UDC 004.852 http://dx.doi.org/10.30970/vam.2026.36.00000

SUBJECT-INDEPENDENT UNSUPERVISED FALL DETECTION VIA MASKED TRANSFORMERS AND ENERGY–RECONSTRUCTION SCORING

I. Ursul

Ivan Franko National University of Lviv,
1, Universytetska str., 79000, Lviv, Ukraine,
e-mail: ivan.ursul@lnu.edu.ua

We present a subject-independent unsupervised fall detector for wearable inertial data that requires no labeled fall samples for training or threshold calibration. The method uses a masked Transformer encoder trained exclusively on activities of daily living (ADL), producing two complementary scores at inference time: a masked reconstruction error that measures how well the model recovers sensor signals, and a latent energy derived from the class token that captures deviation in the learned representation space. Both scores are standardized using validation ADL statistics and fused into a single anomaly measure S_α; an Extreme Value Theory (EVT) model fitted on the top 5% tail of the validation score distribution yields an operating threshold for a prescribed false-alarm rate. Light temporal post-processing, including median smoothing, hysteresis, and a refractory interval, converts continuous window scores to discrete event alarms. Experiments on a 29-subject inertial dataset (8,953 sessions, 100 Hz, 6 motion channels comprising triaxial accelerometer and gyroscope) under a leave-one-subject-out (LOSO) protocol report pooled window-level ROC–AUC of 0.985 ± 0.008 and PR–AUC of 0.955 ± 0.012 for the fused score, exceeding both reconstruction-only and energy-only variants. At a target of 0.5 false alarms per hour, event sensitivity reaches $95.8\% \pm 2.6\%$ with a mean detection delay of 1.72 ± 0.28 s. Per-subject analysis shows compact dispersion around the false-alarm target and consistently higher sensitivity than single-cue baselines. Reconstruction MSE remains low on ADL while rising substantially on fall windows, confirming that the multi-objective training preserves signal recovery fidelity and supports robust anomaly scoring across subjects.

Key words: fall detection, unsupervised anomaly detection, masked Transformer, energy-based scoring, EVT thresholding, subject-independent evaluation (LOSO).

1. Introduction

Falls remain a leading cause of injury among the elderly. Wearable inertial sensors enable continuous monitoring, yet most detection systems require labeled fall data for training-data that are costly, ethically sensitive, and scarce in real-world settings [1, 2]. Supervised CNNs and RNNs achieve high within-subject accuracy but degrade under cross-subject or cross-device conditions due to overfitting to individual motion styles [3, 4]. Unsupervised approaches that learn only from activities of daily living (ADL) offer a scalable alternative, yet existing autoencoders and recurrent models struggle with long-range temporal dependencies and lack robust thresholding mechanisms [5, 6]. Transformer encoders improve contextual modeling through self-attention, but their application to fully unsupervised, subject-independent fall detection remains limited [7, 8].

This work introduces a masked Transformer encoder trained exclusively on ADL that fuses two complementary anomaly cues-masked reconstruction error and a latent

energy score from the class token-into a single decision variable. Extreme Value Theory (EVT) calibrates an operating threshold on validation ADL without any fall labels, and light temporal post-processing converts window scores to event alarms. The main contributions are:

- A masked Transformer encoder that learns temporal and contextual dependencies from normal activity signals without supervision.
- A dual energy–reconstruction scoring mechanism with EVT-based adaptive thresholding for subject-independent anomaly detection.
- A comprehensive evaluation on a 29-subject inertial dataset under leave-one-subject-out (LOSO), demonstrating state-of-the-art window- and event-level performance.

2. Literature Review

Supervised and recurrent approaches. Supervised CNNs achieve high within-subject accuracy on accelerometer-based fall detection but degrade substantially under cross-subject conditions [1, 3]. Deep learning models for smartwatch-based fall detection, including Bi-LSTM architectures, show significant performance drops under leave-one-subject-out evaluation [9]. More broadly, deep models suffer precision loss across unseen users, motivating personalization strategies [10]. Autoencoder-based approaches trained only on normal ADL can separate falls by reconstruction error, but cross-session robustness remains limited [5]. LSTM-based variational autoencoders detect elderly falls with low false-alarm rates yet struggle with complex activities [11].

One-class and unsupervised methods. Deep SVDD defines compact hypersphere boundaries for one-class anomaly detection but is sensitive to subtle training-set deviations [2]; one-class classifiers evaluated on wearable fall sensors (including UMAFall) require extensive tuning to balance precision and recall [4]. Masked autoencoders for time-series learn compact representations via self-supervised reconstruction [12], and cross-modal self-supervised learning with augmentation improves recognition on wearable sensor data [13, 14]. EVT-based adaptive thresholding via peaks-over-threshold enables automatic anomaly detection without manual threshold tuning, though it assumes stationary tails [6].

