A Unified Fall Detection Benchmark and Model for Early, Event-Centric Evaluation

Jerry Liu, Jaelyn Liang, Nidhi Seethapathi

Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139 USA
jerryl07@mit.edu

Abstract

Falls are a leading cause of injury in older adults, and early detection remains difficult in real-world conditions. This paper introduces a unified fall-detection benchmark and TriFallNet, a transformer model for pose-based sensing. The benchmark standardizes video preprocessing to 25 Hz at 256×256 RGB and harmonizes labels to a common fall/ADL convention, enabling reproducible, comparable evaluation across datasets without per-dataset retuning. A single operating point is chosen on validation and carried through to testing to reflect deployment practice. TriFallNet operates on pose-derived kinematics, including joint trajectories, center-of-mass velocity, and torso inclination extracted with MotionBERT, and uses a Skeleton-MixFormer backbone to model temporal structure. Together, the benchmark and TriFallNet close the evaluation gap by making results reproducible, yielding directly comparable evidence for early fall detection in practice.

Introduction

Falls among older adults impose a severe public health and economic burden. In the United States, unintentional falls are the leading cause of injury and injury-related death for adults over 65 (Centers for Disease Control and Prevention 2024b; Kakara et al. 2023). Over a quarter of older Americans (27.6%) fall each year (Centers for Disease Control and Prevention 2024b; Kakara et al. 2023), resulting in over 3 million emergency visits and roughly $50 billion in annual medical costs (Centers for Disease Control and Prevention 2024a). Alarmingly, in 2021 an average of about 100 older adults died each day from falls (Centers for Disease Control and Prevention 2024b; National Center for Injury Prevention and Control (CDC) 2017). Early (pre-impact) detection of falls could enable protective interventions and faster medical response, reducing injuries in elderly care settings.

Existing machine learning fall detectors typically focus on binary post-fall classification and often perform poorly in realistic conditions. Vision-based methods (e.g., CNNs or silhouette analysis) can work in controlled environments, but suffer under occlusion, cluttered backgrounds, and varying illumination (Gutiérrez et al. 2021; Alam et al. 2022). Likewise, single-sensor systems (either wearable or vision) yield high false alarms and degraded accuracy in suboptimal lighting or multi-person scenes (Gutiérrez et al. 2021; Alam et al. 2022). Moreover, prior works generally ignore pre-fall cues and fall context: most rely on single-task classifiers that only predict "fall vs. no-fall" after impact (Gutiérrez et al. 2021; Alam et al. 2022), without inferring cause or likely severity.

Another barrier is the lack of standardized evaluation. Public fall datasets (Le2i, UR-Fall, UP-Fall, MultiCamFall, etc.) differ widely in format and scale (Charfi et al. 2013a; Kwolek and Kepski 2014; Martínez-Villaseñor et al. 2019; Auvinet et al. 2010a). For example, the Le2i dataset contains 143 falls and 79 ADLs recorded at 640×480@25 Hz across six camera views, whereas UR-Fall has 30 falls and 40 ADLs captured with dual Kinect sensors, and UP-Fall provides 17 subjects with multimodal data (two RGB cameras at 640×480@18 Hz plus wearable sensors) (Charfi et al. 2013a; Kwolek and Kepski 2014; Martínez-Villaseñor et al. 2019). These datasets vary in resolution, frame rate, viewpoints, and label conventions, and prior studies report results with inconsistent metrics (frame-level accuracy vs. event-level detection) and ad hoc train/test splits. This fragmentation prevents fair comparison of algorithms across settings (Gutiérrez et al. 2021; Alam et al. 2022; Schneider et al. 2025).

To address these gaps, we propose TriFallNet and a unified evaluation benchmark. TriFallNet is a novel spatio-temporal transformer that ingests 3D skeletal joints (via MotionBERT) plus additional kinematic features, enabling early fall detection (before impact) along with multi-task outputs for fall cause and injury severity (Zhu et al. 2023). The model processes joint trajectories, center-of-mass dynamics, and torso inclination with a Skeleton-Mixformer backbone to capture long-range spatial and temporal dependencies (Xin et al. 2023). In parallel, we introduce a standardized fall benchmark that consolidates multiple staged datasets and introduces consistent event-centric metrics and cross-dataset splits (Charfi et al. 2013a; Kwolek and Kepski 2014; Martínez-Villaseñor et al. 2019; Auvinet et al. 2010a; Alam et al. 2024a; Schneider et al. 2025). Together, these contributions aim to improve robustness and reproducibility in fall detection research while enabling comparisons that transfer across datasets and settings (Gutiérrez et al. 2021; Schneider et al. 2025; Núñez-Marcos and Arganda-Carreras 2024).

Related Work

Vision-Based Fall Detection

Traditional computer vision fall detectors use RGB or depth cameras to analyze body motion. Early methods relied on background subtraction or silhouette analysis with handcrafted features (e.g., bounding-box height/width) to distinguish falls (Charfi et al. 2013a). More recently, deep neural networks have been applied, including convolutional baselines and one-stage person detectors, but these approaches still suffer from occlusions and sensor noise in practical environments (Gutiérrez et al. 2021; Alam et al. 2022). Transformers and hybrid CNN–transformer models have begun to appear; for example, recent work applies transformer backbones to sliding-window RGB clips on staged datasets, improving efficiency but still emphasizing post-impact decisions over explicit pre-impact alerting (Núñez-Marcos and Arganda-Carreras 2024).

