Clinical Utility

A Real-Time Computer Vision Framework for Skeletal Anomaly Detection Integrating Gaussian Filtering, LSTM, and XGBoost

Allied Medical Research Journal

Vol:4 No.4: 79-87

© The Author (s) 2026

Ali Hashir Rana[1], Muhammad Hussain[1], Momal Rana[1], Ausaf Ahmed[1], Ashel Aftab[1], S. Atruba Feroze[1]

[1]*Department of Software Engineering, Iqra University, Karachi, Pakistan*

ABSTRACT

Background: Real-time video anomaly detection in physical surveillance is still limited by the computational cost and illumination noise of traditional 3D Convolutional Neural Networks (CNNs). Detecting sudden kinematic changes, such as a physical struggle, a sudden loss of balance or an unexpected fall, requires lightweight solutions that run on edge hardware, both in security settings and in remote tele-rehabilitation monitoring.

Methods: This work proposes a vision-based skeletal kinematics approach that detects human keypoints in real time with YOLOv8-Pose. Frames are first smoothed with a 2D Gaussian filter, and keypoint coordinates are then converted into kinematic measures: joint velocities, acceleration peaks and joint angles. Temporal patterns were analyzed with an unsupervised Bi-directional Long Short-Term Memory (Bi-LSTM) Autoencoder and a statistical XGBoost classifier. Both models were evaluated on the Custom Binary Surveillance (CBS) dataset, which contains 2,268 standardized 150-frame windows extracted from public and proprietary surveillance videos.

Results: On the hold-out test set, the Bi-LSTM Autoencoder achieved an accuracy of 89.3% (95% CI 86.7% to 91.9%) and an anomaly recall of 91.3%. The full pipeline ran at 55 frames per second (18 ms latency) on an NVIDIA T4 GPU.

Conclusion: Extracting compact 1D kinematic signals from high-dimensional video pixels greatly reduces computation. This lightweight architecture suits edge deployment and may support patient fall screening and movement monitoring, pending validation in clinical settings.

Keywords: *Accidental Falls; Artificial Intelligence; Deep Learning; Telerehabilitation; Video Recording*

Received: June 11th, 2026; **Revised:** August 26th, 2026; **Accepted:** September 8th, 2026

Corresponding Email: ali.63971@iqra.edu.pk

DOI: https://doi.org/10.59564/amrj/04.04/010

INTRODUCTION

The current surveillance technologies are able to generate streams of video, but operators often get tired after 20 minutes of monitoring multiple cameras, making manual observation impossible in transit hubs and urban areas. Video Anomaly Detection (VAD) has therefore become an indispensable computer vision problem[1,2], which has been developed to automatically detect abnormal behaviors, such as an assault, a fall or a theft, in real time, so that a proactive response can be made.

Traditionally, automatic video anomaly detection evolved first from hand-crafted heuristics like Optical Flow and Spatiotemporal Gradients[3,4] and then to dense-pixel deep learning methods, such as 3D-CNN (C3D, I3D[5]) and temporal autoencoder[6-10]. While these architectures can provide rich spatiotemporal representations, processing volumetric RGB blocks proves to be extremely expensive by FLOPs and the requirement of enterprise-class, cluster-type GPU architectures is too high to be implemented at the edge. In addition, dense-pixel networks are always susceptible to environmental artifacts such as compression noise, varying illumination and dynamic background noise[11,12].

This study abstracts visual scenes by simulating their physics of human motion, departing from the dense pixels that have been used in previous studies to overcome these computational limitations. The pipeline extracts the anatomical key point coordinates $(x, y)$and isolates human actors through the single-pass pipeline of YOLOv8-Pose estimation network[13,14].

This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0), which permits others to share, copy, redistribute, and adapt the work for non-commercial purposes, provided the original author(s) and source are credited appropriately. Further details are available on the official AMRJ Open Access policy page: https://ojs.amrj.net/index.php/1/14.

The inter-frame velocity ($v = \Delta p/\Delta t$) is calculated and abrupt velocity spikes are detected to build a compact mathematical derivation of human dynamics. The abrupt physical anomalies (strikes, sudden collapses, physical fights, snatch movements) are kinematic variables that are abrupt in terms of the momentum shift, hence they can be used as a low dimensional input vector of a sequential deep learning model with light processing demand. Biomechanically, there are common ideas in this framework with automated movement analysis. Kinematic indicators of a fall event, a balance disturbance or a physical trauma include rapid joint deceleration, atypical angular flexion and sudden spatial displacement[15,16].
The architecture transforms the visual scene into a kinematic vector, which allows for the identification of abnormal physical movements and may be less affected by variations in lighting, clothing or the use of walking aids in clinical and home settings, although this was not tested on patient populations in the present study.