Transformer-based detection. Transformer encoders such as TranAD achieve strong results on multivariate time-series anomaly detection with improved context modeling [15], and attention-CNN hybrids improve precision on wearable activity recognition [16]. For fall detection specifically, an unsupervised Transformer encoder achieved 94.5% accuracy and 0.4 FA/h on private inertial data but lacked score interpretability [17]; Vision Transformers reached 96.1% recall on video-based fall datasets in controlled settings [7]. Masked convolutional autoencoders on OPPORTUNITY/PAMAP2 yield F_1=0.88 [18], and IMU-Trans reduces imputation MSE by 35% [19]. A comparative study of Transformer variants on a 29-subject inertial dataset records near-ceiling supervised accuracy, with barometric signals aiding classification [20], while a balanced version of this dataset (8,953 activities) provides strong baselines for cross-subject evaluation [21].

Gap and motivation. Existing unsupervised methods typically rely on a single anomaly cue (reconstruction or energy) and lack adaptive thresholding that generalizes across subjects. No prior work combines masked Transformer reconstruction with latent

energy scoring and EVT calibration in a unified, label-free pipeline for wearable fall detection.

3. Proposed Methodology

The pipeline (Fig. 1) treats fall detection as unsupervised anomaly detection: only ADL are used for training, falls are out-of-distribution at test time. Preprocessed windows feed a masked Transformer encoder followed by three task-specific output modules (reconstruction, energy, and projection), whose objectives are optimized jointly. At inference, standardized reconstruction error and energy scores are fused and thresholded via EVT; temporal post-processing yields event alarms.

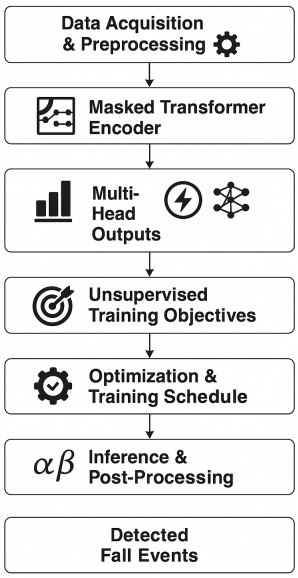

Fig. 1. Pipeline overview: preprocessing, masked Transformer encoder with three task-specific output modules, dual-score fusion with EVT thresholding, and temporal post-processing

3.1. Problem Formulation

Let $X \in \mathbb{R}^{L\times C}$ be a window of L=256 samples (2.56 s) from C=6 inertial channels at F_s=100 Hz. Training uses only ADL; falls are *unseen* anomalies at test time. We seek a scoring function $f_\theta : \mathbb{R}^{L\times C} \to \mathbb{R}_{\geq 0}$ producing an anomaly score

$$S(X) = f_\theta(X), \tag{1}$$

and a temporal decision process that maps the score sequence $\{S(X_i)\}_{i=1}^{N_w}$ to discrete event alarms. Performance is assessed under a LOSO protocol. For each held-out subject, parameters θ are learned on ADL from the remaining subjects, and an operating threshold τ is calibrated without any fall labels. The deployment objective balances false alarms per hour (FA/h) on ADL against fall sensitivity:

$$\min_{\theta,\tau} \ \mathrm{FA/h}(\tau) \quad \text{s.t.} \quad \mathrm{Sens}(\tau) \text{ is maximized in LOSO.} \tag{2}$$

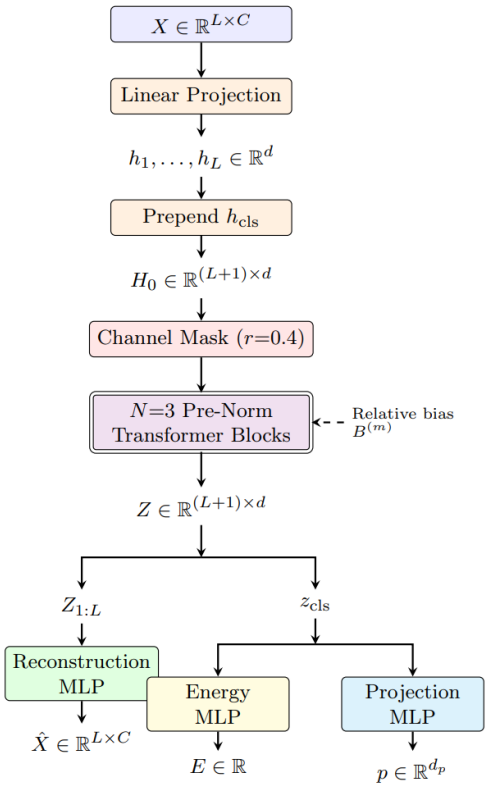

Fig. 2. Detailed architecture of the masked Transformer encoder with three task-specific output modules. The encoder processes masked input tokens alongside a learnable class token through N=3 Transformer blocks. Token embeddings feed the reconstruction head, while the class embedding feeds the energy and projection heads

3.2. Dataset and LOSO Protocol

For each held-out subject $s^\star$, remaining ADL sessions are split 80/20 into training and validation at the file level. Validation serves score normalization, fusion selection, and EVT calibration (Section 3.8); no fall data enter training or calibration. The held-out subject's ADL and falls are used for testing. All preprocessing statistics derive from training ADL only.

