An Efficient Lightweight CNN Autoencoder for Skeleton-Based Video Anomaly Detection

Mostafa Ibrahim Labib
Department of Computer Science, Higher Future Institute for Specialized Technological Studies, Egypt
mostafa.elkhalil@fa-hists.edu.eg (corresponding author)

Fatma Harby Mohamed
Department of Computer Science, Higher Future Institute for Specialized Technological Studies, Egypt
fatma.mohamed@fa-hists.edu.eg

Received: 24 March 2026 | Revised: 26 April 2026 | Accepted: 1 May 2026

 DOI: https://doi.org/10.48084/etasr.18897

ABSTRACT

Anomaly detection based on human motion has been an active area of research in computer vision, evolving from traditional handcrafted methods to deep learning–based representations. This paper proposes a lightweight Convolutional Neural Network (CNN) autoencoder designed for anomaly detection using skeleton image representations. Unlike prior Long Short-Term Memory (LSTM)-based approaches that depend on temporal coordinate sequences, the proposed approach encodes skeletal joint relations as 2D spatial maps, enabling convolutional learning of local and global structure. The model was trained exclusively on normal samples from structured datasets generated from pose-estimation outputs to reconstruct typical poses, with reconstruction error used as an anomaly score. Skeleton images are preprocessed before training using grayscale conversion, Gaussian blurring, and resizing to ensure consistency, reduce noise, and improve efficiency. Experimental results achieved an accuracy of 59.6%, precision of 61.0%, and recall of 95.5%, yielding an overall F1-score of 74.5%. These results demonstrate that the CNN-based autoencoder successfully identifies most anomalous poses while maintaining reasonable precision, validating the benefits of spatial learning and augmentation for skeleton image anomaly detection.

Keywords-Convolutional Neural Network (CNN) autoencoder; data augmentation; lightweight deep learning; quantization

I. INTRODUCTION AND RELATED WORK

Human activity analysis through skeleton-based modeling has emerged as an essential tool for intelligent video surveillance, behavior recognition, and public safety systems. In such applications, the early detection of abnormal movements—such as falls, unsafe postures, or aggressive behavior—is crucial for enabling rapid response and preventing harm [1]. Traditional video anomaly detection methods often depend on pixel-level information or handcrafted spatiotemporal descriptors, which are computationally expensive and highly sensitive to illumination, background noise, and occlusion. To overcome these limitations, skeleton-based representations have gained attention due to their compactness, interpretability, and robustness to environmental variability [2].

In order to capture temporal relationships between joint coordinates, previous research has mostly used sequence-based autoencoders using Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) architectures. Although these models showed potential, they had a number of significant drawbacks. First of all, their capacity to encode the spatial relationships between body joints was limited by functioning directly on 2D coordinate sequences, which resulted in poor generalization to unknown postures and camera views. Second, their ability to learn a variety of regular behavior patterns was hampered by the small dataset size and lack of augmentation approaches. Lastly, there was little quantitative support for the models' assertion that they were "lightweight" in terms of parameter count, inference time, and deployment viability [3].

This study presents an optimized Convolutional Neural Network (CNN) autoencoder that uses skeletal images rather than raw coordinate sequences for anomaly identification in order to overcome these problems. By converting the skeletal joints into spatially structured pictures, this method enables the network to better model body shape by utilizing convolutional feature extraction. The architecture significantly reduces computational complexity while maintaining spatial learning capability through the use of separable convolutional layers. The decoder reconstructs skeleton images using Conv2DTranspose layers, providing high-fidelity reconstruction with minimal overhead [4].

A major improvement over prior work lies in the incorporation of data augmentation at the image level., where the suggested model significantly increases the variety of normal samples by random rotation, translation, zoom, shear, and brightness transformations, improving robustness and recall. Overall, the proposed approach represents a significant step forward in skeleton-based anomaly detection, achieving a balanced trade-off between accuracy, efficiency, and scalability. By integrating spatial learning, augmentation-based generalization, and quantized deployment, this work establishes a robust foundation for future extensions to 3D skeleton data and hybrid CNN–LSTM frameworks for more complex temporal modeling [5].

Computer vision researchers have been actively studying anomaly detection in human motion, moving from manual techniques to deep learning-based representations. Current approaches fall into three general categories: pixel-based video methods, skeleton-sequence models, and image-based spatial learning frameworks. Video anomaly detection systems in pixel-based video methods mostly depended on motion trajectories, optical flow, or manually created spatiotemporal features to identify anomalous patterns. To find abnormalities in dense video streams, techniques such as Social Force Models, Spatiotemporal Interest Points, and 3D Convolutional Networks were used. Although these techniques captured extensive visual information, they were computationally costly and susceptible to noise from partial occlusions, lighting variations, and background clutter [6].