Existing VAD benchmarks are based on abstract, multi-class action data which are more oriented towards longer, macro-scale events (such as explosions or longer episodes of fighting). These schemas dilute the need to focus on the classifiers of threats that are local and fast-moving, such as sudden snatching. To address this, this paper proposes the Custom Binary Surveillance (CBS) Dataset, which combines standardized public clips[17] with proprietary 30-FPS video clips and structures the entire dataset into a binary (Normal vs. Anomaly) schema to ensure the network learns underlying motion physics and not background artefacts.

Research Contributions

The main contributions of this work are four-fold: (1) Design of a lightweight anomaly detection pipeline that performs edge deployment of dense-pixel convolution to single-pass topological keypoint extraction for real-time operation on resource constrained devices; (2) Formulation of a physics-informed temporal feature engineering paradigm by using joint velocity, angular trajectory, and acceleration impulses, to ensure threat sensitivity while allowing detection of anomalous physical movements even under strongly varying environmental lighting; (3) Creation and evaluation of the Custom Binary Surveillance (CBS) dataset, which combines standardized benchmark actions with proprietary real-world localized threats for assessing threat sensitivity and robust generalization; and (4) Extensive empirical evaluation showing accuracy competitive with conventional spatiotemporal deep learning baselines at substantially lower computational cost (55 FPS).

Topological Kinematics and Movement Science

Recent paradigms focus on compact topological extraction[18-20], in order to remove visual noise from the data without going to the quadratic complexity $O(N^2)$that is usually required for heavy Graph Transformers or multi-stage pipelines like OpenPose with ST-GCN.

Markerless single-pass estimators, such as YOLOv8-Pose, are able to localize 2D key points at high frame rates. The ability to abstract the trajectories of the coordinates into explicit physical kinematics has direct application to clinical movement science, patient fall risk assessment and monitoring rehabilitation remotely. In clinical movement analysis and tele-rehabilitation markerless motion capture is becoming more popular. Stenum et al.[15] have shown that 2D video-based pose estimation is able to accurately extract sagittal kinematic gait parameters with reliability, when compared to marker-based motion capture for step time and joint kinematics. To ensure privacy and without sending raw videos, Chang et al.[16] demonstrated that pose estimation performed on local embedded devices can accurately detect fall postures (up to 98.1% accuracy). For automated fall screening, Raza et al.[21] compared pose-estimation-based pipelines, from classical machine-learning classifiers to vision transformers, and reported strong fall-detection results on the UR-Fall, UP-Fall and Le2i datasets. Chang et al.[16] additionally used spine-angle deviation and head acceleration as pre-filters before an LSTM classifier. These precedents suggest that low-order kinematic quantities such as velocity, acceleration and joint angles can describe acute movement deviations[16,21].

METHODOLOGY

The proposed light framework is different from dense-pixel networks, which are suitable for edge deployment, by evaluating the human topological kinematics in real time. The pipeline can be broken down into 4 simple stages as shown in Figure 1: (1) Image Preprocessing and Noise Reduction, (2) Single-Pass Topological Extraction using YOLOv8-Pose, (3) Kinematic Feature Engineering, and (4)

Temporal Sequence Classification. All recordings were handled under institutional data-governance procedures.

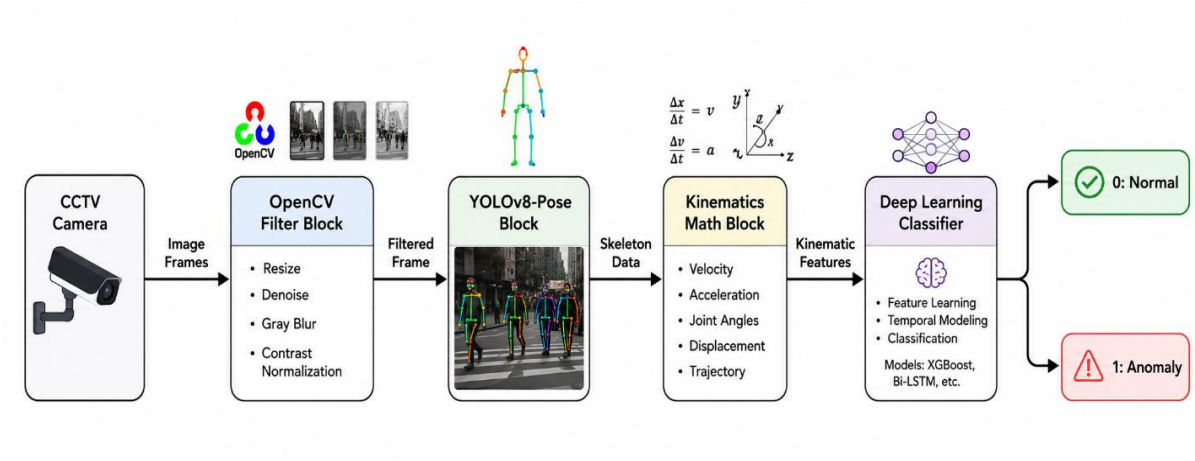


Fig 1: End-to-end architecture of the proposed kinematic anomaly detection framework, showing the progression from pixel-based filtering to topology-driven