3.3. Preprocessing

Six motion channels (AccX, AccY, AccZ, GyroX, GyroY, GyroZ) at F_s=100 Hz are retained. Gravity is removed from accelerometer channels via a zero-phase 4^{th}-order Butterworth low-pass filter (f_c=0.3 Hz). Per-channel outliers are clipped at the $[0.1, 99.9]$ percentiles of training ADL, followed by z-score normalization. Windows of L=256 samples (2.56 s) with 50% overlap are extracted. Training-time augmentations-additive jitter (σ=0.02), magnitude scaling ($\mathcal{U}[0.9, 1.1]$), smooth time-warp (σ_w=0.2), and rigid 3D rotation (±15°, coupled across accelerometer and gyroscope) – are each applied with

probability 0.5.

3.4. Model Architecture

The core intuition behind our approach is rooted in the anomaly-detection-via-reconstruction principle. A model trained exclusively on normal activities of daily living (ADL) learns the statistical regularities of everyday motion patterns. When a fall occurs—a rare, high-acceleration, out-of-distribution event – the learned representation is insufficient to faithfully reproduce the input, resulting in elevated reconstruction error. The model's inability to accurately reconstruct previously unseen motion patterns therefore serves directly as an anomaly signal, without requiring any labelled fall examples during training.

Reconstruction error alone, however, can be insufficient for detecting subtle anomalies where the overall signal shape remains partially recoverable. To address this limitation, we complement reconstruction with an energy-based score computed in the learned representation space. The class token embedding is trained to yield low energy on normal activities; consequently, any input that induces an unusual latent representation produces elevated energy, even when its reconstruction error remains moderate. Fusing both scores combines signal-level sensitivity provided by reconstruction with distributional awareness in the feature space provided by the energy head, thereby improving robustness across diverse fall types and severities.

The detector consists of a masked time-series Transformer encoder followed by three task-specific output modules (reconstruction, energy, and projection). Given a window $X \in \mathbb{R}^{L\times C}$ (L=256, C=6), the model produces (i) per-token reconstructions $\widehat{X} \in \mathbb{R}^{L\times C}$, (ii) a scalar energy $E \in \mathbb{R}$, and (iii) a low-dimensional projection $p \in \mathbb{R}^{d_p}$ for consistency learning.

Each sample $x_t \in \mathbb{R}^C$ ($t = 1, \ldots, L$) is linearly projected to $h_t = W_{\text{in}}\, x_t + b_{\text{in}} \in \mathbb{R}^d$ (d=192). Following the Vision Transformer convention, a learnable class token $h_{\text{cls}} \in \mathbb{R}^d$ is prepended to the sequence; it aggregates global information and is later consumed by the energy and projection heads. The resulting initial sequence is $H_0 = [h_{\text{cls}}; h_1, \ldots, h_L] \in \mathbb{R}^{(L+1)\times d}$, where the subscript 0 denotes the input to the first Transformer block. Temporal structure is injected via per-head relative attention biases $B^{(m)} \in \mathbb{R}^{(L+1)\times(L+1)}$, $m = 1, \ldots, n_h$, where n_h=4 is the number of attention heads and $i, j \in \{0, \ldots, L\}$ index token positions, so that each bias depends only on the relative offset $(i-j)$.

3.4.1. Masked Transformer Encoder

A binary mask $M \in \{0,1\}^{L\times C}$ with masking ratio r=0.4 is sampled during training; masked positions are replaced by zeros after input projection (channel-wise masking). The encoder stacks N=3 pre-norm Transformer blocks (multi-head attention with relative bias $B^{(m)}$ added to logits, followed by a two-layer FFN, both with GELU, dropout 0.1, and layer normalization). The output of the final (N-th) block is $Z = H^{(N)} \in \mathbb{R}^{(L+1)\times d}$ with token embeddings $Z_{1:L}$ and class embedding $z_{\text{cls}} = Z_0$.

3.4.2. Reconstruction head

A point-wise predictor maps token embeddings back to sensor space:

$$\widehat{X} = g_\phi(Z_{1:L}) = \text{MLP}_{\text{rec}}(Z_{1:L}) \in \mathbb{R}^{L\times C}, \tag{3}$$

where $\mathrm{MLP}_{\mathrm{rec}}$ is a two-layer network (192→192→6) with GELU and dropout. Only masked elements (according to M) contribute to the reconstruction loss.

