization` contracts posture, dynamic (`_std`, `_range`, `_energy`), and gaze descriptors by a multiplicative factor in $[0.5, 0.6]$.

- `global behavioral rescaling` multiplies dynamic descriptors by approximately 0.7.
- `progressive fatigue slowdown` multiplies dynamic features by 0.55 and subtracts up to 0.20 from vertical posture descriptors.

4.1.2. **Scenario 2: Motion Stylization or Imitation :**

This scenario captures deliberate alteration of behavioral style. It is instantiated by:

- `inverted_style`, which scales dynamic descriptors (`_std`, `_range`, `_energy`) by either 0.5 or 1.8.
- `hybrid_morpho`, which applies a posture shift of -0.30 to `hmd_pos_y_mean`, multiplies dynamic descriptors by 0.80, and multiplies gaze descriptors by -0.8.
- `nn_counterfactual`, which linearly interpolates each feature vector with a nearest-neighbor behavioral profile using $\alpha = 0.5$.
- `counterfactual style imitation`, which shifts feature vectors toward an alternative behavioral centroid using directional interpolation with $\alpha \approx 0.35$.

  These attacks modify posture, morphology proxies, gesture amplitude, and gaze in a coordinated manner to approximate style convergence.

4.1.3. **Scenario 3: Multimodal Desynchronization :**

This scenario disrupts coordination between hands, posture, and gaze. It is implemented by:

- `multimodal desynchronization`, which scales left-hand dynamic features by 1.30 and right-hand features by 0.70.
- `posture-gaze mismatch`, which applies an additive posture shift of approximately 0.40 while inverting gaze-direction features.
- `sustained temporal drift`, which progressively scales dynamic descriptors toward a factor of 1.05.
- `localized jitter bursts`, which inject Gaussian noise with standard deviation $\sigma = 0.03$ into dynamic features.

  These transformations weaken cross-modal coherence while remaining bounded and realistic.

4.1.4. **Scenario 4: Room Setup and Spatial Configuration Changes :**

This scenario models environmental and calibration effects. It corresponds to:

- `global_scaling`, which multiplies all numerical features by a factor sampled uniformly in $[0.8, 1.2]$.
- `unit_mismatch_pos`, which multiplies positional features matching `_pos_` by 1000, simulating unit inconsistency.

  These perturbations systematically alter spatial magnitudes without modifying semantic feature grouping.
- `fixed_gaze` forces gaze-direction descriptors toward a floor value of 0.05, simulating sustained fixation.
- `expressive_gaze` multiplies gaze-direction features by a factor sampled in $[-2.0, -1.2]$, modeling exaggerated visual exploration.

4.2. **Results on Deep and Classical Tabular Models**

Each perturbation is applied to the original feature set before preprocessing, and the full pipeline (`StandardScaler` $\rightarrow$ `VarianceThreshold` $\rightarrow$ `SelectKBest`) is re-executed prior

to inference. Robustness is measured using balanced accuracy on clean data, balanced accuracy under attack, performance drop (ΔBAcc = $\text{BAcc}_{clean} - \text{BAcc}_{adv}$), flip rate, and mean confidence.

Table 3 reports the most damaging behavioral perturbations, showing for each attack the worst-case degradation across tabular models. Performance degradation is primarily driven by global and structured modifications rather than small localized noise.

The strongest degradation is caused by `unit_mismatch_pos`. This attack multiplies all position-related features (e.g., `hmd_pos_*`, controller positions) by a constant factor (e.g., $\times$1000), simulating a unit mismatch (meters vs. millimeters) or a severe calibration error after sensor reconfiguration. Such systematic spatial distortion drives the worst-case model down to $\text{BAcc}_{adv} = 0.50$ (ΔBAcc = 0.43, Flip = 0.46), revealing strong sensitivity to global coordinate scaling.

Attenuation-based behavior modeled by `total_neutralization` uniformly contracts posture, dynamic, and gaze descriptors (by factors around 0.5–0.6 while preserving sign). This corresponds to a user deliberately minimizing gesture amplitude, reducing expressivity, and limiting movement variability (e.g., adopting restrained, compact motions). Such global behavioral suppression yields large drops (ΔBAcc = 0.39, Flip = 0.50), indicating that discriminative cues are strongly encoded in amplitude and variability statistics.

