SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Explainable Domain-Robust Fall Detection Using Temporal Pose Features

Shabana S[1], T Rajasundari[2], Varshitha D N[3], Ranjan Kumar H S*[4]

[1] Assistant Professor, Department of Computer Science and Engineering, Nitte (Deemed to be University), NMAM Institute of Technology, Nitte, Karnataka, India. E-mail: shabana.s@nitte.edu.in.
[2] Assistant Professor, Department of Artificial Intelligence and Machine Learning, Manakula Vinayagar Institute of Technology, Puducherry, India. E-mail: rajasundariaiml@mvit.edu.in.
[3] Associate Professor, Department of CSE (AI&ML), Vidyavardhaka College of Engineering, Mysuru, Karnataka, India.
E-mail: varshithadn@vvce.ac.in
[4] Associate Professor, Department of Artificial Intelligence and Machine Learning, Nitte (Deemed to be University), NMAM Institute of Technology, Nitte, Karnataka, India. E-mail: ranjan.hs@nitte.edu.in
*Corresponding Author: Ranjan Kumar H. S., E-mail: ranjan.hs@nitte.edu.in.

Abstract
Vision-based fall detection can fail when camera geometry and acquisition conditions change. This study developed an explainable temporal pose pipeline and evaluated it using progressively stricter leakage-safe protocols. YOLO11n-Pose key points were transformed into robust temporal descriptors. Video-independent URFD validation achieved 81.67% balanced accuracy, whereas unchanged transfer to MCFD produced 46.99%, confirming substantial domain shift. Under simultaneous unseen-camera and unseen-scenario testing, invariant-feature learning achieved 62.21%; offline complete-interval summarization increased this upper bound to 71.16%. In a distinct fixed-installation experiment, camera-aware synchronized multi-view fusion achieved 87.23% balanced accuracy, 82.35% sensitivity and 92.11% specificity on unseen chute scenarios. Its improvement over permutation-invariant fusion was not statistically significant. The findings show that temporal evidence and synchronized views can improve recognition under specified deployment conditions, while fixed-camera multi-view performance must not be interpreted as unseen-camera generalization or real-time clinical readiness.

Keywords: fall detection; human pose estimation; domain shift; leakage-safe validation; multi-camera evaluation; explainable machine learning

1. Introduction

Falls are a major public-health concern. The World Health Organization estimates that approximately 684,000 fatal falls occur annually and that 37.3 million falls require medical attention [1]. Contactless video monitoring may support timely intervention without requiring a person to wear or charge a device. However, vision systems are sensitive to viewpoint, occlusion, lighting, background, body scale and pose-estimation quality. These conditions create a gap between controlled benchmark performance and reliable operation in a new room or camera installation. The source project investigated real-time posture classification and fall alerts using person bounding boxes, skeletal keypoints, aspect-ratio rules and persistence logic. That prototype established a practical workflow but did not include dataset-independent statistical evaluation. The present study reconstructs the decision layer, replaces single-frame posture rules with robust temporal descriptors, and performs video-independent, cross-dataset, multi-camera and dual-holdout evaluation. A model that performs well on one dataset can fail when camera geometry, scene, motion distribution and pose confidence change simultaneously. Prior fall-detection work has used silhouettes, depth maps, convolutional networks, skeleton sequences and multi-camera fusion [2]-[7]. Human-pose research also identifies cross-dataset pose estimation as a domain-adaptation problem [8]-[10]. Nevertheless, random frame or clip splits, single-dataset optimization and evaluation of different views of the same simulated event can overestimate generalization. As conceptualized in Figure 1, apparently strong within-dataset performance may deteriorate when a new camera or environment changes pose geometry, motion scale, confidence and background conditions.

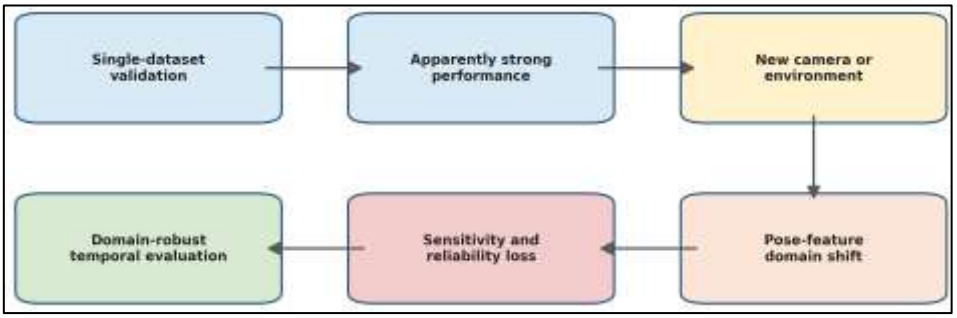

Figure 1. Motivation for domain-robust and leakage-safe evaluation of vision-based fall detection.

Figure 1 summarizes the central research gap: benchmark fitting is not equivalent to deployment generalization. Camera- and environment-induced pose-feature shift can reduce sensitivity and reliability even when the source-domain score is high. This progression motivates the robust temporal representation, explicit domain-shift analysis, and grouped evaluation used in the present study.
This paper addresses that methodological gap through the following contributions:

- An explainable 15-frame representation combining normalized geometry, torso angle, angular change, hip velocity, height reduction, pose motion, posture persistence, confidence, and coverage.
- Nested video-independent URFD evaluation followed by unchanged, eight-camera zero-shot MCFD testing.
- Independent-sequence domain-shift quantification using standardized feature differences and a cross-validated domain classifier.
- A strict 40-fold protocol that simultaneously withholds an entire camera and complete chute scenarios from both fitting and operating-point calibration.
- Cluster-aware confidence intervals and paired permutation tests that treat a complete synchronized chute as the inferential unit.
- A stronger offline full-interval temporal benchmark using distributional and trajectory summaries, reported explicitly as an upper bound rather than an online alarm result.
- Leakage-safe permutation-invariant and camera-aware synchronized multi-view fusion for a fixed installation, with complete chute scenarios withheld from training and threshold calibration.

2. Related Work

2.1 Vision and skeleton-based fall detection

Early camera-based systems described falls through silhouette geometry, motion history, head trajectory and multi-view reconstruction. Auvinet et al. [2], [3] developed an occlusion-resistant multi-camera method and released the MCFD benchmark used in this study. Kwolek and Kepski [6] combined depth maps with an accelerometer and released URFD, which contains 30 fall and 40 activity-of-daily-living sequences. Recent approaches increasingly rely on learned detectors and pose estimators. Skeleton representations can reduce background dependence and provide interpretable joint-level motion, but errors in keypoint localization can propagate into downstream classification [4]. Additional research on YOLO-pose shows the possibility of real-time skeletal inference, but typically focuses on within-dataset evaluation [7].

2.2 Temporal decision making

A fall is a transition, not a static posture. A horizontal body can be a sleeping body or an exercise or completed fall, a true fall may have only a short horizontal phase. Time-level decision fusion and skeleton-sequence models resolve this ambiguity by integrating evidence across multiple frames [4], [5]. The current work uses a robust trailing-window summarization and a persistence rule to model motion and post-fall state while maintaining interpretability.