Pose-Based Fall Detection

An alternative is to first estimate human pose and then detect falls from the resulting keypoint sequences. Pose-based methods isolate the person from background, improving robustness to environmental clutter (Gutiérrez et al. 2021). Many pipelines use 2D or 3D keypoints followed by temporal models such as RNNs, GCNs, or transformers to classify falls. Advances in motion representation learning, such as MotionBERT, provide robust 3D features that capture geometric and kinematic regularities across time (Zhu et al. 2023). Skeleton attention architectures, including Skeleton-Mixformer, strengthen long-range temporal reasoning and joint-to-joint interactions that align with early fall cues like loss of postural control and abrupt center-of-mass shifts (Xin et al. 2023). Nonetheless, most pose-based systems still emphasize post-impact recognition, and lead-time evaluation remains inconsistently reported (Gutiérrez et al. 2021; Alam et al. 2022).

Public Fall Datasets

Research has relied on several staged video datasets. The Le2i Fall Dataset contains falls and non-fall actions recorded at 640×480 and 25 fps across multiple viewpoints and rooms; the UR-Fall Dataset provides 30 falls and 40 ADL sequences captured by two Kinect depth cameras with synchronized accelerometry; the UP-Fall Detection Dataset includes 17 subjects performing activities and falls with two cameras at 18 fps plus wearable and ambient sensors; and MultiCamFall offers eight fixed cameras capturing falls interleaved with confounding ADLs (Charfi et al. 2013a; Kwolek and Kepski 2014; Martínez-Villaseñor et al. 2019; Auvinet et al. 2010a). GMDCSA-24 contributes additional indoor scenarios (Alam et al. 2024a). OmniFall unifies eight staged datasets under a shared taxonomy and adds a staged-to-wild protocol by curating in-the-wild accident footage, but it inherits variability in original capture specs and focuses primarily on taxonomy alignment and generalization analysis (Schneider et al. 2025). Across sources, differences in frame rate, resolution, camera geometry, and annotation style persist, and many datasets lack standardized event timing markers required for pre-impact lead-time measurement at scale.

Evaluation Protocols and Benchmarking

Prior works report results using various metrics and protocols. Some use frame-level accuracy or clip-level F1, while others consider event-level counts. There is no universally accepted train/test split: studies use random folds, cross-subject partitions, or custom scenarios, which complicates comparisons and weakens reproducibility (Gutiérrez et al. 2021; Alam et al. 2022). Consolidation efforts provide common taxonomies and staged-to-wild benchmarking (Schneider et al. 2025), yet two gaps remain: consistent preprocessing that normalizes frame rate and spatial resolution across datasets, and event-centric protocols that quantify both correctness and lead time under constrained false-alarm rates.

Comparison with Prior Work

TriFallNet differs from vision-only transformers by operating on pose-derived kinematics that emphasize human dynamics critical for pre-impact alerts (Zhu et al. 2023; Xin et al. 2023; Núñez-Marcos and Arganda-Carreras 2024). Relative to existing pose-based systems, it uses a Skeleton-Mixformer backbone to model long-range temporal structure and joint interactions that align with anticipatory cues (Xin et al. 2023). On the evaluation side, the proposed unified benchmark complements recent consolidation by enforcing consistent video normalization and label harmonization and by foregrounding pre-impact, event-centric metrics with cross-subject and cross-dataset splits. This design supports transparent, apples-to-apples comparisons across varied recording conditions while directly measuring the early-warning capability required for deployment (Gutiérrez et al. 2021; Schneider et al. 2025).

Methodology

Model: TriFallNet Overview

TriFallNet operates on pose-derived kinematics to emphasize human dynamics that precede impact. We extract per-frame 3D skeletal joints using MotionBERT and compute kinematic descriptors including center-of-mass (CoM) dynamics and torso inclination. These signals are concatenated and processed by a Skeleton-Mixformer backbone adapted for fall detection (Zhu et al. 2023; Xin et al. 2023). The model output a per-frame fall probability for early detection. We adopt selective fine-tuning of upper transformer blocks together with a lightweight framewise classification head. A single operating threshold is chosen on validation data and transferred unchanged to test sets for event analysis.

Inputs and Features

Let the 3D pose sequence be $\{\mathbf{J}_t \in R^{J\times 3}\}_{t=1}^{T}$ with $J = 17$ joints. Denote key joints $\mathbf{j}_t^{\text{pelvis}}$, $\mathbf{j}_t^{\text{lhip}}$, $\mathbf{j}_t^{\text{rhip}}$, $\mathbf{j}_t^{\text{neck}}$, $\mathbf{j}_t^{\text{head}}$, $\mathbf{j}_t^{\text{lfoot}}$, $\mathbf{j}_t^{\text{rfoot}}$. From these, we derive two kinematic features that complement raw trajectories.