The `hybrid_morpho` attack introduces coordinated modifications: a posture shift (e.g., lowering head-height features), a multiplicative reduction of dynamic features, and inversion or amplification of gaze direction. This approximates a user simultaneously changing stance, gesture style, and visual strategy. The resulting cross-modal restructuring produces substantial instability (ΔBAcc = 0.38, Flip = 0.41), suggesting reliance on coherent multimodal patterns.

Gaze-focused perturbations further expose vulnerability. `fixed_gaze` constrains gaze variability to a small value, modeling sustained visual fixation (e.g., staring straight ahead), whereas `expressive_gaze` exaggerates or inverts gaze direction through multiplicative scaling. Both yield ΔBAcc = 0.36 with Flip = 0.386, confirming that eye-movement statistics significantly influence decisions. The `nn_counterfactual` attack models deliberate behavioral imitation. In this scenario, a user attempts to resemble the movement style typically associated with the opposite class by progressively adjusting posture height, gesture amplitude, movement rhythm, and head–hand coordination patterns. Rather than introducing artificial noise or isolated feature corruption, this perturbation represents a coherent adaptation of global movement style across multiple modalities. The resulting shift in behavioral signature increases inter-class overlap, yielding ΔBAcc = 0.27 and Flip = 0.32. These results suggest that even partial and structured imitation of an alternative behavioral profile can meaningfully destabilize classification outcomes.

Overall, two primary failure modes emerge: (i) extreme sensitivity to global spatial scaling and calibration shifts, and (ii) vulnerability to coordinated behavioral restructuring affecting posture, movement amplitude, and gaze strategy. These structured modifications produce substantial accuracy degradation despite strong nominal performance.

4.3. **Results on Temporal Models**

Temporal perturbations are applied at the sequence level within fixed-length windows, targeting sustained temporal drift, inter-limb delays, cross-modal incoherence, and global behavioral rescaling. Unlike tabular models, temporal architectures explicitly encode long-range dependencies between posture, hand dynamics, and gaze signals. Robustness is quantified using balanced accuracy, accuracy drop (ΔBAcc), and prediction flip rate, following Section 4.2.

Table 3. Worst-case robustness across tabular models under behavioral perturbations (for each attack, values are reported for the model with the largest ΔBAcc).

Attack	BAcc_{clean}	BAcc_{adv}	ΔBAcc	**Flip**
Inverted style (`inverted_style`)	0.93	0.75	0.18	0.27
Global scaling (`global_scaling`)	0.91	0.77	0.14	0.23
Unit mismatch (pos) (`unit_mismatch_pos`)	0.93	0.50	0.43	0.46
Total neutralization (`total_neutralization`)	0.91	0.52	0.39	0.50
Hybrid morpho (`hybrid_morpho`)	0.93	0.55	0.38	0.41
Fixed gaze (`fixed_gaze`)	0.93	0.57	0.36	0.39
Expressive gaze (`expressive_gaze`)	0.93	0.57	0.36	0.39
NN counterfactual (`nn_counterfactual`)	0.93	0.66	0.27	0.32

Table 4. Worst-case degradation of temporal models under structured sequence-level perturbations. For each attack, values correspond to the most affected architecture.

Attack	BAcc_{clean}	BAcc_{adv}	ΔBAcc	**Flip**
Global behavioral rescaling	0.89	0.58	0.31	0.47
Multimodal desynchronization	0.88	0.55	0.33	0.49
Posture–gaze mismatch	0.87	0.57	0.30	0.46
Sustained temporal drift	0.86	0.60	0.26	0.42
Progressive fatigue slowdown	0.85	0.63	0.22	0.38
Localized jitter bursts	0.84	0.69	0.15	0.24
Counterfactual style imitation	0.89	0.59	0.30	0.44

Under nominal conditions, temporal models achieve balanced accuracies between 0.84 and 0.89 across architectures (LSTM, CNN–LSTM, ResNet1D, and TimesNet). However, robustness significantly degrades under structured sequence-level perturbations.

Table 4 highlights that the most destructive perturbations are those that disrupt temporal coherence rather than introducing short-lived noise. Multimodal desynchronization produces the largest degradation, reducing balanced accuracy from 0.88 to 0.55 ($\Delta\text{BAcc} = 0.33$, $\text{Flip} = 0.49$). Similarly, global behavioral rescaling yields a 31% drop, confirming strong sensitivity to systematic amplitude and timing distortions.