2.3 Domain shift and evaluation leakage

Camera pose, focal geometry, human appearance, action execution, and keypoint noise change the distribution of pose features. AdaptPose demonstrated cross-dataset adaptation through learnable motion generation [9]; source-free adaptation and dual augmentation have since addressed target specialization and domain generalization for human pose estimation [10], [11]. These studies motivate explicit domain analysis in fall detection. Our emphasis differs from end-to-end pose adaptation: the pose network remains fixed, while the interpretable downstream feature space is measured, filtered, and evaluated under synchronized-scene-safe holdouts.

2.4 Positioning of the present study

The present work is positioned between rule-based explainable detectors and end-to-end sequence networks. Unlike studies emphasizing a high score under a random frame split, it makes the grouping unit explicit, freezes the zero-shot model before external testing, quantifies feature shift, and prevents synchronized recordings of a held-out event from entering training through another camera. Unlike computationally heavier end-to-end adaptation, it preserves an auditable pose-feature representation. Its principal novelty is therefore the combination of interpretable temporal modelling, camera-aware synchronized fusion, and scenario-safe statistical evaluation rather than a new pose backbone.

3. Materials and Methods

Figure 2 provides an overview of the complete study pipeline. Synchronized RGB views are converted into skeletal keypoints and normalized frame-level quantities. Finally, we construct robust temporal descriptors and complete interval trajectory summaries, fuse them across calibrated cameras, and evaluate them using nested chute-grouped selection with untouched outer-test scenarios.

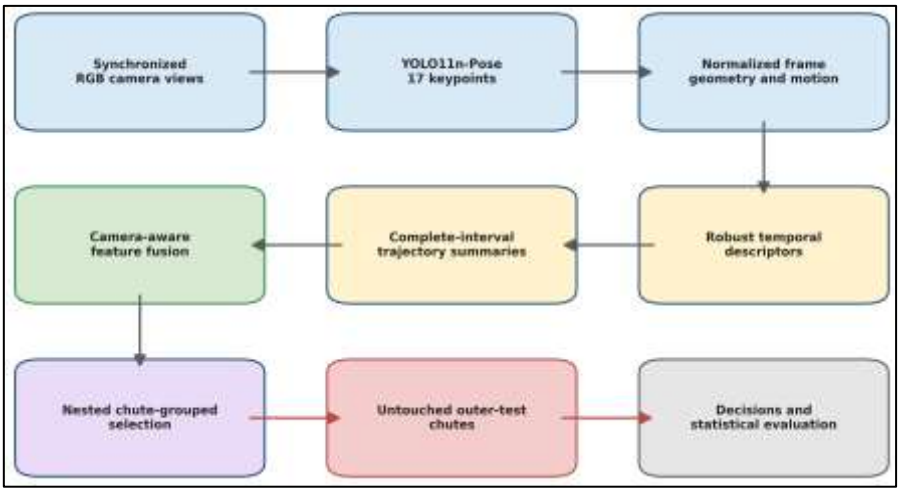

Figure 2. Proposed explainable camera-aware multi-view fall-recognition pipeline.

The arrows in Figure 2 show the processing order and, importantly, the difference between the model development and the final evaluation. The permitted training chutes are used for feature construction, camera-aware fusion, model selection and threshold calibration, and the locked configuration is applied once on the outer-test chutes. The following subsections mathematically and operationally define each component.

3.1 Datasets

URFD was used as the source domain. Its official release contains 70 sequences: 30 simulated falls and 40 activities of daily living [6]. The experiment was reconstructed from RGB frames of camera 0, and all source-domain folds were grouped by complete video rather than by frame.

MCFD was used as the external target domain. It contains simulated falls and daily activities recorded using eight calibrated cameras [2], [3]. Cameras 1-8 provided 23 annotated chute scenarios. Each view was downsampled from 120 to 30 frames per second over the annotated range, with 15 frames of context preceding the annotation. Cameras 1, 2 and 4-8 each contained 69 annotated intervals, whereas camera 3 contained 68, yielding 551 camera-specific intervals (263 positive and 288 negative).

3.2 Pose extraction and frame-level quantities

YOLO11n-Pose was applied at an input size of 640 pixels using the Ultralytics implementation [11]. For each frame, the highest-confidence person detection supplied a bounding box and 17 COCO-format keypoints. Geometrical quantities were normalized by bounding-box dimensions to reduce sensitivity to camera distance. To reduce sensitivity to camera distance and person scale, the key point coordinates were normalized relative to the dimensions of the detected bounding box, as defined in Equation (1). The bounding-box aspect ratio, which provides evidence of horizontal posture, was calculated using Equation (2). Torso orientation was represented by the shoulder-to-hip angle in Equation (3), while downward hip velocity was estimated from consecutive frames using Equation (4). Whole-body pose motion was computed as the median normalized displacement of the available key points according to Equation (5).

$$\tilde{x}_{j,t} = \frac{x_{j,t} - x_t^{box}}{w_t}, \qquad \tilde{y}_{j,t} = \frac{y_{j,t} - y_t^{box}}{h_t} \tag{1}$$

$$AR_t = \frac{w_t}{h_t} \quad (2)$$

$$\theta_t = \operatorname{atan2}\left(\left|y_t^S - y_t^H\right|, \left|x_t^S - x_t^H\right|\right) \quad (3)$$

$$v_t = \frac{y_t^H - y_{t-1}^H}{h_t \Delta t} \quad (4)$$

$$M_t = \underset{j \in K}{\operatorname{median}} \frac{\left\|k_{j,t} - k_{j,t-1}\right\|_2}{h_t} \quad (5)$$

Prior to aggregation, implausible one-frame motion and velocity values were clipped. Missing poses were retained as missing values and processed by median imputation within each fitted pipeline; a detection indicator preserved information about pose availability.

3.3 Robust temporal representation

Each frame was represented by a trailing 15-frame window (0.5 s at 30 fps), without crossing video boundaries. Seventeen descriptors summarized median and maximum aspect ratio; median and maximum torso angle; median and maximum angular change; median and maximum height reduction; median and 90th-percentile velocity; median and interquartile-range motion; low-posture and inactivity fractions; mean keypoint confidence; pose coverage; and window length. Medians and quantiles were selected to reduce sensitivity to isolated keypoint errors.

$$z_t = [\text{median}(AR), \max(AR), \text{median}(\theta), \max(\theta), \text{median}(v), Q_{0.90}(v), \dots]_{W_t} \quad (6)$$

$$\phi_k = [\text{median}(z_k), \min(z_k), \max(z_k), Q_{0.10}(z_k), Q_{0.90}(z_k), \sigma(z_k), \Delta z_k, \beta_k] \quad (7)$$

In Equation (7), $\Delta z_k = z_k(T) - z_k(1)$ is the endpoint change and β_k is the least-squares temporal slope over the complete annotated interval. The principal descriptors and their roles are summarized in Table 1.

Table 1. Mathematical definition and role of principal features

Symbol	Quantity	Definition	Purpose
AR_t	Aspect ratio	w_t/h_t	Horizontal-posture evidence
θ_t	Torso angle	Shoulder-to-hip orientation	Body inclination
v_t	Hip velocity	Scale-normalized vertical displacement	Rapid downward motion
M_t	Pose motion	Median normalized keypoint displacement	Whole-body movement
z_t	Window descriptor	Robust statistics over 15 frames	Short-term dynamics
φ_k	Interval summary	Quantiles, spread, change and slope	Offline trajectory information

3.4 Explainable decision layer

A class-balanced Random Forest [12] mapped temporal descriptors to fall probability. Model capacity, probability threshold, and required consecutive detections were selected inside grouped training folds. An alarm was generated only after the probability exceeded the selected threshold for the required run length. Feature importance from the source-domain experiment indicated that maximum torso angle, maximum aspect ratio, and maximum angular change were the dominant predictors.