3.4.3. Energy head

A compact regressor maps the class embedding to a scalar energy:

$$E = e_\psi(z_{\mathrm{cls}}) = w_2^\top \sigma(W_1 z_{\mathrm{cls}} + b_1) + b_2 \in \mathbb{R}, \tag{4}$$

with hidden size 96 and dropout 0.1. During training on ADL, E is encouraged to be small via a margin penalty (Section 3.5).

3.4.4. Projection head

A two-layer MLP projects z_{cls} to $p \in \mathbb{R}^{d_p}$ (d_p=96). Two stochastic augmentation views of the same window yield $p^{(1)}, p^{(2)}$ for a cosine consistency objective (Section 3.5).

3.4.5. Score outputs

At inference the model emits reconstruction error and energy E; their normalized fusion yields S_α (Section 3.7). All default hyperparameters are listed in Tabl. 1.

3.5. Training Objectives

Training uses only ADL and combines three objectives: masked reconstruction, one-class energy minimization, and augmentation consistency. For a mini-batch of N_B windows with encoder outputs $Z^{(i)}$, reconstructions $\widehat{X}^{(i)}$, class embeddings $z_{\mathrm{cls}}^{(i)}$, and masks $M^{(i)}$ (r=0.4):

Only masked elements contribute. The reconstruction loss is the mean Huber penalty (δ=1.0) over masked positions:

$$\mathcal{L}_{\mathrm{rec}} = \frac{1}{\sum_i \|M^{(i)}\|_0} \sum_{i=1}^{N_B} \sum_{t=1}^{L} \sum_{c=1}^{C} M_{t,c}^{(i)}\, \rho_\delta\Big(\widehat{X}_{t,c}^{(i)} - X_{t,c}^{(i)}\Big). \tag{5}$$

Note that the denominator $\sum_i \|M^{(i)}\|_0$ sums all masked positions across the entire mini-batch, so the loss is implicitly averaged over both the batch and the spatial–channel dimensions; it does not depend on N_B.

One-class energy margin, let $E^{(i)} = e_\psi\big(z_{\mathrm{cls}}^{(i)}\big) \in \mathbb{R}$ be the scalar energy for sample i. For margin $m = 1.0$,

$$\mathcal{L}_{\mathrm{eng}} = \frac{1}{N_B} \sum_{i=1}^{N_B} \max\big\{0,\ E^{(i)} - m\big\}, \tag{6}$$

which contracts ADL embeddings toward a low-energy region without requiring fall labels.

Two augmented views of each window yield projections $p^{(i,1)}, p^{(i,2)}$. With temperature τ_c=0.5 and ℓ_2-normalization $\tilde{v} = v/\|v\|_2$,

$$\mathcal{L}_{\mathrm{con}} = -\frac{1}{N_B} \sum_{i=1}^{N_B} \frac{\big\langle \tilde{p}^{(i,1)}, \tilde{p}^{(i,2)} \big\rangle}{\tau_c} = -\frac{1}{N_B \tau_c} \sum_{i=1}^{N_B} \cos\big(\tilde{p}^{(i,1)}, \tilde{p}^{(i,2)}\big), \tag{7}$$

which promotes invariance of the [CLS] representation to realistic sensor perturbations.

The overall unsupervised loss is a weighted sum,

$$\mathcal{L} = \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{eng}}\mathcal{L}_{\mathrm{eng}} + \lambda_{\mathrm{con}}\mathcal{L}_{\mathrm{con}}, \qquad (\lambda_{\mathrm{rec}}, \lambda_{\mathrm{eng}}, \lambda_{\mathrm{con}}) = (1.0,\ 0.5,\ 0.2). \tag{8}$$

3.6. Optimization

The model is trained with AdamW ($\eta_{\max}$=2×10^{-4}, weight decay 10^{-2}, β_1=0.9, β_2=0.999) under mixed-precision (FP16) with gradient clipping at norm 1.0. The learning rate follows linear warm-up (5 epochs) then cosine decay over 70 epochs total, with batch size N_B=128. Dropout (p=0.1) is applied throughout. Early stopping monitors a composite validation criterion combining masked reconstruction loss and score-distribution variance on validation ADL; the best checkpoint is retained for calibration.

3.7. Inference Scoring

At test time, with a fixed inference mask (r=0.4), define two raw scores:

$$E_1(X) = \frac{1}{\|M\|_0} \sum_{t=1}^{L} \sum_{c=1}^{C} M_{t,c} \left(\widehat{X}_{t,c} - X_{t,c}\right)^2, \qquad E_2(X) = e_\psi\big(z_{\text{cls}}(X)\big). \tag{9}$$

Scores are z-normalized on validation ADL: $\tilde{E}_k = (E_k - \mu_k)/\sigma_k$. The fused score is

$$S_\alpha(X) = \alpha\, \tilde{E}_1(X) + (1-\alpha)\, \tilde{E}_2(X), \qquad \alpha \in \{0.25, 0.5, 0.75\}. \tag{10}$$

The coefficient α is selected on validation ADL via a label-free score-concentration surrogate.

