

Article

Condition-Wise Robustness of Skeleton-Based Gait Sex Classification Under Smartphone Use, Occlusion, and Speed Variations

A Hyun Jung ¹, Yujin Oh ¹, Ye Eun Kong ¹, Min-Hyung Choi ² and Se Dong Min ^{1,3,*}

¹ Department of Software Convergence, Soonchunhyang University, Asan 31538, Republic of Korea; ahjung333@gmail.com (A.H.J.); yujinhappy@naver.com (Y.O.); yeeungong1014@gmail.com (Y.E.K.)

² Department of Computer Science, Saint Louis University, Louis, MO 63103, USA

³ Department of Medical IT Engineering, Soonchunhyang University, Asan 31538, Republic of Korea

* Correspondence: sedongmin@sch.ac.kr

Featured Application

Privacy-preserving sex classification for human–computer interaction and risk-aware assistance from RGB video-derived skeletons in everyday walking scenarios.

Abstract

Skeleton-based gait sex classification can reduce reliance on appearance cues, yet its robustness under everyday walking disturbances remains under-quantified. Using PsyMo 2D pose sequences (90° side view), we render Common Objects in Context (COCO) keypoints into compact grayscale skeleton images, segment sequences into fixed-length 15-frame clips, and classify them with a 3D residual convolutional neural network (CNN) under a subject-wise split shared across four aggregated conditions: overall (A), occlusion/carrying disturbance (B), speed variation (C), and smartphone use (D). To avoid an arbitrary decision rule, we select a global operating threshold on the validation set by sweeping τ to maximize macro-F1, apply it unchanged to the held-out test set, and report a threshold-sensitivity check. Robustness is audited via condition-wise confusion matrices, subgroup precision/recall with 95% subject-level bootstrap confidence intervals, and subject-level probability overlap. To contextualize condition-dependent behavior, we quantify joint-group attribution shifts using Gradient-weighted Class Activation Mapping (Grad-CAM) and examine a coarse arm-swing proxy under smartphone use. Subject-level test accuracy ranged from 0.761 to 0.870 across conditions A–D, with uncertainty summarized by 95% subject-level bootstrap confidence intervals; performance was lowest in B, with increased male→female errors. Overall, these results provide a transparent audit-and-interpretation framework for assessing skeleton-based gait sex classification under realistic walking perturbations in practical evaluation scenarios.



Academic Editor: Lingfeng Shi

Received: 20 January 2026

Revised: 6 February 2026

Accepted: 11 February 2026

Published: 12 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: gait; 2D skeleton; sex classification; robustness; interpretability

1. Introduction

Gait is one of the most frequently repeated whole-body movements in daily life and involves inter-segmental coordination and periodic motor patterns. Owing to these characteristics, gait has long been used as a biometric signal for personal identification [1], and recent research has expanded its scope beyond identity to infer attributes such as age [2], disease status [3], and emotion [4].

Among these attributes, sex has been considered useful either as prior information in medical settings and security and surveillance environments [5] or as a means to rapidly capture user context in human–computer interaction [6] and personalized services [7]. Accordingly, vision-based sex classification has been extensively studied; however, many approaches have relied on appearance cues such as the face, hair, and clothing [8], making them vulnerable to occlusion, changes in apparel, and variations in camera distance and viewpoint. In addition, face-based analysis can impose privacy and ethical burdens [9]. To avoid conflating biological sex labels with sex identity, we use “sex” strictly to denote the binary label provided by the dataset; this work does not claim to infer gender identity, and we explicitly discuss limitations and ethical considerations of binary labeling in Section 4.6.

To address these limitations, gait-based sex classification that leverages motion cues rather than appearance cues has been repeatedly proposed. Studies exploiting the dynamical cues of gait have reported the feasibility of sex classification in controlled settings [10,11], and joint-based skeleton representations are useful because they summarize movement structure without directly revealing facial or clothing details. For example, studies using sensor-derived joint information from devices such as Kinect have reported high classification performance under controlled conditions [12,13]. However, such approaches face practical constraints in everyday deployment due to hardware dependency and installation requirements.

Moreover, a substantial portion of prior work has focused on reporting performance in laboratory or semi-controlled environments; comparatively fewer studies have systematically examined how performance shifts under everyday perturbations and whether such shifts are attributable to weakened motion cues. In particular, quantitative evidence remains limited that (i) measures condition-wise performance degradation under realistic disturbances, (ii) identifies whether misclassification increases asymmetrically across subgroups (e.g., male vs. female), and (iii) links performance changes to weakened or redistributed motion cues rather than unintended appearance leakage [14–16].

Beyond reporting a single aggregate score, an important missing piece is a reproducible and transparent audit framework that quantifies performance degradation under realistic daily-life perturbations, decomposes subgroup-specific failure modes (e.g., male vs. female), and reports statistical uncertainty under subject-wise evaluation splits. Accordingly, we position our contribution as a condition-wise robustness audit and failure-mode characterization for skeleton-based gait sex classification in naturalistic settings, rather than a claim of state-of-the-art superiority across heterogeneous pipelines.

Motivated by these gaps, we include smartphone use as a representative daily-life disturbance in our condition-wise audit. Smartphone texting is a prototypical dual-task that simultaneously imposes cognitive, visual, and upper-limb demands and has been reported to alter gait speed, spatiotemporal parameters, and postural control strategies [17,18]. Because smartphone use inherently mixes cognitive load and upper-limb constraints, we treat it as a combined disturbance and refrain from attributing observed performance changes to a single causal factor without additional controlled experiments.

These issues have been difficult to address, largely because large-scale acquisition of joint information in everyday environments and validation across diverse conditions have been constrained. Recently, however, pose estimation methods that can stably estimate 2D joints in multi-person scenes using RGB videos alone have become widespread [19], and datasets that include representations such as skeletons have been released [20], providing a basis for evaluation and interpretation under naturalistic conditions without additional sensors. At the same time, recent skeleton-based recognition studies have increasingly explored robustness-oriented strategies—such as realistic data augmentation and occlusion-aware modeling—to mitigate nuisance factors such as noisy pose estimates and partial

occlusion [21–23]. Nevertheless, in gait-based sex classification, systematic condition-wise and subgroup-wise quantitative audits of performance and failure modes remain scarce.

Motivated by this gap, we evaluate skeleton-based gait sex classification under everyday disturbances using 2D pose sequences. We render COCO keypoints from the PsyMo dataset into grayscale skeleton images and train a 3D residual CNN to capture spatiotemporal motion patterns for classification. We quantify robustness and failure modes using condition-wise confusion matrices and subgroup-wise precision/recall, and we report statistical uncertainty using subject-level splitting and bootstrap confidence intervals. Finally, we interpret performance degradation by quantifying joint-group importance shifts using Grad-CAM and by examining upper-limb motion changes under smartphone-use conditions using simple kinematic indices. To enable fair and reproducible comparisons, we conduct within-dataset, subject-wise controlled audits across perturbation conditions.

The main contributions of this study are as follows:

- We systematically compare skeleton-based sex-classification performance across four walking conditions—normal walking (baseline), smartphone use, partial occlusion, and speed variation (slow/fast)—and quantitatively assess applicability under everyday conditions.
- We decompose errors using condition-wise confusion matrices and sex-group-wise precision and recall, quantifying under which conditions misclassification increases and in which subgroup.
- Under the smartphone-use condition, we compute changes in upper-limb motion using simple kinematic indices and examine whether these changes relate to prediction confidence and increased misclassification, while acknowledging that cognitive load and upper-limb constraint are confounded in naturalistic dual-task walking.
- Using Grad-CAM-based saliency analysis, we quantify importance at the joint level and across gait segments and show how condition-dependent performance variations relate to weakened motion cues.

Overall, the study provides evidence and a reporting framework for how skeleton-only gait sex classification behaves under everyday perturbations, emphasizing transparent evaluation and failure-mode characterization. The remainder of this paper is organized as follows: Section 2 describes the dataset, preprocessing, model architecture, and evaluation protocol. Section 3 presents condition-wise performance results and error decomposition. Section 4 discusses interpretation findings, limitations, and future directions. Section 5 concludes the paper.

2. Materials and Methods

2.1. Dataset and Experimental Conditions

This study used the PsyMo dataset, which was constructed for analyzing psychological characteristics and gait [20]. PsyMo contains seven walking conditions (NM, CL, BG, WSS, WSF, TXT, PH) recorded from six camera viewpoints and provides metadata including sex, along with silhouette-, skeleton-, and SMPL-based representations for gait data from a total of 312 subjects (199 male and 113 female). In this study, we used the provided 2D skeleton representation and the sex labels, and we restricted the analysis to the 90° side-view data to reduce the influence of appearance variations across viewpoints. The 2D skeleton is provided as keypoints for 17 COCO-format joints (Figure 1); from each frame's keypoints vector, we extracted only the x and y coordinates to construct subsequent inputs. The PsyMo data may include multiple recording units (takes) per subject for each condition. We generated clips from all valid takes corresponding to the target conditions (A–D) while performing the training/evaluation split at the subject level (subject-wise) so that takes or clips from the same subject did not simultaneously appear in both the training and test

sets. Accordingly, even when multiple takes existed per subject, we compared condition-wise performance under a setting that mitigates information leakage and performance overestimation due to repeated recordings.