Image Preprocessing and Noise Reduction

Surveillance streams are often compressed, have sensor noise and poor illumination. The subsequent keypoint detection process is performed by combining the spatial enhancement process with the input frames for stability. A first step is to apply Non-Local Means (NLM) denoising to the search windows with h = 15 to remove the distributed photonic sensor grain, while maintaining high contrast edge boundaries. Frames are then convolved with a 2D smoothing filter (Gaussian filter), to reduce high-frequency variations in the area:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

where σ dictates the blur distribution. As shown in (Figure 2), convolving the input frames suppresses high-frequency background noise while strictly preserving anatomical structural boundaries.

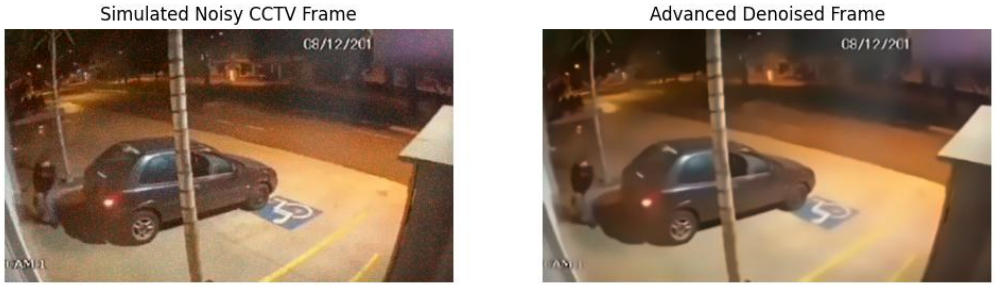


Fig 2: Qualitative comparison of raw nighttime surveillance footage (left) and the stabilized output following the applied 2D Gaussian smoothing operator (right).

Single-Pass Topological Extraction

Unlike multi-stage pipelines that induce severe latency by cascading separate bounding-box detectors and tracking models, single-pass YOLOv8-Pose unifies human detection and keypoint extraction in a single forward inference pass[22-26]. For each frame at time t, the network isolates human subjects and extracts the COCO topology of $K = 17$ anatomical keypoints:

$$J_t = \{(x_{i,t}, y_{i,t}, c_{i,t})\}_{i=1}^{17}$$

where $(x_{i,t}, y_{i,t})$ represent Cartesian joint coordinates and $c_{i,t}$ denotes detection confidence. Joints occluded below confidence threshold τconf are temporally imputed via linear interpolation. Abstracting frames into

discrete coordinates which removes the environmental pixel noise, significantly reducing the downstream memory footprints.

Kinematic Feature Engineering

Physical anomalies such as strikes, falls, or sudden snatching gives out violent changes in momentum and joint geometry that static coordinates fail to capture[16,21]. We instead converted raw coordinates into three distinct and dynamic kinematic features:

Joint Velocity computation

Evaluated via Euclidean displacement across inter-frame temporal step Δt:

$$v_{i,t} = \frac{\sqrt{(x_{i,t}-x_{i,t-1})^2+(y_{i,t}-y_{i,t-1})^2}}{\Delta t}$$

Acceleration Spikes

Derived as the first temporal derivative of velocity to calculate sudden physical impulses in the frames.

$$a_{i,t} = \frac{v_{i,t}-v_{i,t-1}}{\Delta t}$$

As plotted in (Figure 3), normative movement generates a flat acceleration profile, meanwhile an anomalous event registers an acute, high-amplitude spike.