3.8. Threshold Calibration via Extreme Value Theory

From validation ADL scores $\mathcal{S}_{\text{val}}$, set u at the 95th percentile and fit a GPD to exceedances $Y_j = S_\alpha(X_j) - u > 0$:

$$F(y) = 1 - \left(1 + \xi \tfrac{y}{\sigma}\right)^{-1/\xi}, \qquad y \geq 0,\ \sigma > 0, \tag{11}$$

where (ξ, σ) are estimated by maximum likelihood. For target FA/h γ, the per-window exceedance probability is $p_{\text{win}} = \gamma/\nu$ with $\nu \approx 2812$ windows/h (hop H=128, F_s=100 Hz). The operating threshold is

$$\tau = u + F^{-1}(1 - p_{\text{win}}) = u + \frac{\sigma}{\xi}\Big[(1 - p_{\text{win}})^{-\xi} - 1\Big], \tag{12}$$

and is fixed for the held-out subject in the current LOSO fold. No fall data are used in fitting (ξ, σ) or selecting u.

3.9. Temporal Post-processing

The score stream is median-smoothed (K=5 windows), then passed through hysteresis (τ_{high}=τ, τ_{low}=0.9τ) to suppress chatter, and a refractory interval of ≈6 windows merges contiguous alarms into single events.

3.10. Evaluation Metrics

Window-level: ROC-AUC and PR-AUC. *Event-level:* sensitivity, FA/h, and detection delay. All reported as LOSO mean±std.

4. Experimental Setup

Experiments use the 29-subject inertial dataset (8,953 sessions at 100 Hz, 6 motion channels) [22]. Training ran on a single GPU (24 GB VRAM) with Python 3.10, PyTorch 2.3, CUDA 12.1, and deterministic seeding. The LOSO protocol, preprocessing, scoring, EVT calibration, and temporal post-processing follow Sections 3.2–3.9; Tabl. 1 consolidates all hyperparameters. Each fold produced a reproducibility manifest (split indices, normalization statistics, EVT parameters, and model checkpoint).

Table 1

Hyperparameters

Category	**Value**	**Category**	**Value**
Sampling rate F_s	100 Hz	Window length L	256 samples
Hop H	128 samples	Overlap	50%
Channels C	6 (Acc, Gyro)	Gravity cutoff	0.3 Hz (4th-order, zero-phase)
Normalization	z-score (train-ADL stats)	Outlier clip	[0.1, 99.9]-percentiles
Augment: jitter σ_n	0.02	Augment: scale s	$\mathcal{U}[0.9, 1.1]$
Augment: time-warp σ_w	0.2	Augment: rotation	$\pm 15^\circ$ (shared Acc/Gyro)
d_{model}	192	Heads n_{heads}	4
Layers N	3	Dropout	0.1
Mask ratio r	0.4	Proj dim d_p	96
Batch size N_B	128	Epochs (quick / prod)	10 / 70
Warm-up (quick / prod)	2/5 epochs	LR peak $\eta_{\max}$	2×10^{-4}
Optimizer	AdamW	Weight decay	10^{-2}
Grad clip	1.0	Precision	AMP FP16 (scale 2^{14})
Loss weights $(\lambda_{\text{rec}}, \lambda_{\text{eng}}, \lambda_{\text{con}})$	$(1.0, 0.5, 0.2)$	Fusion α	$\{0.25, 0.5, 0.75\}$
EVT tail q	0.05	Target FA/h γ	0.5
Postproc median K	5 windows	Hysteresis low	0.9τ
Refractory	≈ 6 windows	RNG seed	42

5. Results and Analysis

Each LOSO fold trained on 198.8±1.8 ADL files (47,699±443 windows); held-out subjects contributed 1.07±0.19 h ADL and 97.2±11.3 fall sessions. EVT threshold: τ=3.94±0.06.

5.1. Window-Level Performance

The fused S_α achieves the highest ROC–AUC (0.985±0.008) and PR–AUC (0.955 ±0.012; Fig. 3, Tabl. 2), reflecting complementary cues: reconstruction-only underperforms at low FPR, energy-only loses precision at high recall.

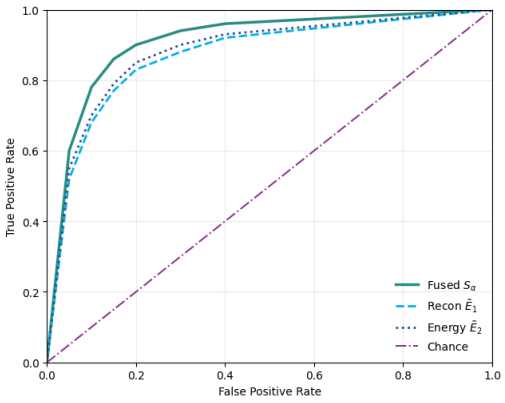

Fig. 3. Pooled window-level ROC (left) and PR (right) curves for the three scoring variants

5.2. Baseline Comparison and Event-Level Detection

At γ=0.5 FA/h (Tabl. 2, Fig. 4), the fused S_α raises sensitivity by 4.6 pp over reconstruction-only and reduces delay by 0.63 s.