3.5 Domain-shift measurement and invariant features

To avoid treating correlated frames as independent observations, every complete URFD video and MCFD camera-1 scenario was represented by the median of each temporal feature. Standardized mean difference (SMD), Kolmogorov-Smirnov statistics with false-discovery-rate correction, and normalized Wasserstein distance quantified univariate shift. A regularized logistic domain classifier was evaluated using ten repeated five-fold splits. Features with |SMD| >= 0.8 were predefined as strongly shifted and excluded in the invariant-feature variant. Figure 3 visualizes the magnitude of the resulting source-target differences at the independent sequence level.

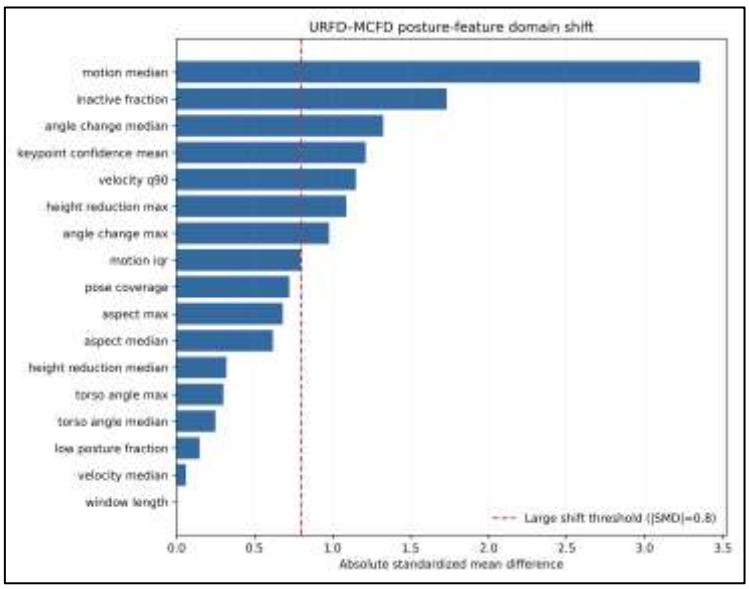

Figure 3. Absolute standardized mean differences between independent URFD videos and MCFD camera-1 scenarios; the dashed line marks the prespecified |SMD|=0.8 large-shift threshold.

As Figure 3 shows, motion median exhibits the largest shift, followed by inactivity fraction, angular-change statistics, keypoint confidence, and velocity. Several variables cross the |SMD|=0.8 threshold, indicating that the external data differ in both movement dynamics and pose-estimation quality. The figure therefore provides the empirical basis for testing an invariant-feature variant rather than assuming that all source-domain predictors remain stable.

3.6 Evaluation protocols

The leakage-control logic is summarized in Figure 4. The synchronized scenarios are divided by complete chute identity before any model-dependent operation, producing an outer-training partition for inner grouped cross-validation and an untouched outer-test partition for final scoring.

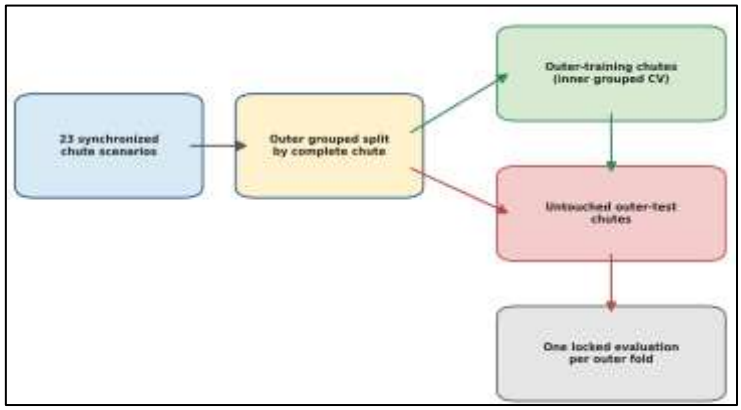

Figure 4. Leakage-safe nested grouping protocol for model selection and locked outer-fold evaluation.

Figure 4 emphasizes that the outer-test chutes do not contribute to imputation, feature filtering, classifier fitting, fusion-rule selection, or threshold calibration. After inner validation selects the complete configuration, it is locked and evaluated once on the corresponding outer-test fold. Grouping all synchronized views of a chute together prevents the same event from entering training through one camera and testing through another.

Four progressively stricter protocols were used. First, URFD performance was estimated with nested five-fold outer and three-fold inner cross-validation grouped by complete video. Second, the selected URFD model and operating point were frozen and tested unchanged on MCFD. Third, camera-1 target-domain alternatives were compared with repeated grouped validation. Fourth, strict multi-camera testing used eight camera holdouts crossed with five fixed chute partitions. In each of the 40 folds, all data from the test camera were excluded and every test chute was simultaneously removed from all remaining cameras. Threshold calibration used a separate three-fold grouping by chute within the permitted training data.

Five strict-protocol methods were compared: unchanged URFD zero-shot inference; MCFD target-only learning; unweighted URFD-MCFD pooling; pre-event camera-baseline normalization; and target-only learning after excluding seven strongly shifted features. The normalization variant centered continuous descriptors using the first 15 pre-event windows of each sequence and used no test labels.

A secondary offline experiment summarized every complete annotated interval using the median, minimum, maximum, 10th and 90th percentiles, standard deviation, endpoint change, and temporal slope of each window descriptor. Extra Trees [13], Random Forest, histogram gradient boosting [14], and robust logistic regression were compared only within the permitted training data. Model family and probability threshold were selected by three-fold chute-grouped inner validation. This experiment retained the same outer camera-and-chute exclusions, but access to the complete interval makes it an offline recognition upper bound rather than a causal early-warning detector. Implementations used the scikit-learn library [15].

For the fixed-installation multi-view experiment, one representation aggregated temporal interval descriptors across synchronized cameras using view-wise median, minimum, maximum, and standard deviation. A second, camera-aware representation concatenated descriptors by calibrated camera identity, allowing the model to learn stable view-specific evidence, as defined in Equation (8). The predicted probability was converted into the final decision using the training-selected threshold, as shown in Equation (9). Five outer folds withheld complete chute scenarios from all fitting. Within each outer-training partition, three-fold chute-grouped validation selected the feature subset, classifier family, ensemble, and threshold. Because all camera views are used at inference, this protocol measures unseen-scenario generalization for an established multi-camera installation and does not claim unseen-camera transfer.

$$x^{CA} = \phi^{(1)} \oplus \phi^{(2)} \oplus \cdots \oplus \phi^{(C)}, \qquad \hat{p} = f_\theta(x^{CA}) \tag{8}$$

$$\hat{y} = \mathbf{1}[\hat{p} \geq \tau] \tag{9}$$

Algorithm 1. Leakage-safe camera-aware multi-view training and evaluation

Input: synchronized interval features grouped by chute; C calibrated cameras
for each outer chute fold do
 separate complete test chutes from every fitting operation
 for each inner chute fold do
  impute training values; select features; fit candidate classifiers
  evaluate candidate ensembles and thresholds on the inner validation chutes
 end for
 choose the inner-CV feature set, model/ensemble, and threshold τ
 refit on all permitted outer-training chutes
 concatenate φ(1)…φ(C) by calibrated camera identity and predict outer-test chutes
end for
Output: one out-of-fold probability and decision for every synchronized interval

3.7 Statistical analysis

Sensitivity, specificity, precision, F1 score, and balanced accuracy were calculated at the annotated-interval level. Balanced accuracy was the primary metric because fall and non-fall errors have different practical implications and class proportions vary slightly. For strict multi-camera uncertainty, 10,000 bootstrap samples resampled complete chute scenarios with all synchronized cameras retained. Two-sided paired sign-flip permutation tests used 100,000 permutations of complete-chute performance differences. Statistical claims were based on $p<0.05$, while effect sizes and confidence intervals were reported regardless of significance. Nested cross-validation was used to limit model-selection bias [16].