Posture–gaze mismatch and counterfactual style imitation also induce substantial degradation (approximately 30% drop), indicating that coherent but altered behavioral strategies significantly increase class overlap. Sustained temporal drift and progressive fatigue cause moderate yet consistent degradation (22%–26%), suggesting that gradual deviations accumulate within recurrent hidden states and amplify representation drift over time.

In contrast, localized jitter bursts lead to comparatively limited degradation ($\Delta\text{BAcc} = 0.15$, $\text{Flip} = 0.24$), demonstrating that short-lived perturbations are less harmful than structured, sequence-wide coherence-breaking transformations.

Overall, despite strong nominal performance (balanced accuracy up to 0.89), temporal architectures exhibit structural vulnerability to coordinated behavioral manipulations. When long-range temporal correlations between modalities are altered, performance approaches near-random levels (0.55–0.60 balanced accuracy), revealing an inherent fragility of temporally dependent models.

5. **Discussion and Conclusion**

This study examined whether a user in virtual reality can intentionally alter behavioral patterns to influence a gender classification system. Across tabular, deep feedforward, and temporal models, robustness was consistently limited under structured and voluntary behavioral perturbations.

Tabular ensembles achieved the highest nominal performance (up to 93% balanced accuracy) and remained stable under mild variability ($\leq 3\%$ drop), but degraded substantially under coherent global changes such as posture neutralization or multimodal desynchronization (drops $>35\%$, flip rates $\approx 50\%$). The PyTorch MLP, despite strong clean performance ($\approx 91\%$), exhibited the largest sensitivity to large-scale behavioral shifts (drops $\approx 50\%$). Temporal architectures showed intermediate robustness: they tolerated localized perturbations but degraded under sustained multimodal distortions, in some cases approaching chance-level performance.

Performance degradation was primarily driven by perturbations that altered global statistical structure or multimodal coordination rather than isolated features. These results indicate that a motivated user can plausibly adjust posture, gesture amplitude, gaze strategy, or synchronization patterns in ways that significantly increase misclassification probability.

Although architectural families differ in baseline accuracy and degradation magnitude, none demonstrated inherent resistance to structured behavioral manipulation. Improving reliability in VR-based gender verification therefore requires robustness-oriented strategies beyond nominal accuracy optimization. Experimental validation with real users remains necessary to assess whether such behavioral adaptations can be intentionally and consistently reproduced in immersive settings.

Reproducibility. Our experiments are performed on publicly available data. All perturbations are applied to predefined feature groups identified through deterministic column-name rules (e.g., head-, hand-, and gaze-related features). When perturbations involve random scaling factors (e.g., sampled uniformly in $[0.8, 1.2]$), fixed random seeds are used to ensure exact reproducibility across runs. After applying the transformations, features are optionally clipped to the minimum and maximum values observed in the training set. This prevents unrealistic numerical artifacts while preserving the statistical support of the original data distribution.

References

[1] Q. J. Wang and R. P. McMahan. "Gender identification of vr users by machine learning tracking data". In: *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE. 2024, pp. 827–828.

[2] M. R. Miller, E. Han, C. DeVeaux, E. Jones, R. Chen, and J. N. Bailenson. "A Large-Scale Study of Personal Identifiability of Virtual Reality Motion Over Time". In: 2023. arXiv: 2303.01430 [cs.CR].

[3] Z. C. Jianmin Dong Youtian Du et al. "Gender recognition using motion data from multiple smart devices". In: *scienceDirect* (2020).

[4] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. "Towards Deep Learning Models Resistant to Adversarial Attacks". In: (2019). arXiv: 1706.06083 [stat.ML].

[5] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. "DeepFool: a simple and accurate method to fool deep neural networks". In: 2016. arXiv: 1511.04599 [cs.LG].

[6] C. Rack, T. Fernando, M. Yalcin, A. Hotho, and M. E. Latoschik. "Who is Alyx? A new behavioral biometric dataset for user identification in XR". In: *Frontiers in Virtual Reality* (2023). Dataset repository: https://github.com/cschell/who-is-alyx.

[7] . *The Full-scale Assembly Simulation Testbed (FAST) Dataset*. arXiv. Dataset repository: https://github.com/xrtlab/FAST-Dataset. 2024.

[8] V. Nair, C. Rack, W. Guo, R. Wang, S. Li, B. Huang, A. Cull, J. F. O'Brien, M. Latoschik, L. Rosenberg, and D. Song. "Inferring Private Personal Attributes of Virtual Reality Users from Head and Hand Motion Data". In: (2023). arXiv: 2305.19198 [cs.HC].