Center of Mass and Velocity.

$$\mathbf{c}_t \;=\; \frac{1}{3}\big(\mathbf{j}_t^{\text{pelvis}} + \mathbf{j}_t^{\text{lhip}} + \mathbf{j}_t^{\text{rhip}}\big), \qquad v_t^{\text{CoM}} \;=\; \frac{\|\mathbf{c}_t - \mathbf{c}_{t-1}\|}{\Delta t}. \tag{1}$$

Torso Inclination.

$$\mathbf{v}_t^{\text{torso}} \;=\; \mathbf{j}_t^{\text{neck}} - \mathbf{c}_t, \tag{2}$$

$$\theta_t \;=\; \arccos\!\left(\frac{\mathbf{v}_t^{\text{torso}} \cdot \mathbf{e}_y}{\|\mathbf{v}_t^{\text{torso}}\|}\right), \quad \mathbf{e}_y = [0,\,1,\,0]^\top. \tag{3}$$

The model input is a fixed-length temporal window of 3D joints (17 joints, 3 coordinates each) concatenated with v_t^{CoM} and θ_t. Features are z-normalized per sequence and, optionally, per subject during training; normalization statistics are computed on training data and reused for validation and test.

Backbone and Heads

Skeleton-Mixformer provides spatial and temporal attention over joints and frames with multivariate topology mixing to capture joint-to-joint and long-range temporal interactions (Xin et al. 2023). For early fall detection, we attach a binary framewise head: a 1D convolution layer followed by a sigmoid that produces $p_t \in [0,1]$ per frame. To adapt efficiently, we freeze most layers and fine-tune the last two transformer blocks together with the Retrospect fusion module and the new head, preserving low-level motion representation while specializing high-level dynamics to fall detection.

Training and Inference

Training uses pooled clips from Le2i, UR-Fall, and GMDCSA-24 with cross-subject splits (70% train, 30% internal validation; subject identities are disjoint across splits). The loss is binary cross-entropy at the frame level. Optimization uses AdamW with separate learning rates for the new head (10^{-3}) and unfrozen transformer blocks (10^{-4}), weight decay 10^{-5}, batch size 64, up to 50 epochs, and early stopping on validation AUC (patience 10). We apply temporal augmentations (random temporal shifts and frame resampling) to improve generalization. During inference, per-frame probabilities are smoothed with a short temporal median to reduce flicker.

Frame-Level Metrics. Let $\hat{y}_t \in [0,1]$ be the predicted probability and $y_t \in \{0,1\}$ the frame label. For a threshold $\tau \in [0,1]$, set $\tilde{y}_t = 1[\hat{y}_t \geq \tau]$. We report accuracy, precision, recall, F1, and AUROC (threshold swept). The operating point $\tau^\star$ is selected on the validation set to maximize F1 (near the ROC top-left).

Event-Level Metrics and Lead Time. For external evaluation $\tau^\star$ is applied unchanged. With annotated impact frame t_{imp} and frame period Δt_{frame}, define

$$t_{\text{det}} \;=\; \min\{\, t \;:\; \hat{y}_t \geq \tau^\star \,\}, \qquad \Delta t \;=\; (t_{\text{imp}} - t_{\text{det}})\,\Delta t_{\text{frame}}. \tag{4}$$

We compute event recall within the window $[t_{\text{imp}} - 3\text{ s},\, t_{\text{imp}}]$ and average false alarms per clip (detections outside any event window), together with mean/median lead time Δt and summary AUROC/AUPRC when calibrated scores are available.

Baseline: Framewise SVM

As a classical baseline, we extract per-frame 2D joint coordinates and confidence scores using AlphaPose (51 features = 17×3) and train SVM classifiers (linear, polynomial, RBF) on Le2i with stratified sampling up to 50 frames per video and 5-fold cross-validation. External evaluation samples 2000 frames from UR-Fall and computes pre-impact event recall in the last 0.5 s before impact. We perform a grid search for hyperparameters and apply recursive feature elimination to assess joint importance.

Interpretability

We apply Integrated Gradients to the early fall detection head to produce joint- and time-resolved attributions. Attributions consistently highlight lower-limb joints (knees, ankles, hips) and the kinematic features (torso inclination and CoM velocity) as discriminative precursors of impact, providing biomechanical insight and guiding error analysis.

Unified Benchmark Overview

This benchmark establishes a unified, deployment-focused protocol for video-based fall detection. It addresses three long-standing impediments to comparability: (1) heterogeneous capture specifications (frame rate, resolution), (2) incompatible label conventions (clip-level tags versus event timing), and (3) ad hoc evaluation practices that confound threshold selection and reporting. Concretely, the framework (i) standardizes all inputs to a common temporal and spatial specification, (ii) harmonizes labels to a binary $\{\textit{fall}, \textit{ADL}\}$ decision while retaining event windows for early-detection analysis, and (iii) fixes dataset splits and metrics under a single interface. A manifest records per-clip provenance, original specifications, and applied transforms to ensure exact reproducibility and auditability.