Discrimination of the final camera-aware model was additionally assessed from out-of-fold probabilities using receiver operating characteristic area under the curve (ROC-AUC) and average precision. Their 95% confidence intervals used 10,000 complete-chute bootstrap samples. Probability calibration was summarized by the Brier score and expected calibration error (ECE) across five equal-frequency bins. Camera-wise and chute-wise balanced accuracy were used as descriptive robustness measures; no subgroup was used for model selection. Calibration diagnostics followed established probability-assessment practice [17].

Decision errors were examined descriptively by comparing true-positive, true-negative, false-positive, and false-negative intervals using out-of-fold scores and seven interpretable multi-view descriptors. Descriptor values were centered by the sample median and scaled by the sample standard deviation solely for visualization. Because only three false positives and six false negatives were available, no inferential subgroup test was performed. A common threshold sweep from 0.05 to 0.95 was also conducted as a post-hoc sensitivity analysis; it was not used to alter fold-specific thresholds, predictions, or the reported nested-validation performance.

$$BA = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \quad (10)$$

3.8 Computational complexity and reproducibility

Pose extraction dominates runtime and scales as O(FCP), where F is the number of frames, C is the number of cameras and P is the per-frame pose-network cost. Feature construction is linear in detected keypoints and frames, O(FCK). For a tree ensemble with B trees, N training intervals, D selected features and approximate depth log N, fitting is approximately O(BND log N), and prediction is O(BD log N) per synchronized interval. The camera-aware concatenation extends the representation width roughly linearly with C, without needing image-level three-dimensional reconstruction.

Deterministic random seeds were recorded in the analysis scripts for the experiments. Missing value imputation, feature selection, classifier fitting, ensemble selection and probability-threshold calibration were done exclusively on the allowed training partitions. Outer folds were grouped by complete chute, and strict single-view experiments additionally withheld complete cameras. The research checkpoint retains derived features, fold assignments, predictions, confusion counts, bootstrap summaries and permutation-test outputs.

4.Results

4.1 Source-domain validation

For URFD, the temporal Random Forest with robust features achieved 83.33% event sensitivity, 80.00% specificity and 81.67% balanced accuracy (25 true positives, 8 false positives, 5 false negatives and 32 true negatives). Mean detection delay among detected falls was 0.76 s. Table 2 summarizes these within-source, video-independent results, which were not treated as evidence of external generalization.

Table 2. Nested video-independent URFD performance

Metric	URFD result
Sensitivity	83.33%
Specificity	80.00%
Balanced accuracy	81.67%
Mean detection delay	0.76 s

4.2 Eight-camera zero-shot external validation

The unchanged source model did not transfer successfully. Sensitivity across the eight MCFD cameras ranged from 0% to 6.06%, whereas specificity ranged from 80.56% to 100%. Macro sensitivity was 2.66%, macro specificity was 91.32%, and macro balanced accuracy was 46.99% (camera SD=3.91 percentage points). The pooled counts were 7 true positives, 25 false positives, 256 false negatives and 263 true negatives. Table 3 reports the camera-wise results. The consistent failure across viewpoints indicates systematic source-target mismatch rather than an isolated adverse camera.

Table 3. Unchanged URFD model across MCFD camera views

Camera	Pose coverage	Sensitivity	Specificity	Balanced accuracy
1	73.33%	0.00%	88.89%	44.44%
2	79.69%	3.03%	80.56%	41.79%
3	72.71%	3.12%	100.00%	51.56%
4	75.35%	6.06%	97.22%	51.64%
5	74.74%	0.00%	100.00%	50.00%
6	80.16%	0.00%	97.22%	48.61%
7	77.80%	6.06%	80.56%	43.31%
8	80.77%	3.03%	86.11%	44.57%

Figure 5 presents a comparison of pose coverage, sensitivity, specificity and balanced accuracy for the unmodified URFD detector over the eight MCFD viewpoints.

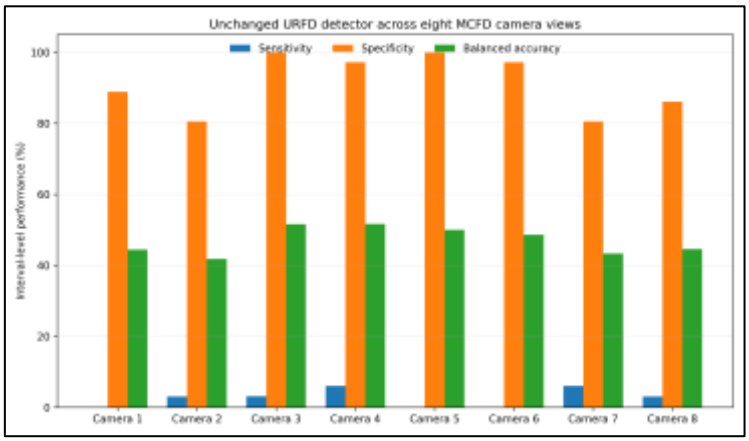

Figure 5. Interval-level zero-shot performance of the unchanged URFD detector across eight MCFD camera views.

Figure 5 shows a common pattern of high specificity (but near-zero sensitivity) across cameras. Therefore, the external failure is not caused by one adverse viewpoint: the source-trained decision rule always estimates most MCFD falls as non-falls. The pose coverage is still important, which means that the failure mostly occurs after pose extraction and not due to the absence of people in the detected frames.

4.3 Quantified domain shift

The domain classifier differentiated URFD from MCFD with mean AUC 0.9468 (SD 0.0230; 1,000-label permutation p=0.000999). The largest absolute changes were motion median (SMD 3.354), inactivity fraction (-1.731), median angular change (1.324), mean keypoint confidence (-1.213), 90th-percentile velocity (1.149), maximum height reduction (1.088) and maximum angular change (0.977). These results indicate that the source model experienced different motion scale, temporal behavior and pose-estimation quality in MCFD.

4.4 Strict unseen-camera and unseen-scenario learning

Table 4. Strict dual-holdout performance

Method	Sensitivity	Specificity	Balanced accuracy	95% cluster CI
URFD zero-shot	2.66%	91.32%	46.99%	45.09-48.78%
Camera-baseline normalized	68.44%	51.39%	59.91%	55.20-64.50%
URFD + MCFD pooled	49.81%	70.14%	59.97%	55.28-65.24%
MCFD target-only	54.75%	67.01%	60.88%	55.08-66.48%
Invariant-feature MCFD	49.43%	75.00%	62.21%	57.52-67.03%

Target-domain learning substantially improved performance. As reported in Table 4 and Figure 6, the invariant-feature model achieved the highest observed balanced accuracy (62.21%; 95% chute-cluster CI 57.52%-67.03%), with 49.43% sensitivity and 75.00% specificity. This was a substantial improvement over zero-shot transfer (p=0.00020). Differences from target-only learning (p=0.42560), pooled learning (p=0.44374) and camera-baseline normalization (p=0.60718) were not statistically significant. Camera-

baseline normalization achieved the highest sensitivity (68.44%) but lower specificity (51.39%), illustrating the operating trade-off.