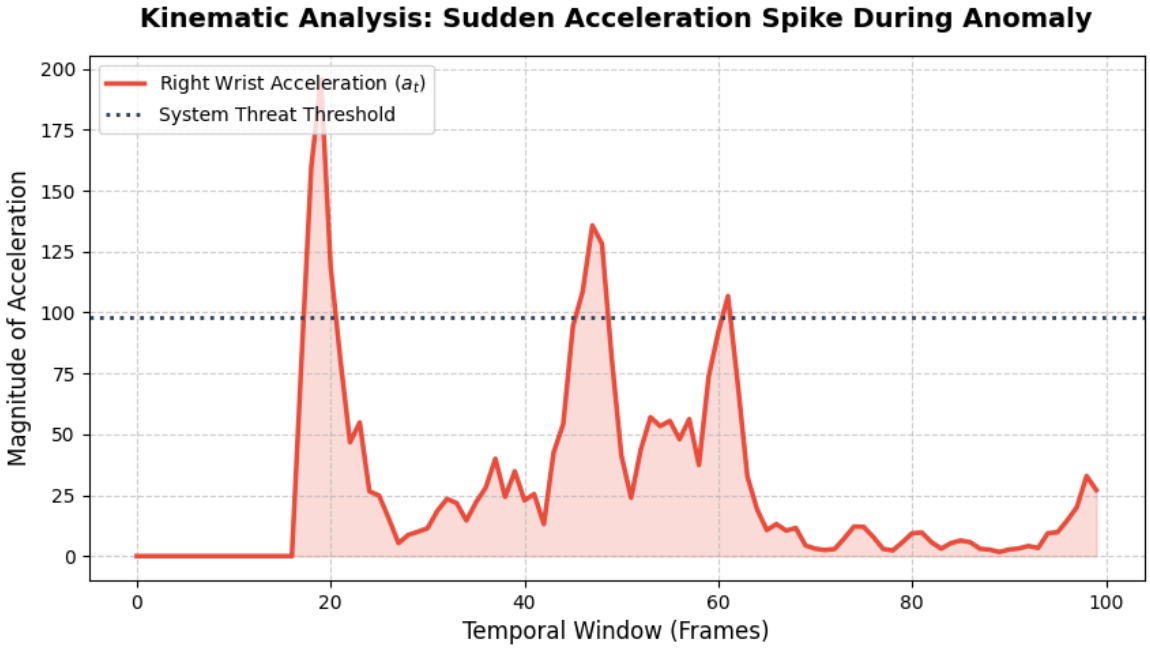


Figure 3: Temporal graph of wrist-joint acceleration $a_{i,t}$ over the first 100 frames of a 150-frame window, highlighting the prominent kinematic spike characteristic of a sudden anomalous event.

Angular Kinematics:

To detect postural collapse or unnatural limb contortion in the frames, the interior angle θ_t between adjacent joint vectors u and v is determined by the usage of dot product:

$$\theta_t = arccos\left(\frac{u \cdot v}{\|u\|\|v\|}\right)$$

The engineered kinematics are concatenated into a compact, low-dimensional vector F_t per frame, that completely bypasses the RGB pixel processing.

Temporal Sequence Modeling and Classification

Kinematic vectors F_t are calculated over sliding temporal windows of length $W = 150$ frames (5 seconds at 30 FPS), forming a sequence matrix $X = [F_1, F_2, \dots, F_W]^T \in R^{W \times D}$ $(D = 12)$. Two distinct paradigms are evaluated as follows:

Tabular XGBoost via Statistical Compression

Matrix X is flattened into static mathematical profiles across five temporal metrics:

$$\phi(X_{:,j}) = [max(X_{:,j}), min(X_{:,j}), \mu(X_{:,j}), \sigma(X_{:,j}), median(X_{:,j})] \quad (6)$$

This outputs a 60-dimensional vector $X_{flat} \in R^{60}$ processed by an eXtreme Gradient Boosting (XGBoost) ensemble of K additive regression trees[27,28]:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(X_{flat,i}), \quad f_k \in F \quad (7)$$

Unsupervised Bi-Directional LSTM Auto encoder

For unsupervised sequence modeling, a Bidirectional Long Short-Term Memory Autoencoder (Bi-LSTM-AE) is applied following principles of recurrent autoencoders for temporal anomaly detection[29]. Forward ($\vec{h}_t$) and backward ($h^{\leftarrow}{}_t$) recurrent layers are used to capture bidirectional temporal dependencies, giving the following latent representation:

$$h_C = [\vec{h}_W \oplus h^{\leftarrow}{}_1] \in R^{2\times H} \quad (8)$$

The decoder reconstructs the kinematic sequence

$$\hat{X} = [\hat{F}_1, \dots, \hat{F}_W]^T \quad (9)$$

Anomalies are registered when the Reconstruction Mean Squared Error (MSE) exceeds a reconstruction-error threshold τ_{rec}:

$$E(X) = \frac{1}{W \cdot D} \sum_{t=1}^{W} \sum_{j=1}^{D} (F_{t,j} - \hat{F}_{t,j})^2 \quad (10)$$

τ_{rec} was set at the 95th percentile of reconstruction error on the normal training windows and was not tuned on the test set.

Dataset Engineering

To address the drawbacks of public datasets that typically only contain footage for long events, we created a new benchmark, called Custom Binary Surveillance (CBS), which consists of public archives17 and proprietary 30-FPS surveillance video. The data represents a range of real-world scenarios in commercial and residential buildings of Karachi (e.g. Gulistan-e-Jauhar and Scheme 33). Ground truth is flattened into a simple binary classification: Normal ($Y = 0$, walking, standing, sitting) and Anomaly ($Y = 1$, assaults, falls, sudden high-velocity snatching). Continuous stream was divided into non-overlapping uniform windows of $W = 150$ frames, mapping raw stream V into discrete arrays $S = \{w_1, \dots, w_N\}$ where $N = \lfloor L/W \rfloor$.