Datasets and Inclusion

The benchmark uses staged, publicly available video datasets frequently adopted in fall detection: Le2i, UR Fall Detection (URFD), GMDCSA-24, and UP-Fall. Training and validation are performed on Le2i, URFD, and GMDCSA-24; UP-Fall is held out as an external test set, and the operating threshold selected on validation is transferred unchanged to all tests. To isolate the vision setting, evaluation relies solely on RGB video. When available, MultiCamFall is included only as auxiliary ADL footage and is excluded from fall-event training and threshold selection (Auvinet et al. 2010b).

Preprocessing and Video Normalization

All videos are resampled to 25 Hz and letterboxed to 256×256 while preserving aspect ratio. Normalized clips and their original specs are logged in a manifest. This simple standardization reduces variance due to capture hardware and keeps inference real-time on commodity GPUs.

Label Ontology and Temporal Alignment

The event ontology distinguishes ADL/non-fall from fall events centered at the impact instant. When reliable timing

Table 1: Datasets used in TriFallNet's benchmark. We train/validate on Le2i, URFD, and GMDCSA-24 and report external test results on UP-Fall. Evaluation is RGB-only with videos normalized to 25 Hz at 256×256.

Dataset	Modality	Views	Notes/Scale
Le2i (Charfi et al. 2013b)	RGB	Multi-room	staged falls/ADL
URFD (Kwolek and Kepski 2014)	RGB	Single	30 falls / 40 ADL; depth/IMU available
GMDCSA-24 (Alam et al. 2024b)	RGB	Single/Multi	160 videos
UP-Fall (Martínez-Villasenõr, Ponce, and et al. 2019)	RGB+wearables	Dual RGB	external test only
MultiCamFall (Auvinet et al. 2010b)	RGB	8 fixed	used as auxiliary ADL

metadata exists, we use it (e.g., URFD's per-frame CSVs can refine impact frames); otherwise, each fall clip receives a single window centered near the annotated impact after normalization. For binary training/evaluation we collapse to $\{fall, ADL\}$ but retain event windows to compute lead-time statistics.

Splits and Protocol

We adopt a cross-dataset protocol: pooled training and validation on Le2i+URFD+GMDCSA-24 with stratified 70/30 split and disjoint subjects per dataset; UP-Fall is held out for external testing. The operating threshold τ is selected on the internal validation split and then transferred *unchanged* to internal test and to UP-Fall for event analysis, mirroring our reported procedure. (Charfi et al. 2013b; Kwolek and Kepski 2014; Alam et al. 2024b; Martínez-Villasenõr, Ponce, and et al. 2019) This avoids threshold overfitting and quantifies distribution shift. *Auxiliary* ADL-only sets (e.g., MultiCamFall) are included only in training when present and never used to select τ.

Models and Training

TriFallNet ingests pose-derived kinematics with a transformer backbone; we keep that architecture unchanged. Training uses pooled clips from Le2i, URFD, and GMDCSA-24; validation AUC is used for early stopping; at inference we apply light temporal smoothing of scores. (Charfi et al. 2013b; Kwolek and Kepski 2014; Alam et al. 2024b) The final τ selected on validation is fixed for all test evaluations, including UP-Fall. (Martínez-Villasenõr, Ponce, and et al. 2019)

Metrics

We report clip-level Accuracy/Precision/Recall/F1 and AUROC. Deployment-relevant metrics are event-centric: (i) Event F1 at a fixed false-alarm budget (e.g., $\leq$1 FA/clip or FA/hour), (ii) AUROC/AUPRC at the event level, and (iii) **lead time** (mean, median, and IQR of $t_{\text{impact}} - t_{\text{alarm}}$). The single τ chosen on validation is used to produce the event trade-off plots and point estimates on UP-Fall, exactly as in our results. (Martínez-Villasenõr, Ponce, and et al. 2019)

Results

Baseline Model

The baseline model achieved some success in classifying each frame as 'fell' or 'not fell'. The quantitative results in Table 2 clearly demonstrate that kernel choice has a substantial impact on SVM performance for frame fall detection.

SVM Models	Accuracy	Precision	Recall	F1-Score
Kernel: Linear	0.81	0.82	0.78	0.80
Kernel: Poly	0.88	0.98	0.77	0.87
Kernel: RBF	0.95	0.96	0.94	0.95
RBF-RFS	0.95	0.95	0.95	0.95

Table 2: The metrics evaluating the different version of the SVM trained on Le2i. The accuracy metric refers to the 5-fold cross validation accuracy. The not fell label corresponds to positive and fell corresponds to negative.

The linear kernel achieves 81% accuracy, indicating that the decision boundary between "fell" and "not fell" frames is not linearly separable. Introducing polynomial features boosts the accuracy to 88%, suggesting that joint scores and coordinates help distinguish between cases. However, the RBF kernel produces the strongest performance (95% accuracy), confirming that a more intricate decision surface is best able to capture the complex spatial relationships inherent in human pose data. Furthermore, for all four versions of the SVM models, the precision is higher than the recall, revealing that the false-positive rate (predicting "fell" when the frame is actually "not fell") exceeds the false-negative rate. In applications that demand more safety, a false positive (erroneously flagging normal activity as a fall) is preferable to a false negative (missing an actual fall), so this bias may be preferable. Indeed, when examining individual error cases, it can be deduced that ambiguous poses, such as sitting abruptly or picking up an object, occasionally trigger fall detection.