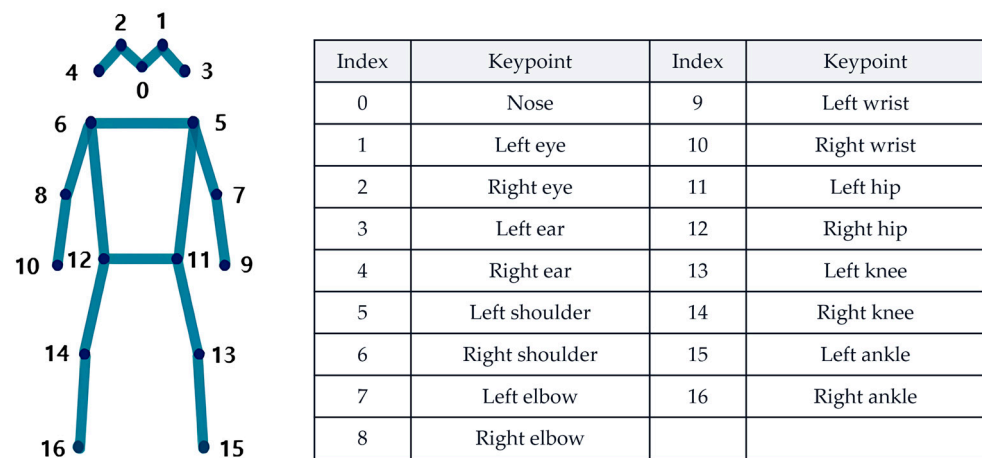


Figure 1. COCO pose format.

In implementation, for each subject and walking type, we used take indices 0 and 1 when available. To keep a consistent temporal budget per take across conditions, we used only the first $T = 45$ usable frames per take and excluded takes shorter than T frames (Section 2.2.3). To facilitate assessment of data balance and condition-wise sample availability, Table 1 summarizes the sample-size distribution in the 90° side-view subset, reporting the number of subjects (male/female) and the number of usable non-overlapping fixed-length clips ($L = 15$) obtained after preprocessing. Because clips are fixed-length and non-overlapping, the number of usable frames is proportional to the clip counts (frames = $15 \times$ clips).

Table 1. Seven walking conditions and sample availability (90° side-view subset only).

Abbreviation	Meaning	Subjects (M/F)	Clips ($L = 15$) ¹
NM	Normal walking	226 (113/113)	1329
CL	Changing clothing	226 (113/113)	1341
BG	Carry bag	226 (113/113)	1344
WSS	Walking speed slow	225 (113/112)	1341
WSF	Walking speed fast	226 (113/113)	1335
TXT	Walking while texting	226 (113/113)	1332
PH	Talking on the phone	226 (113/113)	1323

¹ Clips were generated after preprocessing by truncating each take to the first $T = 45$ frames and splitting it into sequential non-overlapping clips of length $L = 15$ (stride = L); therefore, clip counts indicate usable sample availability per condition, and the equivalent number of usable frames can be derived as $15 \times$ clips.

To compare sex-classification accuracy across walking conditions, we reorganized the seven PsyMo walking conditions (NM, CL, BG, WSS, WSF, TXT, PH) into four aggregated settings based on similar factors, as shown in Table 2. The overall condition (A) included NM, CL, BG, WSS, WSF, TXT, and PH; the partial-occlusion condition (B) included NM, CL, and BG; the speed-variation condition (C) included NM, WSS, and WSF; and the smartphone-use condition (D) included NM, TXT, and PH. Because conditions B–D are defined as subsets of the overall dataset (A), clip counts across conditions are not additive. To ensure fair condition-wise comparisons without confounding by subject composition, we used an identical subject-wise train/test split across conditions A–D (Table 2). For reference, we additionally report the normal-only subset (E), which is derived by restricting

condition A to the NM trials under the same split. Table 2 includes the clip counts for A–D (and E) under $L=15$ and clarifies the shared subject split used throughout the study.

Table 2. Datasets for the aggregated conditions (A–D) and the normal-only subset (E).

Condition ²	Included Data	15-Frame Clips (Train/Test) ¹
Overall dataset (A)	NM, CL, BG, WSS, WSF, TXT, PH	9345 (7440/1905)
Partial-occlusion dataset (B)	NM, CL, BG	4014 (3195/819)
Speed-variation dataset (C)	NM, WSS, WSF	4005 (3192/813)
Smartphone-use dataset (D)	NM, TXT, PH	3984 (3177/807)
Normal-only dataset (E)	NM	1329 (1062/267)

¹ All conditions share the same held-out test-subject set ($n = 46$; 23 male and 23 female); only the number of clips differs by condition. ² Conditions B–D are subsets of A; therefore, clip counts across conditions are not additive.

2.2. Dataset Preprocessing and Input Construction

2.2.1. Skeleton Rendering and Normalization

To standardize scale and translation per frame in a manner consistent with our implementation, we rendered each skeleton using bounding-box-based scaling and center alignment on a fixed 50×50 canvas. Specifically, for each frame, we first computed the bounding box over the 17 joints (minimum/maximum x and y). We then scaled the skeleton uniformly (preserving aspect ratio) so that the bounding box fits within the canvas while leaving a fixed margin and subsequently applied offsets to place the scaled skeleton at the center of the canvas. Finally, we drew joints and inter-joint connections in black on a white background and converted the rendered image to a single-channel grayscale skeleton image (Figure 2). A small constant was used internally to prevent degenerate scaling when the bounding box width/height becomes extremely small.

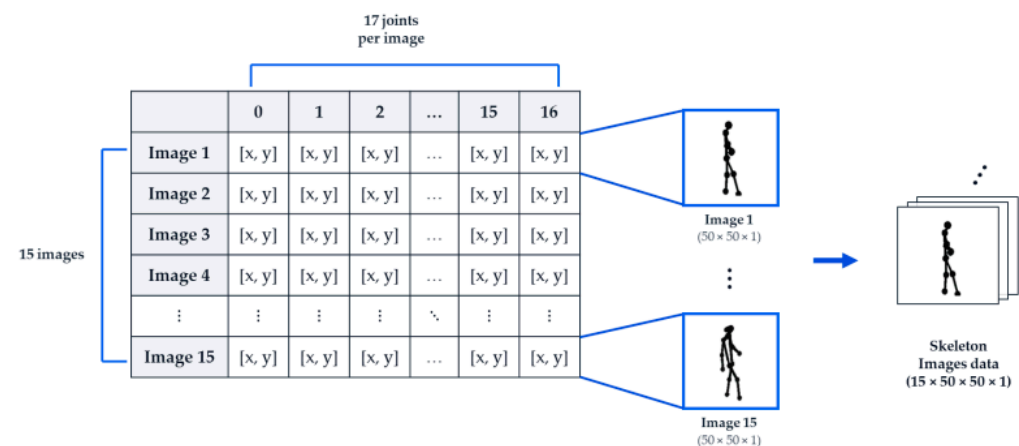


Figure 2. Preprocessing pipeline for constructing model input sequences from 2D joint coordinates. Each frame is rendered by bounding-box-based scaling and center alignment into a $50 \times 50 \times 1$ grayscale skeleton image. A $T = 45$ -frame take is then split into sequential non-overlapping clips of length L (main setting $L = 15$), yielding a final input tensor shape of $(15, 50, 50, 1)$. The stacked icon indicates a sequence of multiple skeleton frames (i.e., an L -frame clip), not a single image.

2.2.2. Rationale for Fixed-Length Clip Sampling

Although gait-cycle-based segmentation and phase alignment can facilitate interpretation, they require prior detection of gait events such as heel strike. Because the input was limited to 2D skeleton-rendered images, we judged the reliability of gait-event detection to be insufficient and therefore constructed inputs as non-overlapping fixed-length clips rather than detecting and segmenting gait cycles. We acknowledge that fixed-length clips

without event-based phase alignment can introduce phase mismatch, particularly under speed-variation conditions. Therefore, we later report a proxy robustness diagnostic based on within-subject prediction instability to quantify how unstable predictions become under such conditions (Section 2.4.1).

2.2.3. Clip Construction and Input Tensor

For each take, we first converted up to the first $T = 45$ frames into 50×50 grayscale skeleton images (Section 2.2.1). Takes with fewer than T usable frames were excluded. We then constructed fixed-length clips by splitting the T -frame sequence into sequential non-overlapping segments of length L with stride L , using start indices $0, L, 2L, \dots$. Any remainder shorter than L at the end of the T -frame window was discarded. Thus, the number of clips per take is $\lfloor T/L \rfloor$. In the main experiments, we used $L = 15$, yielding three non-overlapping clips per take corresponding to frames (1–15), (16–30), and (31–45). The final input tensor had shape (15, 50, 50, 1), and the rendered images were scaled to the range [0, 1] during training by dividing pixel values by 255.

We additionally conducted a clip-length sensitivity (ablation) study with $L \in \{10, 15, 20, 30\}$ under the same subject-wise split and training protocol; the outcomes are reported in Section 3.5 (and Supplementary Materials).

2.2.4. Label Encoding and Subject-Wise Split

We encoded sex labels as 0 for male and 1 for female for training and inference. To mitigate sex imbalance, we randomly removed 86 male subjects (seed = 42), resulting in a balanced cohort of 226 subjects (113 male and 113 female). We then applied a subject-wise stratified split (80:20) to create training and test sets comprising 180 and 46 subjects, respectively (test set: 23 male and 23 female). The per-condition clip counts are summarized in Table 2. The four aggregated conditions (A–D) shared the same subject split; therefore, differences in sample size across conditions reflect the included walking types and the resulting number of valid clips rather than differences in subject composition. The train/test subject IDs and clip counts are provided in Supplementary Table S1.

2.3. Sex Classification Model and Baseline Evaluation

2.3.1. Model Architecture

To classify sex from 15-frame input clips ($15 \times 50 \times 50 \times 1$) composed of 2D skeleton-rendered images, we used a 3D convolutional residual neural network [24,25]. In the initial stage, the network extracts low-level spatiotemporal features using a $7 \times 7 \times 7$ Conv3D (stride 2, same padding, no bias)–Batch Normalization (BN)–ReLU block and a $3 \times 3 \times 3$ MaxPool3D (stride 2, same padding). It then comprises four stages in which the number of channels increases to 64, 128, 256, and 512, respectively, and each stage contains two residual blocks ($\times 2$) to learn hierarchical representations. In each residual block, the main path consists of Conv3D–BN–ReLU–Conv3D–BN, where both Conv3D layers use $3 \times 3 \times 3$ kernels with the same padding. When a stride change or a mismatch between input and output channels occurs, the skip path applies a $1 \times 1 \times 1$ Conv3D (with the same stride, same padding, no bias) followed by BN to match dimensions; the block then combines the two paths via element-wise addition, followed by ReLU.