The normalized CBS dataset is divided into training and test datasets, with 77.7% of the windows used for training and 22.3% of the windows used for testing (Table 1).

Table 1. Composition and Distribution of the CBS Dataset

Data Origin	Composition				
	Classification Schema	Label	Temporal Format	Training Split (77.7%)	Testing Split (22.3%)
Public Baseline	Walk, Sit, Stand	Normal (0)	150-Frame Arrays	400 Windows	110 Windows
Proprietary Custom	Unconstrained CCTV	Normal (0)	150-Frame Arrays	223 Windows	63 Windows
Public Baseline	Hit, Fall, Grab, etc.	Anomaly (1)	150-Frame Arrays	840 Windows	232 Windows
Proprietary Custom	Phone-Snatching	Anomaly (1)	150-Frame Arrays	300 Windows	100 Windows
Total	–	–	–	**1,763**	**505**

A. Ethical Governance and Institutional Exemption

This study was conducted in accordance with observational research ethics and the principles of the Declaration of Helsinki. As a non-interventional computer vision and engineering study involving no clinical trials, medical interventions, or human-subject experimentation, formal clinical trial registration and medical IRB review were not applicable. Administrative permission and data-sharing consent for non-commercial academic use were formally secured directly from the residential apartment society's union office (for the Gulistan-e-Jauhar CCTV footage) and private property owners (for the mobile-recorded snatching clips), alongside publicly available benchmark archives. To ensure strict privacy protection for individuals passing through public or semi-public spaces, all subjects were completely anonymized via automated facial obscuration and topological skeletal abstraction; only dimensionless coordinate trajectories were stored, entirely stripping recognizable biometric profiles.

B. Implementation Details and Evaluation Metrics

The training was performed on an Intel Xeon CPU (2.2 GHz) with an NVIDIA T4 GPU (16 GB VRAM) and PyTorch 2.x, along with OpenCV. YOLOv8-Pose used a confidence threshold $\tau_{conf} = 0.15$ and NMS threshold $\tau_{NMS} = 0.45$. A 1D Gaussian filter ($\sigma = 2$) was used to smooth the temporal coordinates.

XGBoost was trained with $K = 300$ estimators, tree depth of 6, learning rate $\eta = 0.05$, and subsampling of 0.8, also regularized LogLoss:

$$L_{XGB} = -\frac{1}{N}\sum_{i=1}^{N} [y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)] + \sum_{k=1}^{K} \Omega(f_k) \quad (11)$$

The Bi-LSTM Autoencoder (latent dimension $H = 64$, 2 stacked layers) was trained using Adam ($\alpha = 0.001$, L_2 weight decay $\lambda = 10^{-3}$, 30% spatial dropout) for 50 epochs (batch size 32) with ReduceLROnPlateau scheduler.

The C3D and ST-GCN baselines were trained and tested on the same split as the proposed models (1,763 training windows and 505 hold-out windows), in a single run. The C3D baseline operated on the RGB frames of each 150-frame window, whereas the ST-GCN baseline ("Raw") operated on the raw YOLOv8-Pose keypoint coordinates of each window, without the engineered kinematic features. The training hyperparameters of the baselines were not recorded and are therefore not reported.

The accuracy metric (Eq. 12), precision (Eq. 13), recall (Eq. 14), and F1-Score (Eq. 15) are used to measure performance and were calculated using a standard confusion matrix:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (12)$$

$$Precision = \frac{TP}{TP + FP} \quad (13)$$

$$Recall = \frac{TP}{TP + FN} \quad (14)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (15)$$

RESULTS AND DISCUSSION

The abstracted feature representation produced by kinematic (i.e. environment noise) successfully enabled both the classifiers under evaluation to correctly identify anomalous behavior. The convergence to models was smooth, while the dynamic learning rate scheduling and L_2 regularization was stable.

A. Quantitative Performance and Baseline Comparison.

Table 2 shows a summary of the comparative performance of the kinematic models with the

Table 2. Overall Performance Metrics on the CBS Dataset

Model Architecture	Metrics				
	Accuracy	95% CI (Accuracy)	Precision	Recall	F1-Score
Baseline 3D-CNN (C3D)	81.2%	±3.4%	—	—	—
Baseline ST-GCN (Raw)	88.5%	±2.7%	—	—	—
Kinematic XGBoost	82.2%	±3.3%	85.2%	88.3%	86.7%
Bi-LSTM Autoencoder (Proposed)	89.3%	±2.6%	92.4%	91.3%	91.8%

dense-pixel and graph baselines on the hold-out CBS test set.