Despite extensive recursive feature selection (RFS), the retrained SVM with selected features yields a similar accuracy (95%) to the complete feature SVM (RBF), revealing that the decision boundary is already well captured by a small subset of highly discriminative joints. From RFS, the five most important features correspond to the two knee joints, two hip joints, and the neck joint, which follows our physical intuition that lower body joints tend to be the most effective indicators of an impending fall.

The metrics for the baseline SVM do look promising, but there are heavy biases associated with sampling frames from the same set of videos. Therefore, we evaluated the SVM on an external UR Fall video dataset. Firstly, performing binary classification for 2000 frames randomly sampled

from different videos, we still obtained a high accuracy of 92%. However, we only obtained a pre-impact event recall of 65%. In other words, in only 65% of the fall videos, the SVM successfully detected falls before impact. Furthermore, for these 65% of videos, the average lead time for fall detection is only 0.15 seconds, which is essentially insignificant. The low recall and the negligible lead time reveal that the baseline SVM attains high overall frame accuracy because many background frames can be easily classified as "fell" or "not fell". On the contrary, the baseline SVM struggles to reliably anticipate or pinpoint precisely the critical moment of a fall.

TriFallNet

Frame-based Analysis TriFallNet demonstrated robust performance on the fall detection task. The model's training converged quite rapidly, where the final training loss reached 0.431 and the validation loss (calculated for the internal validation split) reached 0.467. The best validation AUC the model attained is 0.855. The details of the calculation of these metrics are described in the Methods section. The training/validation loss curve and AUC curve are shown in Figure 1. Since the gap between the training and validation curves is small, there is minimal overfitting.

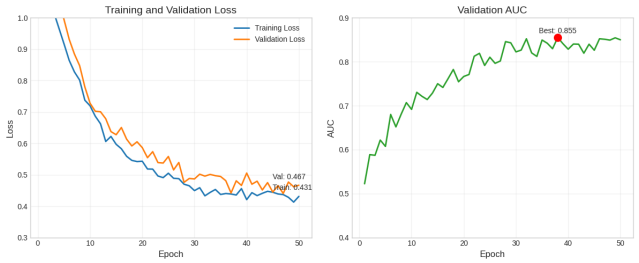

Figure 1: Training and validation loss and AUC curves for TriFallNet over 50 epochs. The best validation AUC attained is 0.855 during epoch 39 and the final validation loss is 0.467. Both curves converge smoothly and have a relatively small gap, indicating effective learning with minimal overfitting.

The receiver operating characteristic (ROC) curve corresponding to the max ROC-AUC of 0.855 is shown in Figure 2, which demonstrates that our model performs significantly better compared to the random guessing line. The optimized detection threshold to maximize the F_1 score is $\tau^* = 0.471$. Using this exact detection threshold, the final average accuracy obtained across the validation set is 80.83%. Furthermore, the model achieved a sensitivity (recall) of 0.850 and a specificity of 0.783 on the validation set. In other words, TriFallNet correctly identified 85.0% of fall events while correctly filtering out 78.3% of non-fall events at this detection threshold. This threshold also yielded a moderate precision of 0.625 and a maximum F_1 score of 0.728. The selected threshold is highlighted in red on Figure 2. One can see that the chosen threshold provided a balanced trade-off, maximizing the F_1 score and aligning with the desired point for reliable prediction.

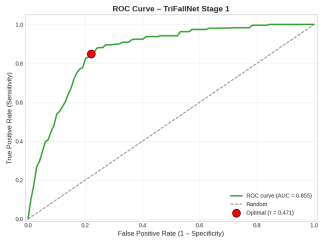

Figure 2: ROC curve for TriFallNet on the validation set. The chosen detection threshold ($\tau = 0.471$) that balances sensitivity and specificity is marked in red.

Event-based Analysis In the frame-based analysis performed on the internal validation split, the detection threshold that maximizes the F_1 score is fixed at $\tau^* = 0.471$. This exact threshold is applied in the event-based analysis to ensure a fair indication of performance on external validation.

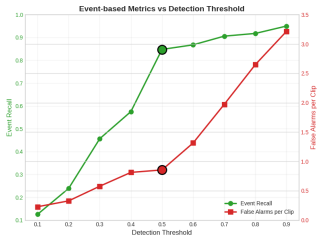

Figure 3: Event-level performance of TriFallNet on the UP-Fall dataset as a function of the detection threshold τ. Event recall (green curve) increases steadily with τ, reaching 85.2% at $\tau = 0.5$, while the average number of false alarms per clip (red curve) remains below one until $\tau \approx 0.5$ and then rises sharply. The chosen detection threshold is $\tau^* = 0.471$ which is achieves the favorable balance between high recall and low false-alarm rate.