For reproducibility, we provide the layer-wise configuration of the 3D residual CNN (number of filters per stage, kernel sizes, strides, padding strategy, and output tensor shapes) and the residual-block specification in a supplementary architecture table (Supplementary Table S4). In addition, we report the total number of trainable parameters (33,291,585; ≈ 33.29 M) based on the implemented architecture.

At the final stage, the network aggregates features using Global Average Pooling 3D, applies dropout ($p = 0.5$), passes them through a fully connected layer (256 units, ReLU),

applies dropout again ($p = 0.5$), and produces a sigmoid output representing the probability of being female ($y = 1$). The overall architecture is shown in Figure 3. This design enables direct learning of gait spatiotemporal patterns, including the temporal dimension, via 3D convolutions, improves training stability through residual connections, and reduces overfitting risk under limited data by avoiding high-dimensional flattening (via Global Average Pooling) and using dropout regularization.

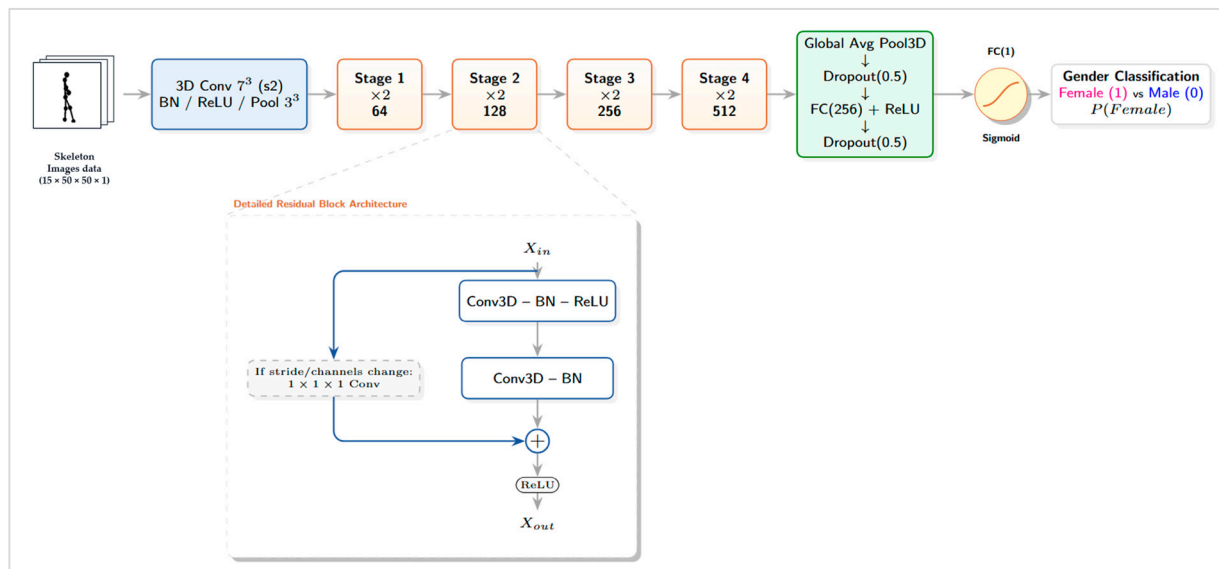


Figure 3. Overview of the sex-classification model architecture. Given a $15 \times 50 \times 50 \times 1$ skeleton clip, the network extracts initial features using a $7 \times 7 \times 7$ Conv3D and a $3 \times 3 \times 3$ MaxPool3D, learns spatiotemporal features through 3D residual blocks ($64 \rightarrow 128 \rightarrow 256 \rightarrow 512$), and outputs the sex (female = 1) probability via Global Average Pooling 3D and fully connected layers with a sigmoid output.

We chose ReLU as the hidden activation because it is a widely used non-saturating function that supports stable gradient-based optimization in deep residual networks, and we used a sigmoid output to model $p(y = 1)$ for binary classification with a probabilistic interpretation under the binary cross-entropy loss. This choice is consistent with the residual-block design, enabling efficient training while providing a probabilistic decision rule at inference.

2.3.2. Training Protocol and Evaluation Metric

We trained the model for 100 epochs with a batch size of 32 using binary cross-entropy loss and the Adam optimizer. We applied ReduceLROnPlateau based on the validation loss (factor = 0.2, patience = 20, minimum learning rate = 1×10^{-5}). During training, we used 20% of the training subjects as a validation set via a subject-wise stratified split (StratifiedShuffleSplit), ensuring that clips from the same subject did not appear in both training and validation sets. To avoid an arbitrary decision rule, we selected an operating threshold τ^* on the validation set by sweeping τ to maximize macro-F1 and then fixed τ^* for evaluation on the held-out test set. Unless otherwise specified, we report headline test results at the subject level: for each test subject, we aggregate clip-level predicted probabilities $p(y = 1)$ by averaging them within the subject and then binarize the subject-level mean probability using τ^* ($\hat{y} = 1$ if $p \geq \tau^*$, otherwise $\hat{y} = 0$). Clip-level results are provided in the Supplementary Material for completeness (Supplementary Figure S2).

We evaluated binary classification performance using accuracy, precision, recall, and F1-score, together with confusion matrices. We also report class-wise precision and recall

(male and female). Let TP, TN, FP, and FN denote the confusion-matrix entries with female (label 1) as the positive class and male (label 0) as the negative class. The metrics are defined as follows (Equations (1)–(4)):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

2.4. Robustness and Interpretation Analyses

2.4.1. Error Decomposition Protocol and Uncertainty Estimation

We computed confusion matrices on the test set for each condition (A–D) to characterize condition-wise misclassification patterns, and we reported accuracy, precision, recall, and F1-score for each condition. To compare sex-group-specific error bias, we additionally reported class-wise precision and recall for male (0) and female (1) under each condition. Defining female (1) as the positive class, we computed the false positive rate (FPR), i.e., the proportion of males misclassified as females, and the false negative rate (FNR), i.e., the proportion of females misclassified as males, defined in Equation (5):

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \quad \text{FNR} = \frac{\text{FN}}{\text{FN} + \text{TP}} \quad (5)$$

We assessed uncertainty in condition-wise performance differences using subject-level nonparametric bootstrap resampling (1000 resamples; 95% percentile confidence intervals) [26]. In each resample, we sampled test-set subjects with replacement and included all clips from the selected subjects when computing metrics, thereby estimating confidence intervals that reflect subject-level variability. We defined confidence intervals as the 2.5th–97.5th percentiles of the resampled distributions and reported 95% confidence intervals for accuracy and sex-group-specific precision/recall.

Proxy diagnostic for phase mismatch (prediction instability). To quantify within-subject prediction stability under each condition without gait-event alignment, we computed a subject-level prediction-instability score defined as the standard deviation of clip-level probabilities $p(y = 1)$ across all test clips from the same subject. We report condition-wise distributions of this score and examine its association with subject-level correctness as supplementary robustness evidence, interpreting it as a proxy diagnostic rather than a direct gait-event measure.

2.4.2. Kinematic Index Computation Under Smartphone Use

Motivated by prior studies reporting that smartphone texting alters spatiotemporal gait parameters and gait strategies [17,18], we computed a simple kinematic index representing arm swing from 2D skeleton coordinates under the smartphone-use condition (D) to evaluate how upper-limb constraints relate to classification performance and prediction confidence. Using bilateral wrist and shoulder joints, we computed, for each 15-frame clip, the left–right amplitude range (max–min) of the wrists in a shoulder-centered coordinate frame. We normalized this value by the mean shoulder width (distance between left and right shoulders) to account for differences in body size and scale, and we defined the clip-level arm-swing index as the mean of the left and right wrist amplitudes. We defined

prediction confidence such that it increases as the predicted probability moves farther from 0.5, and we evaluated the association between the arm-swing index and confidence using Spearman correlation. We further tested differences in the index between correctly and incorrectly classified samples using the Mann–Whitney U test. Note that the smartphone-use subset (TXT/PH) may conflate cognitive load and upper-limb constraints; accordingly, the arm-swing index is used as a coarse kinematic descriptor for interpretation rather than a definitive causal explanation.

2.4.3. Grad-CAM-Based Importance Quantification

To quantitatively relate condition-dependent performance variations to changes in motion cues, we applied a Grad-CAM-based interpretation method [27]. Grad-CAM produces class-discriminative saliency by weighting convolutional feature maps using gradients of the target output, thereby indicating which spatiotemporal regions contributed positively to a given class prediction. In our 3D CNN, we computed Grad-CAM with respect to the last convolutional block's feature maps to capture high-level spatiotemporal evidence.

For each 3D CNN classifier trained under conditions (A–D), we computed Grad-CAM using gradients with respect to the Conv3D feature maps for the model output and applied ReLU to retain only positive contributions. We then applied clip-wise min–max normalization to the resulting CAM and upsampled it to the input resolution (50×50) to obtain a spatiotemporal saliency map for each 15-frame clip.

To compute joint-based importance, we mapped each frame's joint coordinates to the 50×50 coordinate system using the same procedure as for the input skeleton images and accumulated CAM values within a radius r (default: 2 pixels) around each joint. We aggregated joint-wise importance into four joint groups (head, upper, trunk, lower; disjoint4) and normalized by the group sum so that the total group importance summed to 1. We summarized joint-group importance by averaging within each condition and additionally computed averages by walking type.