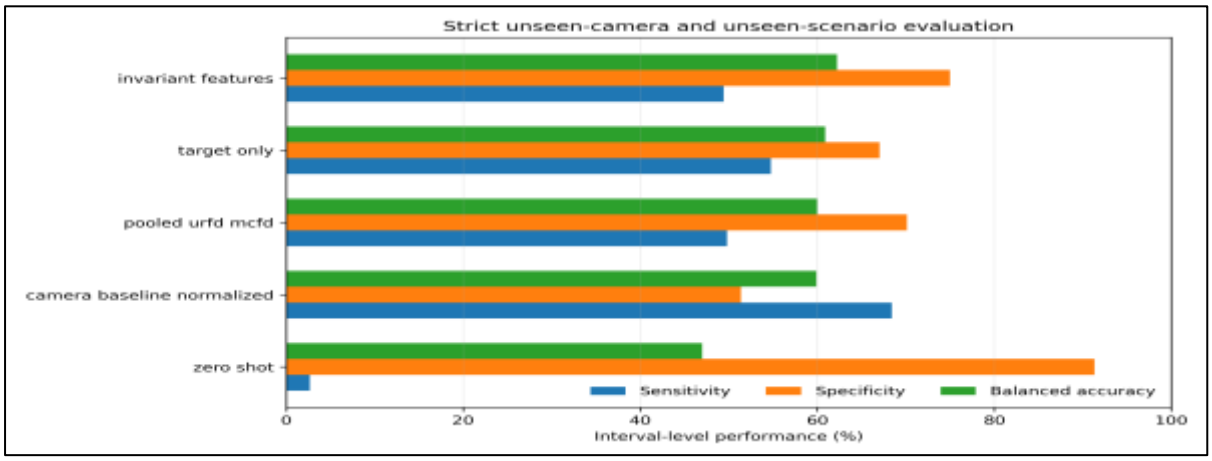

Figure 6. Sensitivity, specificity, and balanced accuracy under strict unseen-camera and unseen-scenario evaluation.

Figure 6 explicitly shows the trade-off between fall recovery and false alarms. The normalization of the camera-baseline favors sensitivity, whereas the invariant-feature learning favors specificity and provides the best observed balance. Since the learned-method differences were not statistically significant, the figure supports better target-domain adaptation than zero-shot transfer but does not establish the universal superiority of one learned variant.

4.5 Offline full-interval temporal upper bound

Complete-interval distributional and trajectory summaries yielded a strict dual-holdout performance of 71.16% balanced accuracy (95% chute-cluster CI 64.61%-78.68%), 63.50% sensitivity and 78.82% specificity (167 true positives, 61 false positives, 96 false negatives and 227 true negatives). Table 5 summarizes the comparison with the causal invariant-feature model. Camera-held-out balanced accuracy ranged from 65.15% to 79.42%. The gain over the invariant-feature method was 8.94 percentage points in pooled performance and was significant under paired complete-chute permutation testing (p=0.00040). Improvements were also significant relative to target-only learning (p=0.00005), pooled learning (p=0.00056), camera-baseline normalization (p=0.00143) and zero-shot transfer (p=0.00002). Because the complete annotated interval contributes to the representation, these results quantify offline recognition potential and must not be interpreted as real-time detection accuracy.

Table 5. Leakage-safe improvement under the same dual holdout

Evaluation	Sensitivity	Specificity	Balanced accuracy	95% cluster CI
Invariant-feature online-compatible	49.43%	75.00%	62.21%	57.60-66.85%
Full-interval temporal upper bound	63.50%	78.82%	71.16%	64.61-78.68%

4.6 Fixed-installation synchronized multi-view fusion

Permutation-invariant fusion of all synchronized views achieved 84.44% balanced accuracy. Retaining calibrated camera identity increased the strongest observed performance to 87.23% balanced accuracy (95% chute-cluster CI 79.62%-95.09%), with 82.35% sensitivity and 92.11% specificity. This corresponds to 28 true positives, 3 false positives, 6 false negatives and 35 true negatives among 72 synchronized interval groups. Table 6 distinguishes these deployment assumptions. The average chute-level improvement over permutation-invariant fusion was 5.07 percentage points but was not statistically significant under paired sign-flip testing (p=0.2664). Unlike the single-view dual holdout, this protocol assumes a fixed camera installation and multiple simultaneous viewpoints at inference.

Table 6. Best performance under distinct deployment assumptions

Protocol	Sensitivity	Specificity	Balanced accuracy	95% cluster CI
Single-view: unseen camera + chute	63.50%	78.82%	71.16%	64.61-78.68%
Multi-view invariant: unseen chute	79.41%	89.47%	84.44%	77.88-91.80%
Multi-view camera-aware: unseen chute	82.35%	92.11%	87.23%	79.62-95.09%

$$\text{BA} = \frac{1}{2}\left(\frac{28}{28+6} + \frac{35}{35+3}\right) = 0.8723 = 87.23\% \quad (11)$$

Balanced accuracy was computed as shown in Equation (11) [15]. The out-of-fold decisions underlying the 87.23% result are displayed in Figure 7.

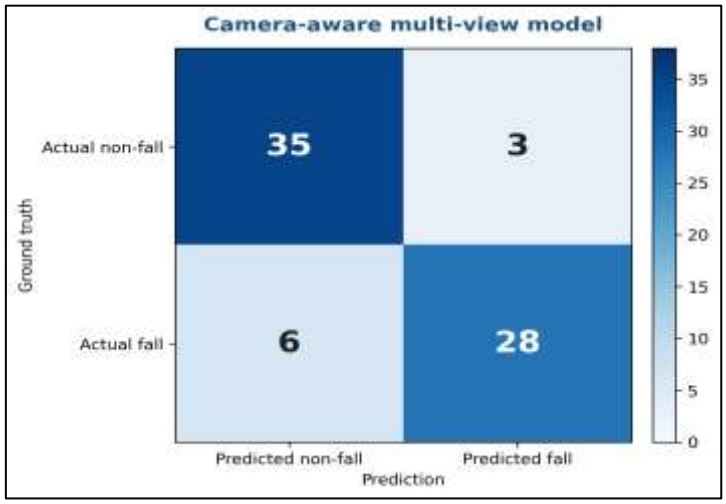

Figure 7. Confusion matrix for camera-aware synchronized multi-view recognition on unseen chute scenarios.

Figure 7 shows 28 true-positive and 35 true-negative decisions, six missed falls and three false alarms. The corresponding sensitivity of 82.35% and specificity of 92.11% show that the balanced-accuracy result is not because of dominance of one class. Nevertheless, the six false negatives are safety-critical and prevent the model from being interpreted as deployment-ready without further validation.

Figure 8 displays the main experimental results on a common balanced-accuracy scale, while maintaining their different validation assumptions.