The value for event recall and false alarms at the detection threshold range is displayed in Figure 3. At a threshold of 0.1, both event recall and false alarms remained very low. However, beyond a threshold of 0.8, event recall climbs to 90% with an average of almost three false alarms per clip. One can easily see that $\tau = 0.5$ is a great optimal threshold for fall detection, as the event recall rose significantly from 0.4 to 0.5 but only increased gradually after 0.5. Moreover, the false alarm ascended gradually before 0.5 and rose dramatically after 0.5. Indeed, at the detection threshold $\tau^* = 0.471$, the exact event recall is 85.0% with approximately 0.8 false alarms per clip.

Most importantly, TriFallNet was often able to detect falls before the person impacts the ground. The average detection lead time was 0.60 seconds prior to ground impact, with a median lead time of 0.53 seconds. For the 90th percentile, the model raised a fall detection alert 1.18 seconds before the fall event. The distribution of lead times for all videos with successful event recall is illustrated in the left histogram of Figure 4.

The temporal evolution of the model's output probability

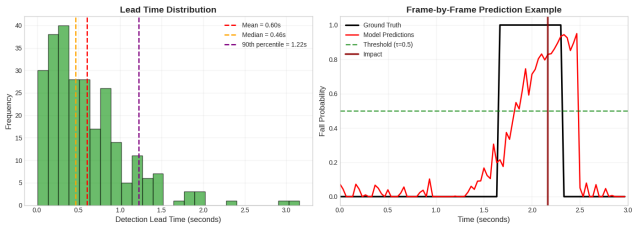

Figure 4: Temporal analysis of fall detection for successful events at $\tau = 0.5$. The left figure is a histogram of detection lead times across all clips in which the model signaled a fall before impact. The median lead time is 0.53 seconds and the 90th percentile is 1.18 seconds, demonstrating anticipatory detection. The right figure is an example of a per-frame fall-probability curve for one representative clip. The model's output hovers near 0.1 before impact, then rises sharply and crosses $\tau = 0.5$ around 0.4 seconds prior to ground contact, illustrating early prediction of the impending fall.

shown in the right of Figure 4 further illustrates the anticipatory behavior of the model. In this particular video clip, the fall probability staggered around 0.1 until 0.5 seconds before impact, at which point the probability curve rises sharply and crosses the decision threshold approximately 0.4 seconds before impact. This early rise in model probability underscores the network's ability to detect pre-fall kinematics rather than merely recognizing the fall after impact.

Feature Importance Analysis Finally, we analyzed the feature importance of each skeletal joint using Integrated Gradients. If we consider the blue joint feature in Figure 5, we observe that the lower body joints were the most influential features for fall detection. In fact, the most important joints contributing to fall prediction were the left and right knees, ankles, and hips. In addition, the lower back joints were among the top ten important features. In contrast, joints in the upper body such as shoulders or arms were ranked much lower in influence. These results align with the physical intuition on the nature of falls, which generally involve a loss of support in the lower limb or sudden leg movements.

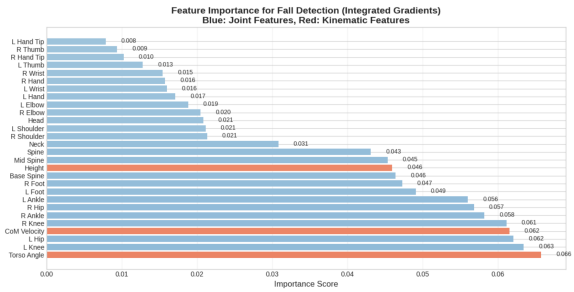

Figure 5: Integrated Gradients attribution for TriFallNet. Reveals that lower body joints contributes more to fall detection compared to upper body joints. Also, kinematic features like CoM velocity and torso inclination angle serve as essential feature in TriFallNet.

In comparison, the three kinematic features (CoM height, CoM velocity, and torso inclination angle) are also significant features. Particularly, the torso angle (ranked number one) and CoM velocity (ranked number four) are discriminatory features that help the model perform fall detection.

Discussion

The results obtained from TriFallNet indicate that a transformer-based model significantly outperforms the baseline SVM model, which relied on single frame 2D pose coordinates. While the binary classification accuracy spikes to 95% for the SVM, its relatively low recall of 65% highlights that it is inadequate for anticipatory fall detection. In fact, the SVM is only effective at classifying frames where the person is obviously not falling or lying on the ground after impact. In contrast, TriFallNet achieves a validation ROC-AUC of 0.855 with a significantly greater recall of 85%. Furthermore, TriFallNet has the ability to predict falls with substantial lead time (mean of 0.6 seconds prior to impact), revealing its strength to capture temporal dependencies within 3D pose movement over time, which is something the static SVM is incapable of. This anticipatory fall detection aligns with prior studies that leverage temporal-based architecture to predict falls early (Chen, Wang, and Yang 2021; Hua, Nan, and Lian 2019). The average 0.6 seconds of lead time is greatly beneficial to real-time monitoring systems as it could trigger immediate alerts to caregivers, emergency responders, or activate smart protective devices (e.g., inflatable hip protectors) before impact occurs.