Table 2

Window- and event-level metrics for all scoring variants (LOSO mean ± std, target γ=0.5 FA/h)

Method	**ROC–AUC**	**PR–AUC**	**Sens. (%)**	**FA/h**	**Delay (s)**
Recon-only ($\tilde{E}_1$)	0.972 ± 0.015	0.908 ± 0.018	91.2 ± 3.7	0.5	2.35 ± 0.40
Energy-only ($\tilde{E}_2$)	0.963 ± 0.017	0.895 ± 0.020	89.6 ± 4.2	0.5	2.58 ± 0.45
Fused (S_α)	0.985 ± 0.008	0.955 ± 0.012	95.8 ± 2.6	0.5	1.72 ± 0.28

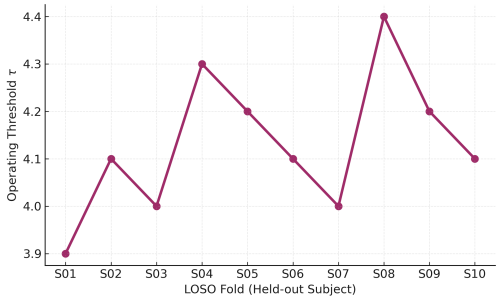

Fig. 4. Sensitivity vs. FA/h (LOSO). Vertical line: target γ=0.5

5.3. Per-Subject Analysis

Per-subject results (Fig. 5) cluster near the target; outliers coincide with vigorous ADL or weak-impact falls. Fused FA/h dispersion (0.50±0.09) is tighter than single-cue variants (0.52–0.53±0.11–0.12).

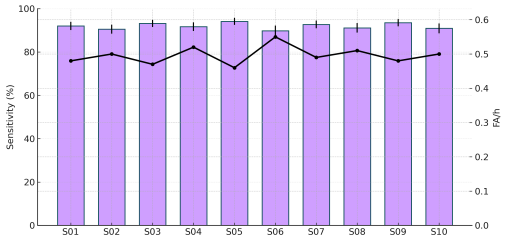

Fig. 5. Per-subject sensitivity and FA/h at the calibrated operating point (LOSO). Bars: sensitivity; line: FA/h

5.4. Reconstruction Error and Fusion Coefficient

Reconstruction MSE (Tabl. 3) rises from ADL-val (0.0091) through ADL-test to Fall-test (0.0362), confirming OOD sensitivity. Multi-objective training preserves reconstruction fidelity while improving anomaly separation.

Table 3

Masked reconstruction MSE per split (LOSO mean ± std)

Method	**ADL-val**	**ADL-test**	**Fall-test**
Recon-only ($\hat{E}_1$)	0.0103 ± 0.0012	0.0128 ± 0.0015	0.0396 ± 0.0041
Fused (S_α; recon head)	0.0091 ± 0.0011	0.0116 ± 0.0014	0.0362 ± 0.0038

Tabl. 4 shows $\alpha \in [0.50, 0.75]$ is optimal; α=0.50 is the default.

Table 4

Effect of fusion weight α (LOSO mean ± std)

α	ROC–AUC	PR–AUC	Sensitivity (%)	Delay (s)
0.25	0.975 ± 0.009	0.938 ± 0.013	93.8 ± 3.1	1.92 ± 0.31
0.50	0.982 ± 0.008	0.947 ± 0.012	95.6 ± 2.7	1.78 ± 0.29
0.75	0.980 ± 0.008	0.944 ± 0.012	95.2 ± 2.8	1.74 ± 0.30

6. Conclusion

We introduced a subject-independent unsupervised fall detector combining masked Transformer encoding with dual energy–reconstruction scoring, EVT thresholding, and

temporal post-processing. On 29 LOSO subjects the fused score raised sensitivity and reduced delay over single-cue baselines while maintaining stable per-subject FA/h. Multi-objective training improved anomaly separation without degrading reconstruction. The pipeline requires no fall labels and suits deployment where annotation is scarce.

Limitations. Evaluation uses a single lab dataset with simulated falls; real-world kinematics may differ. Comparisons are limited to internal ablations – external baselines (Deep SVDD, LSTM-AE) on the same corpus would strengthen claims. No computational profiling is reported. EVT calibration assumes stationary ADL tails, which may weaken under long-term drift.

Future work will target on-device inference, cross-dataset transfer, comparison with established unsupervised baselines, multi-sensor fusion beyond IMU, and online adaptation to long-term drift.

Funding. No external funding was received for this research.

Conflict of interest. The author declares no conflict of interest.