To examine long-range temporal trends in frame importance, we concatenated the frame-wise Grad-CAM sums (raw frame mass) from three consecutive 15-frame clips within the same sequence in chronological order, thereby constructing a 45-frame importance distribution. Because we did not perform event-based phase alignment, we partitioned the 45 frames into three segments (1–15, 16–30, 31–45), defined them as early/mid/late temporal segments, and compared segment-wise importance distributions.

3. Results

3.1. Baseline Sex Classification Performance Across Walking Conditions

Under a subject-wise split and a fixed operating threshold τ^* selected on the validation set ($\tau^* = 0.54$; Section 3.2.1), the 3D residual CNN achieved subject-level test accuracies of 0.8696, 0.7609, 0.7826, and 0.8261 for conditions A–D, respectively (Table 3). All conditions shared the same held-out test subjects (Total = 46; 23 male and 23 female; Table 2). Overall, the best performance was observed for the overall dataset (A), whereas the lowest accuracy among A–D was observed under partial occlusion/carrying (B). Table 3 also reports macro-averaged precision, recall, and F1-score computed under the same τ^* , ensuring consistency with the subject-level confusion matrices (Figure 4) and the error decomposition (Table 4). For reference, the normal-only subset (E; NM trials under the same held-out subject split) achieved an accuracy of 0.7609 (Tables 3 and 4).

Unless otherwise specified, all quantitative results in Tables 3 and 4 and the confusion matrices in Figure 4 are computed at the subject level by first aggregating clip-level predicted probabilities $p(\text{Female} = 1)$ per subject (mean over the subject's test clips) and then

binarizing the subject-level mean probability using the fixed operating threshold τ^* . Clip-level performance and clip-level confusion matrices are provided as supplementary results.

Table 3. Baseline performance across walking conditions (test set, subject-level).

Condition	Accuracy	Precision (Macro)	Recall (Macro)	F1 (Macro)
A	0.870	0.881	0.870	0.869
B	0.761	0.774	0.761	0.758
C	0.783	0.791	0.783	0.781
D	0.826	0.829	0.826	0.826
E	0.761	0.761	0.761	0.761

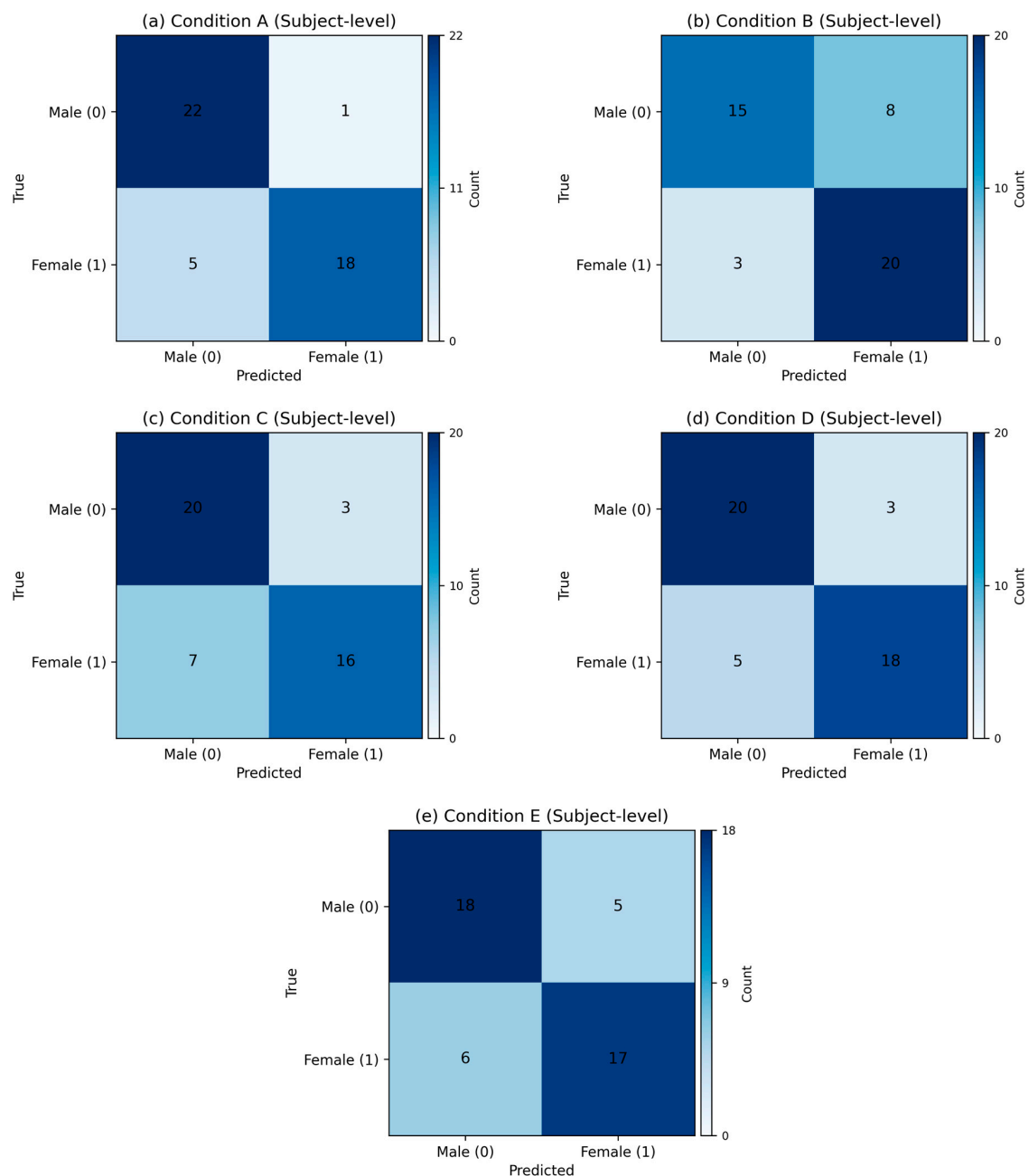


Figure 4. Subject-level confusion matrices (a–e) under the selected operating threshold $\tau = 0.54$. For each subject, clip-level predicted probabilities $p(\text{Female} = 1)$ were averaged and binarized using τ . Counts are aggregated over the held-out test subjects ($n = 46$; 23 male and 23 female). For readability, each panel uses its own color scale.

Table 4. Error decomposition by condition and sex with 95% subject-bootstrap confidence intervals. Confidence intervals were estimated via subject-level nonparametric bootstrap resampling (1000 resamples) on the held-out test subjects; in each resample, test subjects were sampled with replacement, and metrics were recomputed from the resampled subject-level predictions.

Condition	Accuracy (95% CI)	Precision_F (95% CI)	Recall_F (95% CI)	Precision_M (95% CI)	Recall_M (95% CI)	FPR	FNR
A	0.870 (0.783–0.957)	0.947 (0.823–1.000)	0.783 (0.591–0.947)	0.815 (0.652–0.958)	0.957 (0.857–1.000)	0.043	0.217
B	0.761 (0.630–0.870)	0.714 (0.533–0.880)	0.870 (0.706–1.000)	0.833 (0.643–1.000)	0.652 (0.450–0.857)	0.348	0.130
C	0.783 (0.652–0.891)	0.842 (0.650–1.000)	0.696 (0.500–0.870)	0.741 (0.560–0.895)	0.870 (0.720–1.000)	0.130	0.304
D	0.826 (0.717–0.935)	0.857 (0.688–1.000)	0.783 (0.591–0.947)	0.800 (0.621–0.950)	0.870 (0.720–1.000)	0.130	0.217
E	0.761 (0.630–0.891)	0.773 (0.583–0.947)	0.739 (0.548–0.913)	0.750 (0.560–0.913)	0.783 (0.600–0.950)	0.217	0.261

3.2. Error Decomposition by Condition and Sex

This section first specifies the operating threshold used to binarize predicted probabilities (Section 3.2.1) and then reports condition-wise confusion matrices and sex-wise error patterns under the fixed threshold (Section 3.2.2).

3.2.1. Threshold Selection and Sensitivity Analysis for Objective Label Binarization

To avoid an arbitrary probability threshold, we selected a data-driven global operating threshold τ^* on the validation set by sweeping τ in $[0.05, 0.95]$ (step = 0.01) and maximizing macro-F1. For the main setting ($L = 15$), this procedure yielded $\tau^* = 0.54$, which was used for all test-set confusion matrices and derived metrics reported in Table 4 and Figure 4 by binarizing subject-level mean probabilities. As a sensitivity check, we additionally swept τ on the test set and plotted macro-F1 as a function of τ for each condition (Supplementary Figure S1), confirming that the overall trends are stable around τ^* and not driven by a narrow threshold choice. Details of the validation-set threshold sweep are summarized in Supplementary Table S2.

3.2.2. Confusion Matrices and Error Asymmetry

Figure 4 presents condition-wise confusion matrices aggregated over all test subjects under the fixed operating threshold $\tau^* = 0.54$. The matrices visually confirm that the dominant misclassification direction differs by condition, which we quantify using recall-based error rates in Table 4. The dominant error tendency differed by condition (Table 4). In condition A, female recall was lower than male recall (Recall_F = 0.783; Recall_M = 0.957), yielding a higher FNR (0.217) and thus more female→male errors. In contrast, condition B showed reduced male recall (Recall_M = 0.652) and an elevated FPR (0.348), indicating a higher tendency to misclassify males as females under partial occlusion/carrying. Condition C also showed female recall lower than male recall (Recall_F = 0.696; Recall_M = 0.870), yielding a higher FNR (0.304). Condition D exhibited intermediate asymmetry (FPR/FNR = 0.130/0.217), while the normal-only subset (E) showed comparatively moderate recalls (Recall_F = 0.739; Recall_M = 0.783).