* Performance results, with 95% Confidence Intervals (CI) computed using bootstrap resampling of per-window correctness, are evaluated only on the hold-out test set (505 windows: 332 Anomaly, 173 Normal; 22.3% hold-out). Windows classified correctly: C3D 410, ST-GCN 447, XGBoost 415, Bi-LSTM 451. Precision, recall and F1-score are for the Anomaly class (positive class). Confusion-matrix counts are given in the text for the proposed models; — = not available for the baseline models (per-window predictions were not archived).

By extracting localized momentum extremes, statistical XGBoost's F1-score was 86.7%. The unsupervised Bi-LSTM Autoencoder achieved a higher recall (91.3%) than XGBoost (88.3%) and an overall accuracy of 89.3% (95% CI ±2.6%), comparable to ST-GCN (88.5%) given the overlapping confidence intervals. On the 505 test windows, the Bi-LSTM Autoencoder correctly detected 303 of the 332 anomalous windows (TP = 303, FN = 29) and correctly passed 148 of the 173 normal windows (TN = 148, FP = 25), a false-alarm rate of 14.5%. The XGBoost classifier detected 293 of 332 anomalous windows (TP = 293, FN = 39) with 51 false alarms (FP = 51, TN = 122). The key factor in automated surveillance is to maximize a computer's sensitivity to detecting a true physical threat, as this has far more serious consequences in the real world, than false alarms. Compared with heavy spatial-temporal baselines including ST-GCN, the Bi-LSTM Autoencoder, being unsupervised, does not require labelled anomalous examples at training time.

B. Computational Efficiency and Edge Viability.

Inference speed must surpass camera frame rate in order to be real-time. This classification model has only 0.17 million parameters, which is a huge reduction in the number of parameters needed when the high-dimensional video frames are compressed into 1D kinematic feature vectors F_t.

An end-to-end pipeline (YOLOv8 pose extraction, kinematic calculation and classification) runs at 55 FPS on the NVIDIA T4 GPU, easily meeting the 30 FPS requirement of live CCTV feeds. The standalone sequence-classifier ran the 150-frame windows in 5.07 ms on a normal CPU (without GPU acceleration), which is 197.2 fps. The efficiency validates the correctness of this kinematic abstraction, eliminating latency and enterprise level hardware limitations that are generally associated with 3D-CNNs.

C. Real-Time Deployment Simulation

A simulation of continuous deployment was carried out to verify the viability of the system with the use of a First-In-First-Out (FIFO) temporal buffer. The coordinates of the joints were then mapped frame-by-frame to kinematic vectors and the 150 frames in the buffer were filled up before the evaluation of the tensor was performed with respect to the reconstruction error threshold ($\tau_{rec} = 0.4662$).

The system operated efficiently through the switching between states indicating "Normal" (green) and "Anomaly Detected" (red) as illustrated in (Figure 4), switching from one state to another the moment an acute physical momentum occurs. This simulation will verify the pipeline behavior in continuous streaming environment with no memory buffer overflows and no latency drift issues.



Fig 4: Real-time qualitative inference results demonstrating dynamic system behavior during live deployment simulation.

D. Current Limitations and Trajectories of Translational Healthcare.

The proposed kinematic approach yields 55 FPS of inference throughput with a good classification accuracy on Custom Binary Surveillance (CBS) dataset but there are certain drawbacks that can be further pursued. First, there was no evaluation of patients or clinical populations as this investigation was based solely on observational, outdoor, and indoor surveillance recordings, so the diagnostic utility of this framework in formal health care settings has not been established. Second, the current evaluation focused on the binary CBS indicator that selects the sudden, high velocity physical events. This study has not yet assessed performance on general public surveillance tasks (e.g., UCF-Crime[30], ShanghaiTech Campus, UCSD Ped2) where there are more complex multi-class anomalies, which will be explored in future work. Third, the reconstruction-error threshold τ_{rec} was set at the 95th percentile of reconstruction error on the normal training windows, which by construction implies roughly 5% false alarms on data resembling the training windows. The false-alarm rate on the hold-out set was higher (14.5%; 25 of 173 normal windows), which may reflect a distribution shift between the training windows and the test windows, including the unconstrained CCTV footage. Threshold calibration on normal data from the target camera environment would therefore be needed before deployment. This topological formulation will be extended to biomechanics in clinical applications, geriatric tele-monitoring and fall-risk assessment in future studies. Key priorities will be to validate the pipeline against standard functional mobility protocols, such as Timed Up and Go (TUG) test, Berg Balance Scale (BBS) and with supervised clinical trials with patient cohorts. Based on these evaluations, it will be determined if lightweight keypoint tracking is a viable, accessible and camera-based evaluation tool that meets real-time edge processing constraints.