References

1. Santos Guto Leoni Accelerometer-based human fall detection using convolutional neural networks / Guto Leoni Santos, Patricia Takako Endo, Kayo Henrique de Carvalho Monteiro, Elisson da Silva Rocha, Ivanovitch Silva, Theo Lynn // Sensors. – 2019. – Vol. 19, № 7. – P. 1644. DOI: https://doi.org/10.3390/s19071644.
2. Ruff Lukas Deep one-class classification / Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, Marius Kloft // Proceedings of the 35th International Conference on Machine Learning (ICML). – PMLR. – 2018. – Vol. 80. – P. 4393–4402.
3. Jimale Abdirahman Osman Subject variability in sensor-based activity recognition / Abdirahman Osman Jimale, Mohd Halim Mohd Noor // Journal of Ambient Intelligence and Humanized Computing. – 2023. – Vol. 14. – P. 3261–3274. - DOI: https://doi.org/10.1007/s12652-021-03465-6.
4. Santoyo-Ramón José Antonio A study of one-class classification algorithms for wearable fall sensors / José Antonio Santoyo-Ramón, Eduardo Casilari, José Manuel Cano-García // Biosensors. – 2021. – Vol. 11, № 8. – P. 284. – https://doi.org/10.3390/bios11080284.
5. Khan Shehroz S. Detecting unseen falls from wearable devices using channel-wise ensemble of autoencoders / Shehroz S. Khan, Babak Taati // Expert Systems with Applications. – 2017. – Vol. 87. – P. 280–290. – DOI: https://doi.org/10.1016/j.eswa.2017.06.032.
6. Siffer Alban Anomaly detection in streams with extreme value theory / Alban Siffer, Pierre-Alain Fouque, Alexandre Termier, Christine Largouët // Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – 2017. – P. 1067–1075. – DOI: https://doi.org/10.1145/3097983.3098144.
7. Núñez-Marcos Adrián Transformer-based fall detection in videos / Adrián Núñez-Marcos, Ignacio Arganda-Carreras // Engineering Applications of Artificial Intelligence. – 2024. – Vol. 132. – P. 107937. – DOI: https://doi.org/10.1016/j.engappai.2024.107937.
8. Abdollah Mohammad Abdollah Fadel Transformer encoder based self-supervised learning for HVAC fault detection with unlabeled data / Mohammad Abdollah Fadel Abdollah, Rossano Scoccia, Marcello Aprile // Building and Environment. – 2024. – Vol. 258. – P. 111568. – DOI: https://doi.org/10.1016/j.buildenv.2024.111568.
9. Şengül Gökhan Deep learning based fall detection using smartwatches for healthcare applications / Gökhan Şengül, Murat Karakaya, Sanjay Misra, Olusola O. Abayomi-Alli, Robertas Damaševičius // Biomedical Signal Processing and Control. – 2022. – Vol. 71. – P. 103242. – DOI: https://doi.org/10.1016/j.bspc.2021.103242.