Across conditions, sex-wise recall was not symmetric (Table 4). In condition A, female recall was lower than male recall (0.783 vs. 0.957), yielding a higher FNR (0.217). In condition B, male recall decreased (0.652), and the FPR increased (0.348), indicating an elevated male→female misclassification tendency under occlusion/carrying. In conditions

C and D, overall accuracy was higher than B (0.783 and 0.826, respectively), but residual asymmetry remained (Table 4).

In addition, we visualized condition-wise predicted probability distributions by computing, for each subject in the test set, the mean of clip-level predicted probabilities $p(\text{Female} = 1)$, thereby obtaining a subject-level mean probability, and plotting sex-stratified histograms (Figure 5). Across conditions, the two distributions were generally separated, but the overlap pattern differed by perturbation. In condition A, a non-negligible portion of female subjects extended into the region below the operating threshold τ^* , consistent with the higher FNR (female $\rightarrow$ male errors) observed in Table 4. In condition B, the male distribution extended further above τ^* , increasing overlap on the high-probability side, which aligns with the elevated FPR and reduced male recall under partial occlusion/carrying. Conditions C and D showed intermediate overlap, while the normal-only reference subset (E) exhibited comparatively moderate separation.

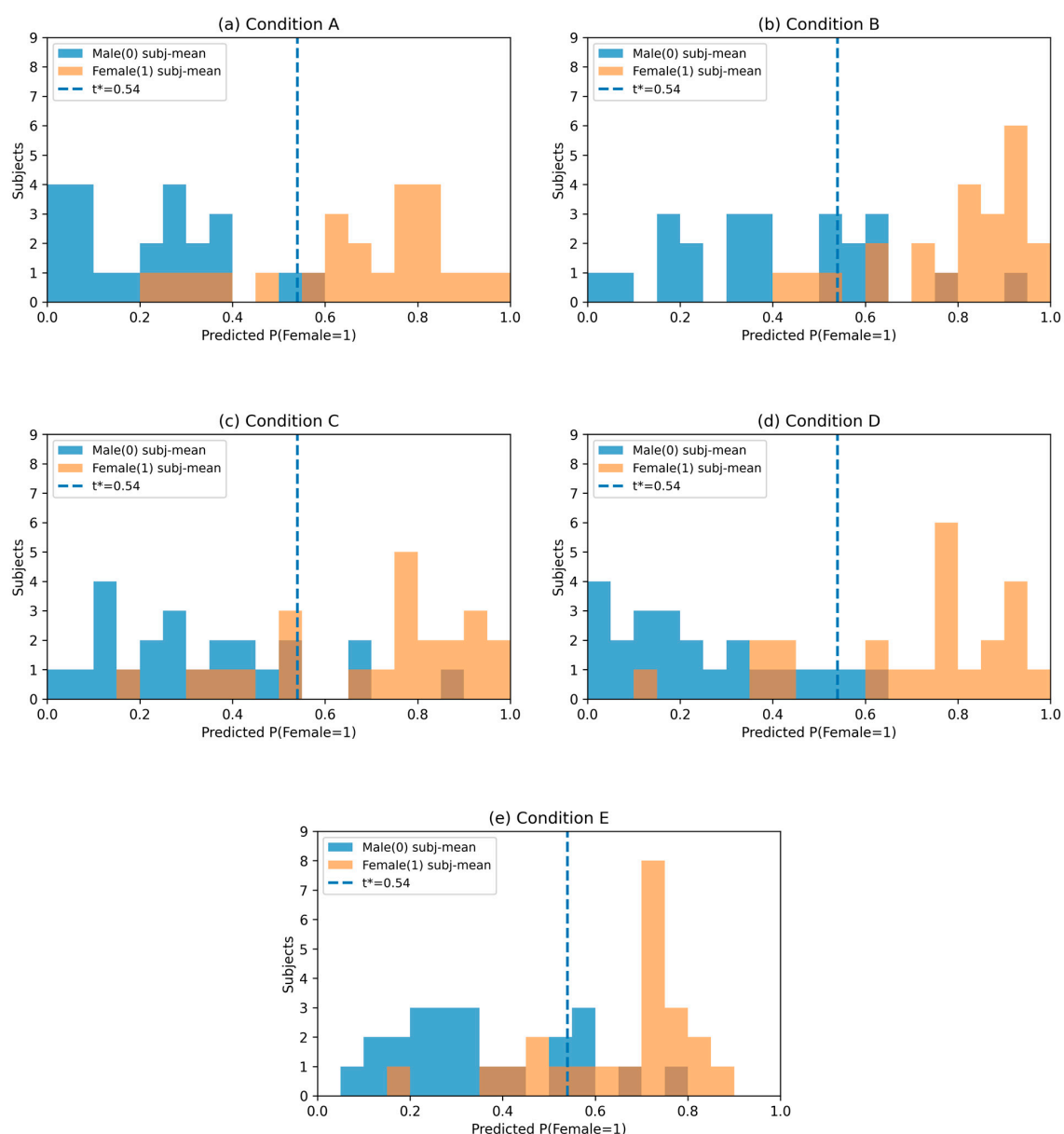


Figure 5. Subject-level mean predicted probability distributions by condition and sex (A–E). For each test subject, we computed the mean of clip-level predicted probabilities $p(\text{Female} = 1)$ and plotted sex-stratified histograms for (a) A, (b) B, (c) C, (d) D, and (e) E. The dashed vertical line indicates the selected operating threshold $\tau^*(0.54)$.

Proxy diagnostic for phase mismatch (within-subject prediction instability): To provide a lightweight surrogate for potential gait-cycle misalignment without explicit event detection, we quantified within-subject prediction instability from clip-level probabilities (Supplementary Table S5; Figure S3). Across conditions, instability increased under perturbations, with the largest variability observed in condition B (Figure S3; Table S5). Condition C also showed higher instability than the normal-only subset E, which is consistent with the possibility that speed-induced temporal variability interacts with fixed-length sampling without gait-cycle alignment.

3.3. Kinematic Evidence Under Smartphone Use

Among the condition-D test clips ($n = 807$), 806 clips were included in the kinematic analysis after excluding one clip with insufficient keypoint data for computing the arm-swing index (used_rows = 806). Under the smartphone-use condition (D), we evaluated the association between the 2D skeleton-based arm-swing index (arm_swing) and prediction confidence (confidence; $2 | p - 0.5 |$). Spearman correlation analysis showed a statistically significant but very weak negative association ($\rho = -0.0860$, $p = 0.0146$). A Mann–Whitney U test comparing arm_swing distributions between correctly classified and misclassified groups showed no significant difference ($p = 0.9713$; correct $n = 601$; wrong $n = 205$). The mean arm_swing was 2.4876 overall, 2.5022 in the correctly classified group, and 2.4447 in the misclassified group ($\Delta\text{mean} = 0.0575$).

3.4. Grad-CAM-Based Importance Distribution Across Conditions

We compared Grad-CAM-based importance distributions across conditions A–D. The joint-group (head, upper, trunk, lower) importance mass ratios are summarized in Table 5 (imp_all). In condition A, the importance was almost entirely concentrated on the lower-limb group (0.982), with negligible contributions from the upper body (0.014), trunk (0.003), and head (0.000). Condition B showed a more distributed pattern, although the lower-limb group remained the largest contributor (0.341), followed by upper (0.291), trunk (0.199), and head (0.169). In condition C, the head group dominated the Grad-CAM mass (0.892), while the other groups contributed at substantially lower levels (upper 0.051, trunk 0.039, lower 0.018). Condition D exhibited the most balanced allocation across groups, with slightly higher emphasis on the upper body (0.300) than on the lower (0.267), head (0.227), and trunk (0.206). These ratios were computed over valid clips for each condition (Table 5).

Table 5. Joint-group Grad-CAM mass ratios by condition (imp_all).

Condition	n^2	Head ¹	Upper	Trunk	Lower
A	1899	0.000	0.014	0.003	0.982
B	817	0.169	0.291	0.199	0.341
C	810	0.892	0.051	0.039	0.018
D	802	0.227	0.300	0.206	0.267
E	265	0.117	0.266	0.192	0.425

¹ Mean computed over valid clips (rounded to three decimal places). ² n denotes the number of valid clips used to compute the joint-group distribution.

Figure 6 shows the proportions of the summed Grad-CAM importance over three segments of a 45-frame timeline (1–15, 16–30, 31–45) for each condition. Overall, the temporal mass distribution was fairly balanced across segments, with only modest condition-dependent shifts. In condition A, the mid segment (16–30) contributed the smallest proportion, whereas the early (1–15) and late (31–45) segments accounted for relatively larger shares. In conditions B–D, the mid segment tended to take a slightly larger share than the early and late segments, indicating a mild mid-segment emphasis under perturbations.

Notably, condition C did not exhibit a strong late-segment dominance; instead, its mid and late segments were comparable.

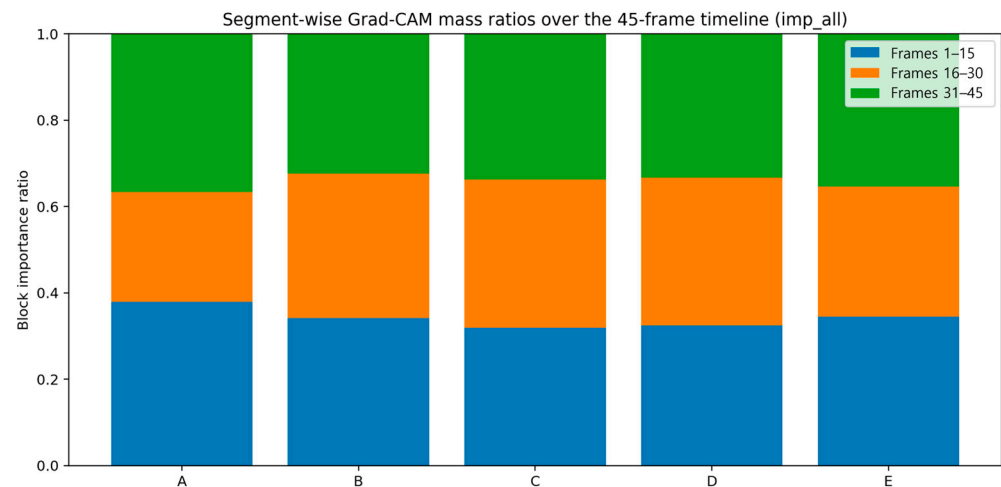


Figure 6. Segment-wise Grad-CAM mass ratios over the 45-frame timeline (imp_all). Each 45-frame sequence was partitioned into three consecutive 15-frame segments (early: 1–15, mid: 16–30, late: 31–45). For each sequence, Grad-CAM importance was summed within each segment and normalized by the total importance of the 45-frame sequence, and the resulting segment-wise proportions were compared across conditions.