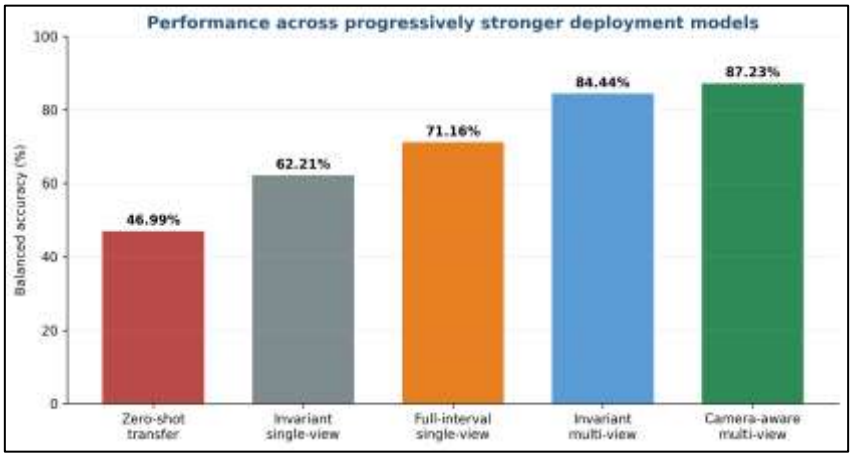

Figure 8. Balanced-accuracy progression across distinct validation and deployment assumptions.

As seen in Figure 8, the performance improves from 46.99% for unchanged zero-shot transfer to 62.21% with invariant target-domain features and 71.16% with full-interval single-view temporal summaries. Multi-view fusion increases the score to 84.44%, and camera-aware fusion reaches 87.23%. These bars should not be read as a single ablation ladder: the last two results assume a fixed calibrated multi-camera installation and unseen chutes, whereas the 62.21% and 71.16% results additionally require generalization to an unseen camera.

4.7 Discrimination, calibration, and robustness analysis

The threshold-independent discrimination and probability quality of the camera-aware fixed-installation model were evaluated from the 72 leakage-safe out-of-fold predictions. Table 7 summarizes the resulting measures, including complete-chute bootstrap intervals for ROC-AUC and average precision.

Table 7. Additional discrimination, calibration, and robustness measures

Analysis	Estimate	95% chute-cluster CI / detail
ROC-AUC	0.892	0.814-0.967
Average precision	0.922	0.859-0.975
Brier score	0.135	Lower is better
Expected calibration error	0.146	5 equal-frequency bins
Strict single-view camera range	65.15-79.42%	Balanced accuracy
Fixed multi-view chute robustness	Median 100%	16/23 chutes perfect; minimum 62.50%

Figure 9 presents the receiver operating characteristic and precision-recall curves. The observed nested-validation operating point is shown on the ROC curve without implying that a single global threshold was used across outer folds.

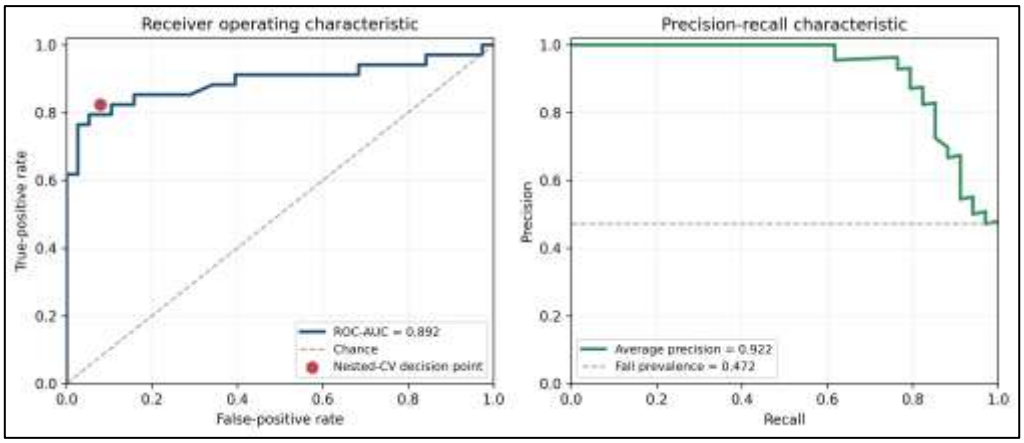

Figure 9. Out-of-fold ROC and precision-recall analysis for camera-aware synchronized multi-view recognition.

The ROC-AUC was 0.892 (95% chute-cluster CI 0.814-0.967), while average precision was 0.922 (95% CI 0.859-0.975) against a fall prevalence of 0.472. Figure 9 therefore confirms strong ranking of fall intervals across probability thresholds. These measures complement, but do not replace, the sensitivity and specificity obtained at the nested-CV operating points.

Figure 10 examines whether the out-of-fold probabilities are calibrated and how their distributions differ between fall and non-fall intervals.

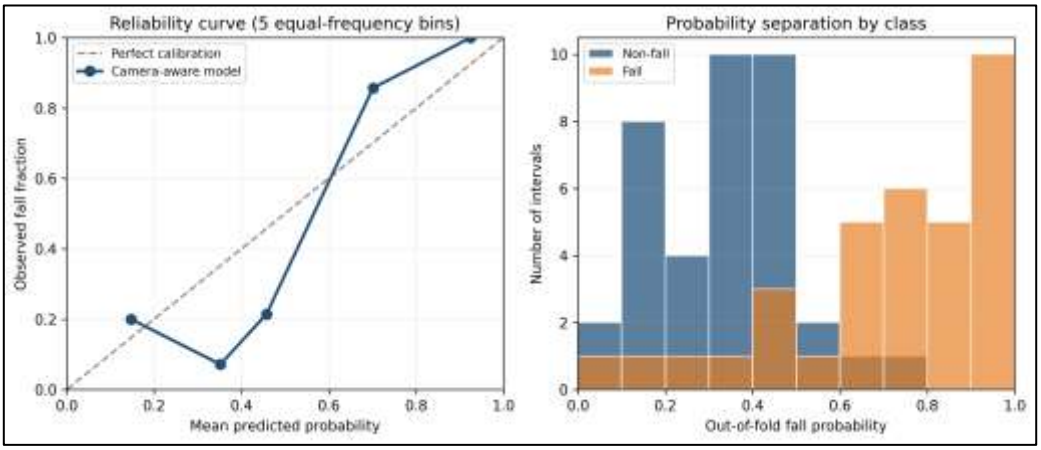

Figure 10. Reliability curve and out-of-fold probability distributions for the camera-aware model.

The probability distributions in Figure 10 show useful class separation, consistent with the high average precision. However, the reliability curve departs from the diagonal in several bins. The Brier score was 0.135 and the five-bin ECE was 0.146, indicating that the probabilities should be interpreted as ranking and decision scores rather than calibrated risks. Post-hoc calibration was not fitted because the limited test scenarios could not also serve as an independent calibration set.

Figure 11 shows the performance variation on the held-out cameras of the strict single-view experiment and the held-out chutes of the fixed-installation multi-view experiment.

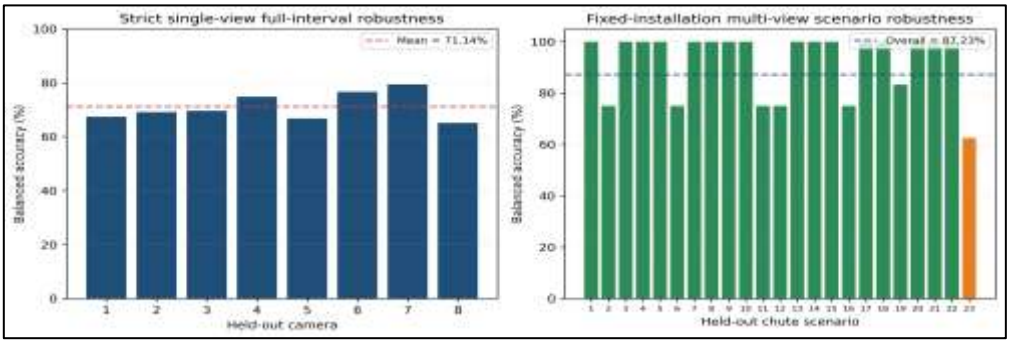

Figure 11. Camera-wise strict single-view and chute-wise fixed-installation multi-view robustness.