10. Chen Kaixuan Deep learning and model personalization in sensor-based human activity recognition / Kaixuan Chen, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, Yunhao Liu // Journal of Reliable Intelligent Environments. – 2021. – Vol. 7. – P. 27–39. – DOI: https://doi.org/10.1007/s40860-021-00167-w.
11. Yi Myeung-Kyun Fall detection of the elderly using denoising LSTM-based convolutional variant autoencoder / Myeung-Kyun Yi, Kwangho Han, Seong Oun Hwang // IEEE Sensors Journal. – 2024. – Vol. 24, № 11. – P. 18556–18567. – DOI: https://doi.org/10.1109/JSEN.2024.3389808.
12. Liu Qian TS-MAE: A masked autoencoder for time series representation learning / Qian Liu, Junchen Ye, Haohan Liang, Leilei Sun, Bowen Du // Information Sciences. – 2025. – Vol. 689. – P. 121499. – DOI: https://doi.org/10.1016/j.ins.2024.121499.
13. Deldari Shohreh CroSSL: Cross-modal self-supervised learning for time-series through latent masking / Shohreh Deldari, Dimitris Spathis, Mohammad Malekzadeh, Fahim Kawsar, Flora D. Salim, Akhil Mathur // Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM). – 2024. – P. 152–160. – DOI: https://doi.org/10.1145/3616855.3635795.
14. Su Kunming RI-MAE: Rotation-invariant masked autoencoders for self-supervised point cloud representation learning / Kunming Su, Qiuxia Wu, Panpan Cai, Xiaogang Zhu, Xuequan Lu, Zhiyong Wang, Kun Hu // Proceedings of the 39th AAAI Conference on Artificial Intelligence. – 2025. – P. 7015–7023. – DOI: https://doi.org/10.1609/aaai.v39i7.32753.
15. Tuli Shreshth TranAD: Deep transformer networks for anomaly detection in multivariate time series data / Shreshth Tuli, Giuliano Casale, Nicholas R. Jennings // Proceedings of the VLDB Endowment. – 2022. – Vol. 15, № 6. – P. 1201–1214. – DOI: https://doi.org/10.14778/3514061.3514067.
16. Khatun Mst. Alema Deep CNN-LSTM with self-attention model for human activity recognition using wearable sensor / Mst. Alema Khatun, Mohammad Abu Yousuf, Sabbir Ahmed, Md. Zia Uddin and others // IEEE Journal of Translational Engineering in Health and Medicine. – 2022. – Vol. 10. – P. 1–16. – https://doi.org/10.1109/JTEHM.2022.3177710.
17. Ursul Ivan Elderly fall detection using unsupervised transformer model / Ivan Ursul // Electronics and Information Technologies. – 2024. – № 26. – DOI: https://doi.org/10.30970/eli.26.7.
18. Cheng Dongzhou MaskCAE: Masked convolutional autoencoder via sensor data reconstruction for self-supervised human activity recognition / Dongzhou Cheng, Lei Zhang, Lutong Qin, Shuoyuan Wang, Hao Wu, Aiguo Song // IEEE Journal of Biomedical and Health Informatics. – 2024. – Vol. 28, № 5. – P. 2687–2698. – DOI: https://doi.org/10.1109/JBHI.2024.3366228.
19. Avdan Goksu IMU-Trans: Imputing missing motion capture data with unsupervised transformers / Goksu Avdan, Sinan Onal, Chao Lu // Neural Computing and Applications. – 2025. – Vol. 37, № 7. – P. 5699–5717. – DOI: https://doi.org/10.1007/s00521-024-10892-9.
20. Ursul Ivan Benchmarking of transformer-based architectures for fall detection: a comparative study / Ivan Ursul // Technology Audit and Production Reserves. – 2025. – Vol. 3, № 2 (83). – P. 62–70. – DOI: https://doi.org/10.15587/2706-5448.2025.329398.
21. Ursul Ivan A balanced big dataset for sensor-based fall detection: enhancing model accuracy and robustness / Ivan Ursul // International Journal of Computing. – 2025. – Vol. 24, № 2. – P. 407–419. – DOI: https://doi.org/10.47839/ijc.24.2.4025.
22. Ursul Ivan Sensor-based fall detection dataset with 8,953 activities from 29 subjects [Electronic resource] / Ivan Ursul. – 2025. – Figshare. DOI: https://doi.org/10.6084/ m9.figshare.28287482.

Article: received 04.02.2026
revised 25.02.2026
printing adoption 04.03.2026

НЕЗАЛЕЖНЕ ВІД СУБ'ЄКТА ВИЯВЛЕННЯ ПАДІНЬ БЕЗ УЧИТЕЛЯ ЗА ДОПОМОГОЮ МАСКОВАНИХ ТРАНСФОРМЕРІВ ТА ОЦІНЮВАННЯ НА ОСНОВІ ЕНЕРГІЇ–РЕКОНСТРУКЦІЇ

І. Урсул

Львівський національний університет імені Івана Франка,
вул. Університетська 1, Львів, 79000, Україна,
e-mail: ivan.ursul@lnu.edu.ua

Представлено незалежний від суб'єкта детектор падінь без учителя для носимих інерціальних даних. Метод використовує маскований трансформер-кодувальник, натренований виключно на повсякденних активностях (ADL), з двома оцінками на етапі застосування моделі: помилкою маскованої реконструкції та прихованою енергією класового маркера. Оцінки стандартизуються за контрольними ADL і об'єднуються як S_α; модель екстремальних значень (EVT), побудована на верхніх 5% хвоста розподілу, визначає операційний поріг для фіксованого рівня хибних тривог. Проста часова постобробка (медіанне згладжування, гістерезис, рефрактерний інтервал) перетворює оцінки у виявлені події. Експерименти на наборі з 29 суб'єктів (8 953 сесії, 100 Гц; 6 каналів руху) за протоколом виключення одного суб'єкта (LOSO) показали об'єднаний ROC–AUC 0,985±0,008 та PR–AUC 0,955±0,012 для комбінованої оцінки, що перевищує варіанти лише з реконструкцією або лише з енергією. При цільовому рівні 0,5 хибних тривог на годину чутливість виявлення подій становить 95,8%±2,6% із середньою затримкою 1,72±0,28 с. Аналіз окремих суб'єктів демонструє компактний розкид навколо цільового рівня хибних тривог та вищу чутливість порівняно з варіантами на основі одного показника. Помилка реконструкції залишається низькою на ADL і зростає на падіннях, що підтверджує збереження якості відновлення сигналу та придатність для оцінювання аномалій. Запропонований підхід не потребує міток падінь для навчання чи калібрування.

Ключові слова: виявлення падінь, виявлення аномалій без учителя, маскований трансформер, енергетичне оцінювання, порогове калібрування EVT, міжсуб'єктне узагальнення (LOSO).