[9] Q. J. Wang, A. G. Moore, N. N. Chawla, and R. P. McMahan. "Cross-Domain Gender Identification Using VR Tracking Data". In: *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. Oct. 2024.

[10] . "VR Eye Tracking Data for Gender Identification: A Look at Same-Domain and Cross-Domain Scenarios". In: *ACM (Digital Library entry)*. Dec. 2025.

[11] N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma. "Adversarial classification". In: New York, NY, USA: Association for Computing Machinery, 2004. ISBN: 1581138881.

[12] S. Pelekis, T. Koutroubas, A. Blika, et al. "Adversarial machine learning: a review of methods, tools, and critical industry sectors." In: Springer Nature Link, 2005, pp. 641–647.

[13] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. "Practical Black-Box Attacks against Machine Learning". In: 2017. arXiv: 1602.02697 [cs.CR].

[14] A. Kantchelian, J. D. Tygar, and A. D. Joseph. "Evasion and Hardening of Tree Ensemble Classifiers". In: (2016). arXiv: 1509.07892 [cs.LG].

[15] S. M. Fazle Karim et al. "Adversarial Attacks on Time Series". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[16] T. S. Yunfeng Diao1 et al. "BASAR: Black-box Adversarial Attack on Skeleton-based Action Recognition". In: *computer vision foundation* (2021).

[17] O. Y. M. M. Kurniawan A. "Experiments on Adversarial Examples for Deep Learning Model Using Multimodal Sensors". In: *Sensors* 22.22 (2022), p. 8642.

[18] T. L. Giang Bui Brittany Morago et al. "Integrating videos with LIDAR scans for virtual reality". In: *Proceedings of the IEEE Conference on Virtual Reality (VR)*. 2016.

[19] M. R. A. L. J. H. Tanner Hobson Jeremiah Duncan et al. "Alpaca: AR Graphics Extensions for Web Applications". In: *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 2020.

[20] N. P. Simon Eberz Andrea Patané et al. "Broken Hearted: How to Attack ECG Biometrics". In: *Proceedings of the Network and Distributed System Security Symposium (NDSS)*. 2017.

[21] N. N. C. A. J. I. Alec G. Moore Tiffany D. Do and R. P. McMahan. *FAST Dataset: Full Body Tracking Data from VR Users*. 2023.

[22] C. Rack, F. Sieper, L. Schach, M. Yalcin, and M. E. Latoschik. *Who is Alyx? (GitHub Repository)*. GitHub Repository. 2022.

[23] N. R. R. P. M. Alec G. Moore Tiffany D. Do. "Identifying Virtual Reality Users Across Domain-Specific Tasks". In: *IEEE* (2023).

[24] X. Wang, B. Lafreniere, and J. Zhao. "Exploring Visualizations for Precisely Guiding Bare Hand Gestures in Virtual Reality". In: New York, NY, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300.

[25] J. T. Wei Huo Ke Wang et al. "Gait Recognition via Motion Difference Representation Learning and Salient Feature Modeling". In: *IEEE Transactions on Human-Machine Systems* (2025).

[26] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: 12.null (2011).

[27] I. Guyon and A. Elisseeff. "An introduction to variable and feature selection". In: 3 (2003).

[28] S. R. Varma S. "Bias in error estimation when using cross-validation for model selection". In: *BMC Bioinformatics* 7.1 (2006), p. 91.

[29] G. C. Cawley and N. L. Talbot. "On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation". In: 11 (2010). ISSN: 1532-4435.

[30] Y. Z. Fuyu Zhu Hua Wang. "Deep residual networks for time series classification". In: *IEEE 6th Information Technology,Networking,Electronic and Automation Control Conference (ITNEC)* (2023).

[31] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long. "TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis". In: (2023). arXiv: 2210.02186 [cs.LG].

[32] M. Awais, F. Shamshad, and S.-H. Bae. *Towards an Adversarially Robust Normalization Approach*. 2020. arXiv: 2006.11007 [cs.LG].

[33] T. Kim, J. Kim, Y. Tae, C. Park, J.-H. Choi, and J. Choo. "Reversible Instance Normalization for Accurate Time-Series Forecasting against Distribution Shift". In: *International Conference on Learning Representations*. 2022.

[34] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.