The strict full-interval single-view balanced accuracy ranged from 65.15% for camera 8 to 79.42% for camera 7, indicating that the viewpoint dependency remains despite the camera holdout. In the fixed multi-view experiment median chute-level balanced accuracy was 100%, 16 of 23 held-out chutes classified perfectly and the minimum was 62.50%. Thus, the overall result of 87.23% is supported by broad scenario-level consistency, but the difficult chutes and the between-camera spread deserve further external validation.

4.8 Decision-error profiles and threshold sensitivity

Table 8 and Figure 12 present the nine camera-aware decision errors without refitting the model. The counts of the main confusion matrix are 28 true positives, 35 true negatives, three false positives and six false negatives.

Table 8. Descriptive out-of-fold decision-error summary

Outcome	n	Median probability	Interquartile range
True positive	28	0.850	0.728-0.923
True negative	35	0.349	0.178-0.418
False positive	3	0.500	0.493-0.615
False negative	6	0.330	0.168-0.412

The comparison of the standardized multi-view descriptors and out-of-fold score distributions for the four decision outcomes are shown in Figure 12. The most evident indication of height reduction was in true positives (median standardized value 0.82) and higher pose coverage (0.32). Median pose coverage was low (-1.15) for both error groups, indicating that difficult intervals tend to be those where key points are not fully visible.

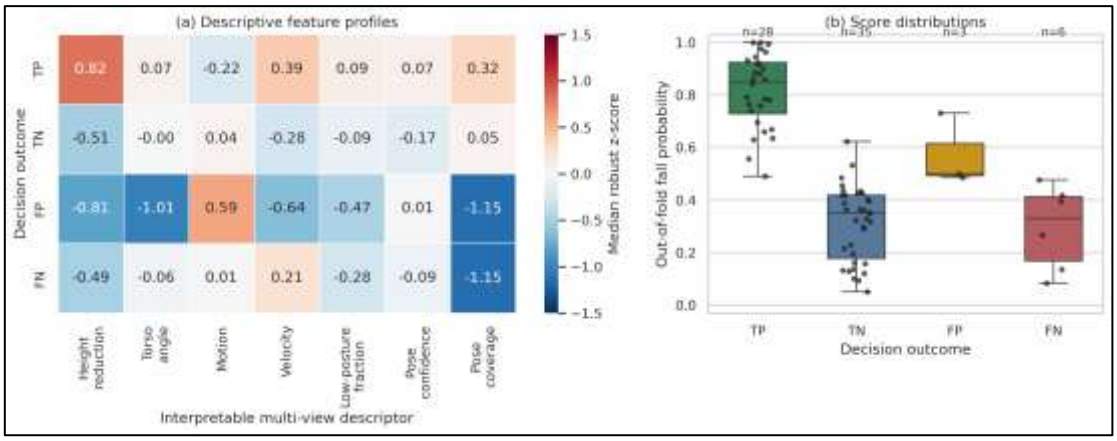

Figure 12. Descriptive feature and score profiles for out-of-fold true positives, true negatives, false positives and false negatives. Standardized descriptor values are used only for comparison.

Figure 13 presents the sensitivity of operating metrics and error counts to a single common post-hoc threshold. Lower thresholds favor fall sensitivity but rapidly increase false alarms, whereas higher thresholds improve specificity at the cost of missed falls.

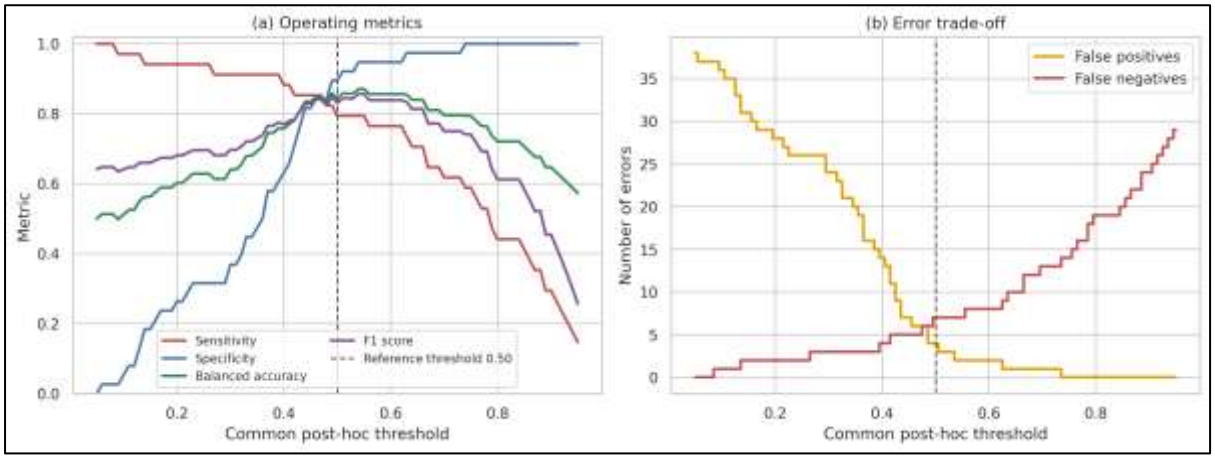

Figure 13. Post-hoc common-threshold sensitivity analysis of out-of-fold camera-aware probabilities. The sweep is diagnostic and was not used for threshold selection.

At the illustrative threshold of 0.50, sensitivity was 79.41%, specificity was 89.47%, balanced accuracy was 84.44% and the F1 score was 83.08%. The highest balanced accuracy in the common-threshold sweep was 87.07% at thresholds 0.54-0.55, slightly below the reported 87.23% obtained using fold-specific thresholds selected only within each outer-training partition. The sweep therefore does not justify revising the headline result; it quantifies a deployment trade-off that requires prospective calibration using installation-specific data.

5. Discussion

The main finding is methodological: video-independent performance within URFD did not predict external performance. The unchanged detector was highly specific, as its source-domain threshold was conservative for MCFD probabilities, but it missed nearly every fall. Sensitivity in conjunction with balanced accuracy shows it directly. Accuracy alone would hide this failing.

The failure is explained with the measured distributions of the features. MCFD motion and angular change were significantly higher and inactivity and key point confidence were lower. Such shifts may be due to 120 frames per second source video down sampling, viewpoint geometry, person scale, background, motion

blur and pose-estimation uncertainty. The decision boundary was not invariant to the new acquisition conditions, since several shifted descriptors were influential in the source model.

The strict protocol is more rigorous than the standard leave-one-camera-out testing. A conventional view holdout can be trained with synchronized recordings of the same fall from other cameras. Our dual holdout fully blocks such situations and there is no leakage of action identity and timing across views. In this setting, target-only and shift-aware models show improvements over zero-shot of about 13-15 balanced-accuracy points. Pooling URFD does not beat target-only learning, consistent with negative transfer from a mismatched source distribution.