3.5. Clip-Length Ablation

To address concerns about incomplete gait-cycle coverage with fixed-length clips, we conducted a clip-length ablation study ($L = 10, 15, 20$, and 30 frames) under the same subject-wise split and training protocol. Averaged over conditions A–D, the mean macro-F1 peaked at $L = 15$ (0.7307), followed by $L = 10$ (0.7107) and $L = 20$ (0.6894), while the mean macro-F1 decreased at $L = 30$ (0.6180). Consistent with the main setting ($L = 15$), the ablation results indicate that performance is highest around $L = 15$ (Table S3).

4. Discussion

This study analyzed the robustness of sex classification under gait perturbations that frequently occur in everyday environments, using 2D skeleton sequences from the PsyMo dataset. By applying a subject-wise stratified split with balanced sexes and configuring the aggregated conditions (A–D) and the normal-only reference subset (E) to share the same held-out test-subject set, we evaluated performance under a realistic generalization setting. Under the fixed operating threshold selected on the validation set ($\tau^* = 0.54$), the 3D residual CNN achieved subject-level condition-wise performance across A–D, with the lowest accuracy under the partial occlusion/carrying-related subset (B) and higher performance under the remaining conditions (A, C, and D) (Table 3). Beyond average accuracy, the confusion matrices (Figure 4) and the subject-level probability histograms (Figure 5) show that both the dominant misclassification direction and the degree of distributional overlap around τ^* vary by condition, indicating condition-dependent changes in failure modes and subgroup-conditional errors, most clearly in condition B (Table 4; Figures 4 and 5). Clip-level performance and clip-level confusion matrices are provided in the Supplementary Material for completeness.

Importantly, the gaps observed between the normal-only subset (E) and the disturbance-focused subsets (B–D) under the same subject split can be interpreted as a proxy comparison between relatively clean (“raw”) and perturbed (“harsh”) data. Accordingly, we discuss not only average accuracy but also how failure modes and subgroup-specific errors change

as nuisance factors become more dominant, using the condition-wise confusion matrices (Figure 4) and probability overlap patterns (Figure 5) as direct evidence.

4.1. Why the Largest Performance Degradation Occurred Under Partial Occlusion/Carrying (Condition B)

Condition B showed the clearest robustness degradation among the disturbance-focused settings. The confusion matrix (Figure 4) and the sex-wise error decomposition (Table 4) jointly indicate that the main failure mode in B is an increased tendency to misclassify males as females (higher FPR with reduced male recall), which is consistent with the possibility that occlusion/carrying destabilizes pose estimation and suppresses or distorts skeleton-derived upper-body motion cues. In the subject-level probability histograms (Figure 5), this manifests as a stronger intrusion of the male probability mass into the higher-probability region, increasing overlap on the “female” side relative to other conditions. Together, Figures 4 and 5 suggest that under occlusion/carrying the model’s decision boundary may become less reliable for male subjects, amplifying male→female errors.

In a skeleton-based pipeline, occlusion/carrying can manifest as (i) missing keypoints or coordinate jitter, (ii) reduced temporal consistency across frames, and (iii) changes in arm or upper-body posture due to task constraints. This interpretation is consistent with the pose-estimation literature, where occlusion is recognized as a central challenge in 2D pose estimation: it can reduce the reliability of joint localization and increase estimation uncertainty [19]. Quantitative reports further show that pose-estimation performance degrades as crowdedness/occlusion becomes more severe, supporting the view that occlusion translates into measurable performance loss [28].

From a modeling perspective, condition B likely weakens spatiotemporal discriminative cues that are relatively stable under normal walking. In our Grad-CAM analysis, the speed-variation condition (C) exhibited a head-dominant importance profile, whereas condition B showed reduced head contribution and a relative shift of importance toward the upper/trunk/lower groups (noting that interpretation in condition B requires caution due to potential pose-estimation unreliability under occlusion/carrying). This shift suggests that when head/upper-body keypoints become less reliable (e.g., due to occlusion or arm constraints) or when those cues are attenuated by perturbations, the network may compensate by relying more on comparatively observable lower-limb or trunk cues. Accordingly, improving robustness under occlusion may benefit from weighted inputs using keypoint confidence, occlusion-aware augmentation, temporal smoothing, or structural changes such as using coordinate- or graph-based representations rather than low-resolution skeleton-rendered images (see Section 4.6). Moreover, occlusion-induced keypoint jitter may be amplified by low-resolution (50×50) rasterization, potentially obscuring fine-grained motion cues in condition B.

4.2. Effects of Walking Speed Variations (Condition C): Phase Mismatch and Speed–Sex Confounding

Whereas occlusion primarily degrades input observability, speed variation constitutes a different perturbation that alters motion dynamics and temporal structure even under the same observation scheme. Therefore, interpreting condition C requires focusing on whether (i) gait-pattern changes and (ii) the temporal alignment strategy jointly influenced performance. Condition C (speed variation) achieved subject-level accuracy of 0.783 (Table 3) and exhibited a female→male-leaning error tendency (Table 4): female recall was lower than male recall (Recall_F = 0.696 vs. Recall_M = 0.870), yielding FPR/FNR = 0.130/0.304. This pattern suggests that speed variation in this setup is associated not only with general distribution shift and temporal inconsistency but also with non-negligible subgroup-conditional

error asymmetry, which is consistent with the possibility that speed-induced temporal variability interacts with fixed-length sampling without gait-cycle alignment.

Walking speed strongly modulates gait features and affects overall spatiotemporal gait parameters and joint kinematics. Prior biomechanics studies have reported that controlling for speed and body size can reduce sex differences previously observed in some spatiotemporal parameters, whereas certain whole-body joint-kinematic differences may persist even under dimensionless speed conditions [29]. In this study, it is likely that (i) speed-induced changes in gait strategy interacted with (ii) the use of fixed-length 15-frame clips without gait-cycle alignment or phase normalization based on events (e.g., heel strike). We omitted cycle alignment because heel-strike event detection was difficult to perform reliably using only 2D skeleton-rendered inputs, and event-detection errors could negate potential benefits of phase alignment. However, this design choice leaves a limitation whereby performance under speed variation may degrade due to phase mismatch; thus, future work should consider lightweight phase normalization based on periodicity estimation or weak alignment that does not require explicit event detection.

As a proxy diagnostic of within-subject temporal inconsistency that may arise from fixed-length sampling without gait-cycle alignment, we report subject-level prediction-instability metrics (e.g., mixing and $\text{std}(p)$) and the subject-level distance to the fixed operating threshold $|\text{mean}(p) - \tau^*|$ across conditions in the Supplementary Material (Figure S3 and Table S5; $\tau^* = 0.54$). The instability metrics indicate that perturbations can increase within-subject variability, most notably in condition B, while E shows the lowest $\text{std}(p)$. These results do not constitute gait-event alignment; rather, they provide reproducible evidence that fixed-length sampling can be sensitive to temporal variability. As the reviewer suggested, future work should evaluate lightweight weak-alignment strategies (e.g., periodicity-based normalization or proxy-event anchoring) once detector reliability can be validated under 2D skeleton noise. With fixed-length 15-frame clips, the sampled window may cover only a portion of a gait cycle, and the relative gait phase can vary across walking speeds and subjects when event-based segmentation is not used. This phase variability can weaken the correspondence between the learned spatiotemporal patterns and biomechanical gait events. Consistent with this, the segment-wise Grad-CAM mass ratios over the 45-frame timeline (Figure 6) are distributed across early (1–15), mid (16–30), and late (31–45) segments, with only modest condition-dependent shifts; thus, the current segment-level analysis alone does not clearly localize event-timed differences. As the reviewer suggested, future work should test whether gait-event-aware alignment or weakly event-anchored temporal normalization (e.g., periodicity-based normalization or proxy event cues that avoid explicit heel-strike detection) can reduce phase mismatch and improve robustness under speed perturbations.

4.3. Smartphone Dual-Task (Condition D): High Accuracy and Limitations of a Simple Arm-Swing Index

Condition D represents a prototypical real-world perturbation in which cognitive, visual, and upper-limb tasks are simultaneously imposed (dual-task), and prior work has reported gait-strategy changes under such conditions [30,31]. In our study, condition D achieved subject-level accuracy of 0.826, close to the best-performing condition A (0.870) (Table 3). Nevertheless, high accuracy did not imply perfectly symmetric subgroup behavior: Table 4 shows $\text{Recall_M} (0.870) > \text{Recall_F} (0.783)$ with a non-negligible FNR of 0.217, indicating residual female $\rightarrow$ male errors under the fixed threshold. Consistently, Figure 5d shows that the female distribution retains some mass near or below τ^* , which explains why the FNR remains non-negligible despite reasonably separated distributions.

Motivated by prior studies linking smartphone texting to altered gait strategies, we further examined whether a simple arm-swing amplitude proxy derived from 2D skeletons

relates to model uncertainty and errors under condition D [32]. The arm-swing index showed a statistically significant but very weak negative association with prediction confidence (Spearman $\rho = -0.0860$, $p = 0.0146$), and it did not differ between correctly and incorrectly classified clips (Mann–Whitney U $p = 0.9713$). Taken together, these results suggest that this specific amplitude-range proxy alone does not explain misclassification or provide a strong account of the model’s confidence variation under smartphone use.