To situate these model findings, the work introduces a unified benchmark that removes confounds that typically blur cross-paper comparisons: heterogeneous capture specifications, inconsistent label granularity, and ad hoc thresholding. All videos are normalized to 25 Hz and 256×256 RGB; labels are mapped to a shared, event-centric ontology; and a single operating point, selected once on validation, is transferred unchanged to internal and external tests. A manifest links each result to clip provenance and preprocessing, enabling exact reproduction and targeted ablations (e.g., frame-rate or resolution sensitivity). Where timing metadata exist, standardized event windows support reporting of lead time alongside clip-level classification, aligning evaluation with early-warning use cases.

TriFallNet extends these studies by integrating not just joint coordinates but also kinematic features like center of mass velocity and torso inclination angle, something not done by prior transformer-based fall detection models. Integrated Gradients analysis revealed that these features provide strong discriminative power and should be exploited to a greater extent in identifying impending falls. As an extension of these kinematic features, other degrees of freedom in the torso such as sideway bending or rotation can be included in the model. Future works include adding joint kinematic parameters such as joint angles, joint angles' velocity, and acceleration in combination with a more advanced set of kinematic features to improve fall detection. Moreover, a valuable future step would be training both TriFallNet and baseline SVM exclusively on the most significant features

and then comparing their performance directly to our current implementation.

Compared to the existing literature, TriFallNet demonstrated a competitive performance. For instance, while previous pose-based fall detection methods mainly achieved accuracy within the range of 80-90% (Hua, Nan, and Lian 2019; Maldonado-Mendez et al. 2022), our results also lie within the range of 80-90% yet demonstrated more success in anticipating falls. However, despite these strengths, TriFallNet's precision remains a moderate 0.625. This reveals great room for improvement in distinguishing true fall events from false positives. To combat this issue, future studies can improve precision through implementing more sophisticated spatial-temporal architecture or more refined thresholding strategies.

Additionally, TriFallNet benefits significantly from transfer learning with its pretrained Skeleton-MixFormer backbone. Although training from the scratch comparison was not feasible due to limited data, transfer learning in theory should have contributed to the observed performance. An ablation study is necessary to identify precisely which components of TriFallNet contribute most to its success, guiding future adjustments to either freeze or fine-tune elements within the Skeleton-MixFormer backbone.

Nonetheless, there are limitations inherent in this approach. Firstly, our methods depend on the generalizability of pretrained models, which might introduce bias toward certain movements prevalent in the original training datasets. Another limitation is the relatively limited size of the fine-tuning datasets collected from different data sources, which might restrict model performance due to underfitting. Furthermore, the dataset presents multiple biases that can limit the real-world applicability of the model. Primarily, several datasets used in training, such as Le2i and UP-Fall, contain enacted or simulated falls, which may not fully capture the biomechanics present in a genuine fall. In addition, the majority of the dataset captured falls of only young or middle-aged adults. Therefore, some methods to improve demographic representativeness, especially for the elderly populations, are crucial to guarantee generalizability.

An important practical consideration is the end-to-end system latency, determined primarily by the speed of the pose estimation backbone and fall detection pipelines. Preliminary measurements indicate an inference frame rate of approximately 20 FPS, corresponding to an average latency of roughly 50 milliseconds per frame, which should support real-time fall detection. Deploying such a system for real-time application requires a further consideration of trade-offs between false alarm rates (FAR) and the critical consequences of missed falls, which is something we tackled when we balanced event recall and false alarms. However, quantitative measures may be insufficient and a human-in-the-loop review process is required to validate the system before complete autonomy.

Conclusion

The results from TriFallNet reveal that combining 3D pose coordinates with calculated kinematic features enables accurate fall prediction and anticipatory fall warning. Integrated Gradients analysis shows that the lower body joints, center of mass kinematics, and torso inclination angles are the most discriminatory features for fall detection, which align with the physical intuition of falling. Understanding how each feature shapes fall mechanics can guide the development of more reliable preventive fall-detection systems. Furthermore, our approach can be integrated into real-time monitoring systems in places like elderly homes to trigger immediate alerts to ensure rapid response to falls. In fact, even less than one second of lead time can mitigate injury severity by enabling protective measures. Finally, our findings generate insights into the specific biomechanics behind falling and bolster the development of fall prevention strategies.

The benchmark itself is intended as an enabling resource: by standardizing inputs, labels, and operating points, it yields apples-to-apples comparisons and transparent, auditable metrics that reflect deployment realities. As additional naturalistic video, broader subject cohorts, and consolidated timing annotations are incorporated, the protocol will further tighten the link between laboratory evaluations and real-world early-warning systems, accelerating the development of reliable, low-latency fall detection.

References

Alam, E.; Sufian, A.; Dutta, P.; and Leo, M. 2022. Vision-Based Human Fall Detection Systems Using Deep Learning: A Review. *Computers in Biology and Medicine*, 146: 105626.