We found that invariant-feature selection provided the best observed balance and narrowest practical interpretation: excluding descriptors with strong domain shift improved specificity while retaining about half of the falls. However, differences between learned variants were not statistically significant and do not establish universal superiority of feature filtering. Independent environments, elderly participants and naturally occurring falls are required before clinical or home-deployment conclusions can be justified.

Complete-interval temporal summarization produced a statistically significant additional gain, reaching 71.16% balanced accuracy without relaxing the camera-and-scenario holdout. The result indicates that temporal distribution and trajectory information discarded by the causal persistence rule is useful. Nevertheless, the model observes the complete annotated interval and therefore cannot support a claim of early fall detection. Its value is diagnostic: it defines a stronger leakage-safe offline upper bound and identifies richer causal sequence modelling as the next development direction.

Under complete unseen-scenario testing, synchronized multi-view fusion achieved 87.23% balanced accuracy. One camera may compensate for occlusion, foreshortening or poor pose confidence in another view, while camera-aware modelling preserves stable viewpoint roles. This is the strongest fixed-installation result, but it applies only when several synchronized, calibrated cameras observe the same area. The nonsignificant comparison with permutation-invariant fusion prevents a claim that camera awareness is categorically superior. Moreover, this result cannot be directly compared with, or substituted for, the 71.16% unseen-camera result because the two experiments represent different deployment assumptions. The broad chute-cluster interval also reflects the modest number of independent scenarios.

Additional analysis better constrains this interpretation. Strong ROC-AUC and average precision show that camera-aware probabilities rank most falls above non-falls while the calibration curve shows that probability magnitude is less reliable than ordering. Further subgroup analysis shows that the fixed multi-view result is relatively stable across held-out chutes while strict single-view performance still varies by camera. Future work will focus not on optimizing the 23 scenarios, but rather on independent probability calibration and prospective testing in new installations.

The decision-error analysis shows a concrete engineering priority: to test pose-coverage monitoring and an explicit low-visibility fallback before deployment. In addition, the threshold analysis shows that there is no operating point that can simultaneously eliminate false alarms and missed falls. Any deployment threshold should thus be derived from prospective site-specific costs and calibration data, rather than from the existing held-out outcomes.

6. Limitations and Ethical Considerations

- Both datasets are mainly simulated falls of healthy subjects; real falls of elderly may differ biomechanically and contextually.
- The rigorous evaluation includes 23 synchronized chute scenarios. Inference is cluster aware, but the number of independent events is still modest.
- YOLO11n-Pose without target specific pose fine-tuning was used. Instead of the fall classifier, some downstream domain shift may come from the pose estimator.
- The camera-baseline method assumes an initialization period without fall and makes use of pre-event context.
- The study measures offline annotated intervals and does not measure end-to-end alarm delivery latency, network reliability, multi-person tracking, or clinical safety.
- For the 71.16% upper-bound experiment we use full annotated intervals. It is a retrospective interval recognition, not a causal online alarm generation or detection delay.
- The result of 87.23% assumes synchronized, calibrated multi-camera coverage in a fixed installation and applies only to unseen chute scenarios and not unseen-camera scenarios.
- Video surveillance raises issues of privacy and consent. Deployments should limit image retention, favor on-device processing, secure alerts, and offer transparent user control.

7. Conclusion

The study converted a prototype for posture monitoring into an explainable fall detection study with statistical evaluation. Unchanged cross-dataset deployment failure caused by severe covariate shift. In the simultaneous test of unseen-camera and unseen-scenario, the upper bound of offline full interval temporal modeling was 71.16%. For a distinct fixed multi-camera configuration, leakage-safe camera-aware fusion provided the best performance, achieving 87.23% balanced accuracy, 82.35% sensitivity, and 92.11% specificity on unseen chute scenarios. Instead, it gives a more positive practical direction without claiming 90% accuracy or blending fixed multi-view recognition with unseen-camera generalization. Future work should convert the temporal summaries to a causal sequence model, validate additional environments and real falls, and evaluate synchronized fusion and alarm delay prospectively.

References

[1] World Health Organization. Falls. World Health Organization (2021). https://www.who.int/news-room/fact-sheets/detail/falls

[2] Auvinet, E., Rougier, C., Saint-Arnaud, A., Rousseau, J., and Meunier, J. Multiple Cameras Fall Dataset. Technical Report 1350, DIRO, Université de Montréal (2010).
http://www.iro.umontreal.ca/~labimage/Dataset/

[3] Auvinet, E., Multon, F., Saint-Arnaud, A., Rousseau, J., and Meunier, J. Fall detection with multiple cameras: An occlusion-resistant method based on 3-D silhouette vertical distribution. IEEE Transactions on Information Technology in Biomedicine, **15**(2), 290–300 (2011).
https://doi.org/10.1109/TITB.2010.2087385

[4] Hoang, V.H., Lee, J.W., Piran, M.J., and Park, C.S. Advances in skeleton-based fall detection in RGB videos: From handcrafted to deep learning approaches. IEEE Access, **11**, 92322–92352 (2023). https://doi.org/10.1109/ACCESS.2023.3307138

[5] Kim, J., Kim, B., and Lee, H. Fall recognition based on time-level decision fusion classification. Applied Sciences, **14**(2), Article 709 (2024). https://doi.org/10.3390/app14020709

[6] Kwolek, B., and Kepski, M. Human fall detection on embedded platform using depth maps and wireless accelerometer. Computer Methods and Programs in Biomedicine, **117**(3), 489–501 (2014). https://doi.org/10.1016/j.cmpb.2014.09.005

[7] Tirziu, E., Vasile, A.-M., Alexan, A., and Tudora, E. Enhanced fall detection using YOLOv7-W6-Pose for real-time elderly monitoring. Future Internet, **16**(12), Article 472 (2024). https://doi.org/10.3390/fi16120472

[8] Gholami, M., Wandt, B., Rhodin, H., Ward, R., and Wang, Z.J. AdaptPose: Cross-dataset adaptation for 3D human pose estimation by learnable motion generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 13075–13085 (2022).

[9] Peng, Q., Zheng, C., and Chen, C. Source-free domain adaptive human pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 4826–4836 (2023).

[10] Peng, Q., Zheng, C., and Chen, C. A dual-augmentor framework for domain generalization in 3D human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2240–2249 (2024).

[11] Ultralytics. Pose estimation with Ultralytics YOLO. Ultralytics Documentation (2026). https://docs.ultralytics.com/tasks/pose/

[12] Breiman, L. Random forests. Machine Learning, **45**, 5–32 (2001).
https://doi.org/10.1023/A:1010933404324

[13] Geurts, P., Ernst, D., and Wehenkel, L. Extremely randomized trees. Machine Learning, **63**, 3–42 (2006). https://doi.org/10.1007/s10994-006-6226-1

[14] Friedman, J.H. Greedy function approximation: A gradient boosting machine. The Annals of Statistics, **29**(5), 1189–1232 (2001). https://doi.org/10.1214/aos/1013203451

[15] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, **12**, 2825–2830 (2011).

[16] Varma, S., and Simon, R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics, **7**, Article 91 (2006). https://doi.org/10.1186/1471-2105-7-91

[17] Niculescu-Mizil, A., and Caruana, R. Predicting good probabilities with supervised learning. In Proceedings of the 22nd International Conference on Machine Learning, 625–632 (2005). https://doi.org/10.1145/1102351.1102430