CONCLUSION

This paper introduces a lightweight skeletal kinematics framework for real-time video anomaly detection to overcome the computational burden and environmental noise of the existing 3D-CNNs. The pipeline transforms the video streams into topological coordinate trajectories using the YOLOv8-Pose, and calculates the basic physical derivatives such as velocity, angular deviation, and acceleration impulse within a small 1D feature representation. This representation was found to be effective through experimental evaluation on the Custom Binary Surveillance (CBS) dataset, where tabular XGBoost achieves an F1-score of 86.7% and the unsupervised Bi-LSTM Autoencoder achieves a 91.3% anomaly recall rate and an accuracy of 89.3%. Beyond municipal edge surveillance, converting dense pixel data into lightweight kinematic vectors offers a computationally efficient basis for remote movement analysis, with potential applications in automated fall detection and tele-rehabilitation monitoring once clinically validated.

Acknowledgment

The authors thank the Department of Software Engineering, Iqra University, Karachi, for providing the resources and environment needed for this study.

Author Contributions

AHR: Conception and design, data collection, model development, and manuscript writing
MH: Data collection, data preprocessing, and kinematic feature engineering
MR: Dataset annotation, data analysis, and interpretation
AAh: Model training, baseline experiments, and manuscript editing
AAf: Literature review, results visualization, and manuscript editing
SAF: Supervision, critical revision, and final approval
All authors read and approved the final manuscript.

Conflict of interest: The authors declare no conflict of interest.

Funding: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Data availability: The public component of the CBS dataset is available from the Kaggle repository cited in the text. The proprietary CCTV footage cannot be shared for privacy reasons; the derived, anonymized keypoint trajectories are available from the corresponding author on reasonable request.

Ethical approval: Not applicable. This was a non-interventional engineering study with no clinical procedures and no direct contact with the individuals recorded.

REFERENCES

1. Alarfaj M, Pervaiz M, Ghadi YY, al Shloul T, Alsuhibany SA, Jalal A, et al. Automatic Anomaly Monitoring in Public Surveillance Areas. Intell Autom Soft Comput. 2023;36(1):1141–1158. https://doi.org/10.32604/iasc.2023.027205
2. Cabanillas Carbonell M, Sallari Rivera J, Santivañez Muñoz J. Artificial intelligence in video surveillance systems for suspicious activity detection and incident response: A systematic review. Adv Sci Technol Res J. 2025;19(3):389–405. https://doi.org/10.12913/22998624/196795
3. Xiao J, Ji G. Divide and Conquer in Video Anomaly Detection: A Comprehensive Review and New Approach. In: 2023 China Automation Congress (CAC).