Importantly, the smartphone-use subset (TXT/PH) does not disentangle cognitive load from upper-limb constraints, which limits causal interpretation; therefore, the kinematic analysis should be interpreted as descriptive rather than causal, and future work should isolate these factors via controlled task designs. To prevent overinterpretation, the weak/non-separable findings should be understood as evidence that this specific amplitude-based proxy may not capture the motion cues used by the classifier under smartphone use, rather than as evidence that upper-limb motion is irrelevant.

A single-view 2D amplitude proxy cannot represent the full complexity of 3D motion, and discriminative upper-limb cues may involve not only amplitude but also phase relationships, left–right asymmetry, and trunk–arm coordination. Therefore, future work should re-evaluate informative cues under smartphone use using robustness-aware keypoint quality filtering/normalization, frequency-domain features, inter-joint coupling metrics, and embedding-based interpretation from skeleton-native models.

4.4. Meaning and Uncertainty of Sex-Wise Error Asymmetry: The Need for Reliability and Bias Checks

A key contribution of this study is demonstrating that average accuracy alone is insufficient for assessing robustness, as evidenced by condition-dependent sex-wise error asymmetry, most notably with opposite dominant error directions in A and B. If perturbations amplify a tendency to misclassify males as females, applications such as HCI personalization or user-context inference may accumulate systematic errors that disadvantage a specific subgroup even when overall performance appears acceptable.

Methodologically, we used subject-level nonparametric bootstrap confidence intervals to estimate uncertainty while accounting for clustered data, where multiple clips are generated per subject [33]. This choice is important because treating clips as independent samples can underestimate uncertainty. Bootstrap methods also provide a classical methodological basis for uncertainty quantification of complex statistics [26]. Future work should go beyond threshold-based (0.5) binarized metrics and examine probability calibration (e.g., reliability diagrams, ECE) and condition-wise shifts in probability distributions to evaluate whether a single threshold is appropriate across all conditions.

4.5. Grad-CAM Interpretation: Reallocation of Joint Cues and Cautions in Interpretation

When error asymmetry is observed, the next step is to inspect, by condition, what the model used as evidence and connect it to plausible causes at a hypothesis level. Grad-CAM, as applied here, produces class-discriminative localization maps by using gradients flowing into the final convolutional layer. Using the joint-group Grad-CAM mass ratios (Table 5), we observed condition-dependent allocation patterns rather than a fixed evidential basis. In condition A, importance was almost entirely concentrated on the lower-limb joint group, suggesting that the classifier primarily relied on lower-body motion cues under this setting. In condition C, the head joint group dominated the attribution mass, indicating that the salient cues can shift substantially under speed variation. Condition B showed a more distributed profile across joint groups, with the largest share still assigned to the lower-limb group but with non-negligible contributions from the upper and trunk groups. Condition D exhibited the most balanced allocation, with a mild emphasis on the upper-body group relative to other regions. Together, these patterns suggest that perturbations may change

the discriminative motion cues available to the model and/or alter pose quality in ways that redistribute which joint groups become most informative in the 50×50 skeleton-rendered representation. Accordingly, Grad-CAM should be interpreted as a hypothesis-generating diagnostic of model behavior rather than as evidence of biomechanical causality.

Moreover, we did not directly validate the Grad-CAM-derived joint-group importance against independent biomechanical measurements (e.g., joint-angle ranges, spatiotemporal parameters, or laboratory-grade kinematics). Therefore, the current interpretation should be viewed as model-centric evidence rather than external biomechanical validation. Future work should test whether condition-dependent saliency shifts correspond to independent kinematic/biomechanical measures, for example, by comparing saliency changes with spatiotemporal gait variability, inter-joint coordination proxies, or task-controlled perturbation experiments, rather than interpreting Grad-CAM as direct biomechanical causality. Importantly, the kinematic proxy used in this revision is derived from the same 2D skeleton source and therefore does not constitute an independent biomechanical measurement. Future work will incorporate cross-modal or external kinematics (e.g., laboratory-grade motion capture/IMU, or dataset-provided 3D body models where available) to validate whether saliency shifts correspond to independently measured gait variables.

4.6. Practical Considerations, Limitations, and Future Work

From a practical standpoint, skeleton-based approaches can preserve movement structure while reducing reliance on appearance cues (e.g., face or clothing). PsyMo enables reproducible benchmarking by releasing relatively anonymized gait representations such as silhouettes, 2D/3D skeletons, and SMPL. However, privacy is not binary: because sensitive-attribute inference may remain possible even from skeletons, additional protection strategies may be required, and anonymization-oriented research has been proposed. Therefore, when sex classification is used in user profiling or surveillance contexts, contextual review—including application purpose, ethics, and data-minimization principles—is essential alongside technical performance.

This study has several limitations: (i) reliance on the dataset's binary sex labels and self-reported annotations (which may not reflect sex identity and may introduce labeling bias). Accordingly, our results should be interpreted as performance on the dataset-provided binary sex label rather than as inference of gender identity, and future work should consider more inclusive labeling schemes and governance when available, (ii) evaluation on a single dataset (PsyMo) and a single side-view setting, which may limit generalizability across acquisition settings and populations, (iii) absence of direct benchmarking against alternative pipelines (e.g., Kinect-based skeletons or full-image sex classification baselines), making it difficult to quantify the relative advantage of skeleton rendering under matched conditions, (iv) potential increased sensitivity to speed variation due to fixed-length clip sampling without phase alignment, (v) the correlational and model-centric nature of Grad-CAM interpretation without external biomechanical validation, and (vi) possible information loss due to low-resolution (50×50) rasterization of skeletons that may obscure fine-grained dynamics; accordingly, a resolution sensitivity check (e.g., 50×50 vs. higher rasterization resolutions with appropriate regularization) will help determine whether the observed performance and interpretation trends are robust to rasterization granularity. Importantly, we do not claim superiority over appearance-based full-image pipelines; rather, our focus is to provide a condition-wise robustness audit and interpretable failure-mode analysis under a privacy-preserving skeleton representation, which is not readily available in typical full-image baselines.

Future work will include (i) a resolution sensitivity check (e.g., 50×50 vs. higher rasterization resolutions with appropriate regularization) to determine whether the ob-

served performance and interpretation trends are robust to rasterization granularity, (ii) improving robustness via keypoint confidence and occlusion-aware learning, (iii) benchmarking skeleton-native representations (coordinate- or graph-based) that avoid discretization artifacts while preserving temporal detail, enabling clearer attribution of errors to representation choice rather than input discretization, (iv) addressing speed perturbations through lightweight phase alignment without explicit event detection, (v) extending evaluation to multi-view settings and more diverse demographic distributions, (vi) cross-dataset validation and domain-shift evaluation on additional gait datasets to assess robustness beyond PsyMo, and (vii) including matched baselines (e.g., full-image models and Kinect-based skeleton pipelines) and, where available within the same dataset, alternative released representations (e.g., silhouettes or SMPL) under an identical subject split to quantify privacy–robustness–performance trade-offs fairly; this will also help contextualize the subject-level accuracy range (0.761–0.870). Moreover, future studies should report condition-wise subgroup-conditional performance and examine whether calibration or domain-adaptation strategies can mitigate systematic misclassification patterns.

5. Conclusions

Using a subject-wise split, we compared performance and error patterns across four aggregated walking conditions (A–D) that reflect common real-world perturbations. Under a subject-wise split, subject-level accuracy ranged from 0.761 to 0.870 across conditions A–D, with the lowest performance under partial occlusion/carrying (Condition B). Error decomposition revealed condition-dependent sex-wise error asymmetry, most notably under occlusion/carrying, where male recall decreased and the tendency to misclassify males as females increased. Together, these findings quantify not only average performance changes under perturbations but also how subgroup-wise error imbalance can be amplified depending on disturbance type.

Error decomposition revealed condition-dependent sex-wise asymmetry: condition B showed increased male→female errors (higher FPR with reduced male recall), whereas condition C showed increased female→male errors (higher FNR with reduced female recall) under the fixed threshold τ^* . In contrast, under smartphone use (D), *arm_swing* was very weakly but significantly associated with confidence, while it did not differ between correct and incorrect predictions. Together, these findings quantify not only average performance changes under perturbations but also how group-wise error imbalance can be amplified across conditions. This work provides an empirical basis for improving both accuracy and reliability through model and training refinements, thereby supporting more generalizable and trustworthy skeleton-based gait sex classification in everyday settings.

Future work should further examine input-quality factors, including pose-estimation reliability, and test whether phase normalization, as well as calibration and condition-adaptive learning, can improve performance stability and mitigate sex-wise error asymmetry under perturbations.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/app16041830/s1>, Table S1: Train/test subjects and clip counts per aggregated condition under the shared split; Table S2: Validation-set threshold selection summary (macro-F1 maximization); Table S3: Clip-length sensitivity analysis ($L = 10, 15, 20, 30$); Table S4: Layer-wise configuration of the ResNet-18-style 3D residual backbone with an MLP classification head (input: $15 \times 50 \times 50 \times 1$); Table S5: Summary statistics of the instability proxy on the held-out test set (median and interquartile range [Q1–Q3] across subjects; $\tau^* = 0.54$); Figure S1: Threshold-sensitivity analysis on the test set (macro-F1 vs. τ); Figure S2: Clip-level confusion matrices (A–E) under the fixed operating threshold $\tau^* = 0.54$; Figure S3: Subject-level prediction instability (instability proxy) across walking conditions (A–E) under the fixed operating threshold $\tau^* = 0.54$.