Alam, E.; Sufian, A.; Dutta, P.; Leo, M.; and Hameed, I. A. 2024a. GMDCSA-24: A Dataset for Human Fall Detection in Videos. *Data in Brief*, 57: 110892.

Alam, E.; Sufian, A.; Dutta, P.; Leo, M.; and Hameed, I. A. 2024b. GMDCSA-24: A Dataset for Human Fall Detection in Videos. *Data in Brief*, 57: 110892.

Auvinet, E.; Rougier, C.; Meunier, J.; St-Arnaud, A.; and Rousseau, J. 2010a. Multiple Cameras Fall Dataset. Technical Report Technical Report 1350, DIRO, Université de Montréal, Montréal, Canada.

Auvinet, E.; Rougier, C.; Meunier, J.; St-Arnaud, A.; and Rousseau, J. 2010b. Multiple Cameras Fall Dataset. Technical Report Technical Report 1350, DIRO, Université de Montréal, Montréal, Canada.

Centers for Disease Control and Prevention. 2024a. Economics of Injury and Violence Prevention: Cost of Older Adult Falls. https://www.cdc.gov/injury-violence-prevention/economics/index.html. Accessed 2025-10-22.

Centers for Disease Control and Prevention. 2024b. Older Adult Falls Data. https://www.cdc.gov/falls/data-research/index.html. Accessed 2025-10-22.

Charfi, I.; Mitéran, J.; Dubois, J.; Atri, M.; and Tourki, R. 2013a. Optimized Spatio-Temporal Descriptors for Real-Time Fall Detection: Comparison of Support Vector Machine and Adaboost-Based Classification. *Journal of Electronic Imaging*, 22(4): 041106.

Charfi, I.; Mitéran, J.; Dubois, J.; Atri, M.; and Tourki, R. 2013b. Optimized Spatio-Temporal Descriptors for Real-Time Fall Detection: Comparison of Support Vector Ma-

chine and Adaboost-Based Classification. *Journal of Electronic Imaging*, 22(4): 041106.

Chen, Z.; Wang, Y.; and Yang, W. 2021. Video Based Fall Detection Using Human Poses. *arXiv preprint arXiv:2107.14633*.

Gutiérrez, J.; Lagriffoul, F.; Orcesi, F.; et al. 2021. Comprehensive Review of Vision-Based Fall Detection Systems. *Sensors*, 21(3): 947.

Hua, M.; Nan, Y.; and Lian, S. 2019. Falls Prediction Based on Body Keypoints and Seq2Seq Architecture. In *arXiv preprint arXiv:1908.00275*.

Kakara, R. S.; Bergen, G.; Burns, E. R.; and Stevens, M. R. 2023. Nonfatal and Fatal Falls Among Adults Aged $\geq$65 Years — United States, 2020–2021. *MMWR Morbidity and Mortality Weekly Report*, 72(35): 938–943.

Kwolek, B.; and Kepski, M. 2014. Human Fall Detection on Embedded Platform Using Depth Maps and Wireless Accelerometer. *Computer Methods and Programs in Biomedicine*, 117(3): 489–501.

Kwolek, B.; and Kepski, M. 2014. Human Fall Detection on Embedded Platform Using Depth Maps and Wireless Accelerometer. *Computer Methods and Programs in Biomedicine*, 117(3): 489–501.

Maldonado-Mendez, C.; Hernandez-Mendez, S.; Torres-Muñoz, D.; and Hernandez-Mejia, C. 2022. Fall detection using features extracted from skeletal joints and SVM: Preliminary results. *Multimedia Tools and Applications*, 81: 27657–27681.

Martínez-Villaseñor, L.; Ponce, H.; Brieva, J.; et al. 2019. UP-Fall Detection Dataset: A Multimodal Approach. *Sensors*, 19(9): 1988.

Martínez-Villasenõr, L.; Ponce, H.; and et al. 2019. UP-Fall Detection Dataset: A Multimodal Approach for Fall Detection. *Sensors*. Dataset includes two RGB cameras and wearable sensors; we use RGB only.

National Center for Injury Prevention and Control (CDC). 2017. Every 20 Minutes an Older Adult Dies from a Fall in the United States. https://stacks.cdc.gov/view/cdc/46789. STEADI infographic.

Núñez-Marcos, A.; and Arganda-Carreras, I. 2024. Transformer-Based Fall Detection in Videos. *Engineering Applications of Artificial Intelligence*, 132: 107937.

Schneider, D.; Marinov, Z.; Baur, R.; Zhong, Z.; Düger, R.; and Stiefelhagen, R. 2025. OmniFall: A Unified Staged-to-Wild Benchmark for Human Fall Detection. arXiv:2505.19889.

Xin, W.; Miao, Q.; Chen, B.; et al. 2023. Skeleton MixFormer: Multivariate Topology Representation for Skeleton-Based Action Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*. Ottawa, Canada: ACM.

Zhu, W.; Ma, X.; Liu, Z.; Liu, L.; Wu, W.; and Wang, Y. 2023. MotionBERT: A Unified Perspective on Learning Human Motion Representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.