IEEE; 2023:8553–8558. https://doi.org/10.1109/CAC59555.2023.10451821
4. Kong Y, Fu Y. Human action recognition and prediction: A survey. Int J Comput Vis. 2022;130(5):1366–1401. https://doi.org/10.1007/s11263-022-01594-9
5. Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the Kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017:6299–6308. https://doi.org/10.1109/CVPR.2017.502
6. Abdalla M, Javed S, Al Radi M, Ulhaq A, Werghi N. Video anomaly detection in 10 years: a survey and outlook. Neural Comput Appl. 2025;37(22):26321–26364. https://doi.org/10.1007/s00521-025-11659-8
7. Zhu L, Wang L, Raj A, Gedeon T, Chen C. Advancing Video Anomaly Detection: A Concise Review and a New Dataset. Adv Neural Inf Process Syst. 2024;37:2856–2875. https://doi.org/10.52202/079017-2856
8. Duong HT, Le VT, Hoang VT. Deep Learning-Based Anomaly Detection in Video Surveillance: A Survey. Sensors. 2023;23(11):5024. https://doi.org/10.3390/s23115024
9. Wu P, Pan C, Yan Y, Pang G, Yan Q, Wang P, et al. Deep learning for video anomaly detection: A review. IEEE Trans Neural Netw Learn Syst. 2026;37(7):3010-3030. https://doi.org/10.1109/TNNLS.2025.3647892
10. Hao Y, Li J, Wang N, Wang X, Gao X. Spatiotemporal consistency-enhanced network for video anomaly detection. Pattern Recognit. 2022;121:108232. https://doi.org/10.1016/j.patcog.2021.108232
11. Singh R, Pal A, Mishra S, Chowdhury A. Enhancing Situational Awareness: Anomaly Detection Using Real-Time Video Across Multiple Domains. IEEE Access. 2025;13:73680–73696. https://doi.org/10.1109/ACCESS.2025.3560874
12. Betgeri S, Lodha P, Oak A, Parikh A, Lodha A. AlertCam-An Effective Anomaly Detection System using Computer Vision. In: 2025 International Conference on Emerging Smart Computing and Informatics (ESCI). IEEE; 2025:1–6. https://doi.org/10.1109/ESCI63694.2025.10988334
13. Jocher G, Chaurasia A, Qiu J. YOLO by Ultralytics. GitHub repository; 2023. Available from: https://github.com/ultralytics/ultralytics
14. Vijayakumar A, Vairavasundaram S. YOLO-based Object Detection Models: A Review and its Applications. Multimed Tools Appl. 2024;83(28):83535–83574. https://doi.org/10.1007/s11042-024-18872-y
15. Stenum J, Rossi C, Roemmich RT. Two-dimensional video-based analysis of human gait using pose estimation. PLoS Comput Biol. 2021;17(4):e1008935. https://doi.org/10.1371/journal.pcbi.1008935
16. Chang WJ, Hsu CH, Chen LB. A pose estimation-based fall detection methodology using artificial intelligence edge computing. IEEE Access. 2021;9:129965–129976. https://doi.org/10.1109/ACCESS.2021.3113824
17. Nield J. CCTV Action Recognition Dataset. Kaggle Repository; 2024. Available from: https://www.kaggle.com/datasets/jonathannield/cctv-action-recognition-dataset
18. Mishra PK, Mihailidis A, Khan SS. Skeletal Video Anomaly Detection Using Deep Learning: Survey, Challenges, and Future Directions. IEEE Trans Emerg Top Comput Intell. 2024;8(2):1073–1085. https://doi.org/10.1109/TETCI.2024.3358103
19. Krishnan SR, Amudha P. Spatio-temporal anomaly detection for real-time video surveillance using SpatioGuard-YOLO framework. Multimed Tools Appl. 2026;85(4):542–563. https://doi.org/10.1007/s11042-026-21696-7
20. Anantathanavit R, Suthirat S, Su PC. AIGuard: Anomaly Detection in Surveillance Videos with YOLOv8. In: 2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE; 2025:2008–2013. https://doi.org/10.1109/APSIPAASC65261.2025.11249213
21. Raza A, Yousaf MH, Ahmad W, Velastin SA, Viriri S. Human fall detection using pose estimation: From traditional machine learning to vision transformers. Eng Appl Artif Intell. 2025;143:109809. https://doi.org/10.1016/j.engappai.2024.109809
22. Nasir R, Jalil Z, Nasir M, Alsubait T, Ashraf M, Saleem S. An enhanced framework for real-time dense crowd abnormal behavior detection using YOLOv8. Artif Intell Rev. 2025;58(7):202. https://doi.org/10.1007/s10462-025-11206-w
23. Shamim R, Alfurhood BS, Agarwal T, Mallik BB. YOLOv8 for Anomaly Detection in Surveillance Videos. In: Mathematical Modeling for Computer Applications. Wiley; 2024:245–262. https://doi.org/10.1002/9781394248438.ch18
24. Shi T. Anomaly Detection in Logistics Warehouses Based on YOLOv8. In: 2024 3rd International Conference on Smart City Challenges & Outcomes for Urban Transformation (SCOUT). IEEE; 2024:24–30. https://doi.org/10.1109/SCOUT64349.2024.00016
25. Natha SAS, Siraj M, Alsaif SA, Farooq F, Shah A, Mahmud M. A fusion approach of YOLOv8 and CNN-Transformer for End-to-End road anomaly detection. Sci Rep. 2025;15(1):45341. https://doi.org/10.1038/s41598-025-29718-4
26. Aishwarya K, Sowmiya V, Suthan R, Deboral C, Tanusri S, Sahana V. A Deep Learning Strategy for Abnormal Object Detection by YOLOv8 Model. In: 2024 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS). IEEE; 2024:1–5. https://doi.org/10.1109/ICPECTS62210.2024.10780243
27. Tsanaullailla L, Hamami F, Suakanto S. Smart Surveillance in Retail Stores: Detection, Classification, and Documentation of Suspicious Behavior Utilizing YOLOv11 and XGBoost. In: 2025 4th International Conference on Electronics Representation and Algorithm (ICERA). IEEE; 2025:127–132. https://doi.org/10.1109/ICERA66156.2025.11087345
28. Shamnath, Belwal M. Human Suspicious Activity Detection Using Ensemble Machine Learning Techniques. In: 2022 2nd International Conference on Intelligent Technologies (CONIT). IEEE; 2022:1–5. https://doi.org/10.1109/CONIT55038.2022.9848183
29. Park D, Hoshi Y, Kemp CC. A multimodal anomaly detector for robot-assisted feeding using an LSTM-based variational autoencoder. IEEE Robot Autom Lett. 2018;3(3):1544–1551. https://doi.org/10.1109/LRA.2018.2801475
30. Sultani W, Chen C, Shah M. Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018:6479–6488. https://doi.org/10.1109/CVPR.2018.0067