Author Contributions: Conceptualization, A.H.J. and S.D.M.; methodology, A.H.J.; software, A.H.J.; validation, A.H.J. and Y.O.; formal analysis, A.H.J.; investigation, A.H.J., M.-H.C. and Y.E.K.; resources, S.D.M.; data curation, A.H.J. and Y.O.; writing—original draft preparation, A.H.J.; writing—review and editing, A.H.J., Y.O., Y.E.K. and S.D.M.; visualization, A.H.J., M.-H.C. and Y.E.K.; supervision, S.D.M.; project administration, S.D.M.; funding acquisition, S.D.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Research Foundation of Korea (NRF) funded by the Ministry of Education (Grant No. NRF-2021R1I1A3059769); by BK21 FOUR (Fostering Outstanding Universities for Research) funded by the Ministry of Education (Grant No. 5199990914048); and by the Soonchunhyang University Research Fund.

Institutional Review Board Statement: Not applicable. This study used a publicly available dataset (PsyMo), and no new data were collected from human participants by the authors.

Informed Consent Statement: Not applicable. The study used a publicly available dataset (PsyMo) provided by the dataset authors.

Data Availability Statement: The PsyMo dataset is available from the original authors/source (see Ref. [17]) under the terms and conditions described therein. The code used for training and evaluation is available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

PsyMo tag abbreviations: BG, bag carrying; CL, changing clothing; NM, normal walking; PH, talking on the phone; TXT, walking while texting; WSF, walking speed fast; WSS, walking speed slow. BN, batch normalization; CI, confidence interval; CNN, convolutional neural network; COCO, Common Objects in Context; FNR, false negative rate; FPR, false positive rate; Grad-CAM, Gradient-weighted Class Activation Mapping; HCI, human–computer interaction; L, clip length; ReLU, rectified linear unit; SMPL, Skinned Multi-Person Linear model; TP/TN/FP/FN, true positive/true negative/false positive/false negative (positive class = female (label 1), negative class = male (label 0)).

References

1. Nixon, M.S.; Tan, T.; Chellappa, R. *Human Identification Based on Gait*; Springer: Boston, MA, USA, 2006. [\[CrossRef\]](#)
2. Xu, C.; Makihara, Y.; Ogi, G.; Li, X.; Yagi, Y.; Lu, J. The OU-ISIR Gait Database Comprising the Large Population Dataset with Age and Performance Evaluation of Age Estimation. *IPSJ Trans. Comput. Vis. Appl.* **2017**, *9*, 24. [\[CrossRef\]](#)
3. Franco, D.; Russo, M.; Amboni, M.; Ponsiglione, A.; Di Filippo, A.; Romano, M.; Amato, F.; Ricciardi, C. The Role of Deep Learning and Gait Analysis in Parkinson’s Disease: A Systematic Review. *Sensors* **2024**, *24*, 5957. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Bisogni, C.; Cimmino, L.; Nappi, M.; Pannese, T.; Pero, C. Walk as you feel: Privacy preserving emotion recognition from gait patterns. *Eng. Appl. Artif. Intell.* **2024**, *128*, 107565. [\[CrossRef\]](#)
5. Galiyawala, H.; Raval, M.S.; Patel, M. Person Retrieval in Surveillance Video using Height, Color and Gender. *arXiv* **2018**, arXiv:1810.05080. [\[CrossRef\]](#)
6. Zhang, W.; Smith, M.L.; Smith, L.N.; Farooq, A. Gender and gaze gesture recognition for human–computer interaction. *Comput. Vis. Image Underst.* **2016**, *149*, 32–50. [\[CrossRef\]](#)
7. Naham, A.-Z.; Wang, J.; Raaed, A.-S. Multi-Task Learning and Gender-Aware Fashion Recommendation System Using Deep Learning. *Electronics* **2023**, *12*, 3396. [\[CrossRef\]](#)
8. Li, B.; Lian, X.-C.; Lu, B.-L. Gender classification by combining clothing, hair and facial component classifiers. *Neurocomputing* **2012**, *76*, 18–27. [\[CrossRef\]](#)
9. Sun, Z.; Liu, Z. Ensuring privacy in face recognition: A survey on data generation, inference and storage. *Discov. Appl. Sci.* **2025**, *7*, 441. [\[CrossRef\]](#)
10. Yu, S.; Tan, T.; Huang, K.; Jia, K.; Wu, X. A Study on Gait-Based Gender Classification. *IEEE Trans. Image Process.* **2009**, *18*, 1905–1910. [\[CrossRef\]](#)

11. Hu, M.; Wang, Y.; Zhang, Z.; Zhang, D. Gait-Based Gender Classification Using Mixed Conditional Random Field. *IEEE Trans. Syst. Man Cybern. B Cybern.* **2011**, *41*, 1429–1439. [\[CrossRef\]](#)
12. Azhar, M.; Ullah, S.; Ullah, K.; Syed, I.; Choi, J. A Gait-Based Real-Time Gender Classification System Using Whole Body Joints. *Sensors* **2022**, *22*, 9113. [\[CrossRef\]](#)
13. Azhar, M.; Ullah, S.; Raees, M.; Rahman, K.U.; Rehman, I.U. A real-time multi view gait-based automatic gender classification system using kinect sensor. *Multimed. Tools Appl.* **2023**, *82*, 11993–12016. [\[CrossRef\]](#)
14. Parashar, A.; Ding, W.; Shekhawat, R.S.; Rida, I. Deep Learning Pipelines for Recognition of Gait Biometrics with Covariates: A Comprehensive Review. *Artif. Intell. Rev.* **2023**, *56*, 8889–8953. [\[CrossRef\]](#)
15. Mughal, A.B.; Khan, R.U.; Bermak, A.; ur Rehman, A. Person Recognition via Gait: A Review of Covariate Impact and Challenges. *Sensors* **2025**, *25*, 3471. [\[CrossRef\]](#)
16. Chen, Y.; Peng, K.; Roitberg, A.; Schneider, D.; Zhang, J.; Zheng, J.; Liu, R.; Chen, Y.; Yang, K.; Stiefelbogen, R. Exploring Self-supervised Skeleton-based Action Recognition in Occluded Environments. *arXiv* **2023**, arXiv:2309.12029. [\[CrossRef\]](#)
17. Schabrun, S.M.; van den Hoorn, W.; Moorcroft, A.; Greenland, C.; Hodges, P.W. Texting and Walking: Strategies for Postural Control and Implications for Safety. *PLoS ONE* **2014**, *9*, e84312. [\[CrossRef\]](#)
18. Bruyneel, A.-V.; Reinmann, A.; Gafner, S.C.; Sandoz, J.-D.; Duclos, N.C. Does texting while walking affect spatiotemporal gait parameters in healthy adults, older people, and persons with motor or cognitive disorders? A systematic review and meta-analysis. *Gait Posture* **2023**, *100*, 284–301. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Cao, Z.; Simon, T.; Wei, S.-E.; Sheikh, Y. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 22–25 July 2017; pp. 7291–7299. [\[CrossRef\]](#)
20. Cosma, A.; Radoi, I. PsyMo: A Dataset for Estimating Self-Reported Psychological Traits from Gait. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2024; pp. 4603–4613. [\[CrossRef\]](#)
21. Cormier, M.; Schmid, Y.; Beyerer, J. Enhancing Skeleton-Based Action Recognition in Real-World Scenarios through Realistic Data Augmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), Waikoloa, HI, USA, 4–8 January 2024; pp. 290–299.
22. Mathe, E.; Vernikos, I.; Spyrou, E.; Mylonas, P. Leveraging Artificial Occluded Samples for Data Augmentation in Human Activity Recognition. *Sensors* **2025**, *25*, 1163. [\[CrossRef\]](#)
23. Singh, R.; Sharma, A. Occluded skeleton-based multi-stream model using Part-Aware Spatial–Temporal Graph Convolutional Network for human activity recognition. *Eng. Appl. Artif. Intell.* **2025**, *156*, 111183. [\[CrossRef\]](#)
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [\[CrossRef\]](#)
25. Yang, J.; Sun, S.; Chen, J.; Xie, H.; Wang, Y.; Yang, Z. 3D-STARNET: Spatial–Temporal Attention Residual Network for Robust Action Recognition. *Appl. Sci.* **2024**, *14*, 7154. [\[CrossRef\]](#)
26. Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* **1979**, *7*, 1–26. [\[CrossRef\]](#)
27. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. [\[CrossRef\]](#)
28. Li, J.; Wang, C.; Zhu, H.; Mao, Y.; Fang, H.-S.; Lu, C. CrowdPose: Efficient Crowded Scenes Pose Estimation and a New Benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 10863–10872. Available online: https://openaccess.thecvf.com/content_CVPR_2019/html/Li_CrowdPose_Efficient_Crowded_Scenes_Pose_Estimation_and_a_New_Benchmark_CVPR_2019_paper.html (accessed on 20 January 2026).
29. Bruening, D.A.; Baird, A.R.; Weaver, K.J.; Rasmussen, A.T. Whole body kinematic sex differences persist across non-dimensional gait speeds. *PLoS ONE* **2020**, *15*, e0237449. [\[CrossRef\]](#)
30. Bovonsunthonchai, S.; Ariyadomkit, R.; Susilo, T.E.; Sangiamwong, P.; Puchaphan, P.; Chandee, S.; Richards, J. The impact of different mobile phone tasks on gait behaviour in healthy young adults. *J. Transp. Health* **2020**, *19*, 100920. [\[CrossRef\]](#)
31. Crowley, P.; Madeleine, P.; Vuillerme, N. The effects of mobile phone use on walking: A dual task study. *BMC Res. Notes* **2019**, *12*, 352. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Umberger, B.R. Effects of suppressing arm swing on kinematics, kinetics, and energetics of human walking. *J. Biomech.* **2008**, *41*, 2575–2580. [\[CrossRef\]](#)
33. Field, C.A.; Welsh, A.H. Bootstrapping clustered data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2007**, *69*, 369–390. [\[CrossRef\]](#)

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.