Recent deep architectures such as Convolutional LSTMs, 3D CNNs, and joint appearance-skeleton methods have improved spatiotemporal modeling, but their high complexity makes them unsuitable for real-time or resource-constrained environments. This limitation motivated the transition toward skeleton-based anomaly detection, which abstracts human motion into a compact, interpretable representation [7].

Skeleton-based representations minimize redundancy while maintaining motion semantics by describing the human body as a structured collection of joints and connections. Encoder-decoder LSTMs, for example, have been utilized to reconstruct typical motion patterns and identify abnormalities based on reconstruction errors in several studies on skeleton-based augmentation [8]. However, pure sequence-based models have certain drawbacks, such as:

- They rely significantly on sequence length and are vulnerable to missing or noisy frames.
- They operate on 1D coordinate sequences, failing to fully capture spatial joint relationships.
- They require large, clean datasets, which are rarely available for specialized anomaly detection tasks.

These constraints drove the need for a more generalizable, spatially aware model that could handle smaller, unbalanced datasets.

In contrast to sequence-based models, CNNs excel at extracting spatial hierarchies of structured data. Recent studies have proposed transforming skeleton coordinates into heatmaps or binary images, allowing convolutional architectures to learn joint relationships directly [9]. Such representations are particularly beneficial for tasks like pose classification, action recognition, and anomaly detection, where geometric relationships between joints carry critical information.

Convolutional Autoencoders (CAEs) support unsupervised feature learning by reconstructing inputs to identify anomalies, and lightweight variants using separable convolutions, batch normalization, and Conv2DTranspose layers maintain efficiency without sacrificing representational strength [10].

Unlike conventional skeleton-sequence models, our approach captures both local and global spatial structures within skeleton images while maintaining a compact model footprint. By combining spatial representation, data augmentation, and deployment-focused optimization, this work moves skeleton-based anomaly detection closer to a scalable, practical, real-world solution.

II. METHODOLOGY

In order to detect abnormalities in skeleton-based images with high efficiency and generalization, the proposed methodology presents an improved CAE framework that illustrates the overall system pipeline, including data preparation, model construction, training, evaluation, and quantization, as shown in Figure 1. The model creates a fully automated pipeline for anomaly detection by concentrating on spatial feature learning, robust data augmentation, and quantization for lightweight deployment.

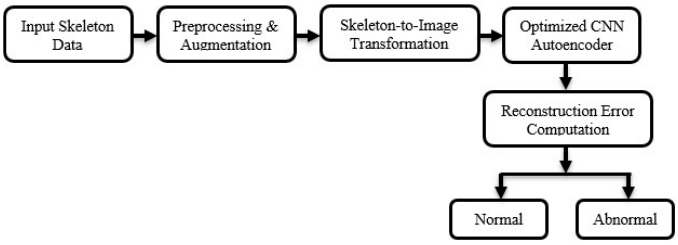

Fig. 1. Skeleton anomaly detection framework.

A. Dataset Preparation and Structure

The dataset used in this study consists of grayscale skeleton images generated from pose-estimation outputs derived from publicly available benchmark datasets and organized into training and testing sets. It includes five yoga and exercise poses, where Tree and Warrior II are classified as normal poses, whereas Dog, Goddess, and Plank are considered anomalous, as illustrated in Figure 2. The utilized benchmark datasets include NTU RGB+D 120, available at [11] and further discussed in [12], and the ShanghaiTech Campus Skeleton Subset, available at [13] and further discussed in [14].

The dataset structure supports easy extension to larger standard skeleton datasets such as NTU RGB+D 120 [11, 12] or the ShanghaiTech Campus Dataset [13, 14] for scalable experimentation. To ensure reproducibility and transparency in data utilization, the system automatically examines and logs the dataset composition, including the number of images per class and verifying category consistency.

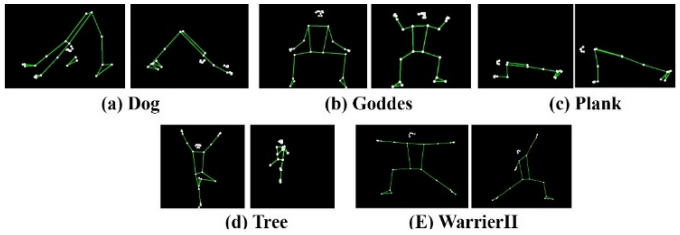

Fig. 2. Samples of five pose dataset categories.

B. Image Processing

All skeleton images undergo an optimized preprocessing pipeline before the training phase to guarantee normalization, consistency, and noise reduction. These procedures include grayscale conversion, Gaussian blur for noise reduction, resizing images to 128 × 128 pixels for balancing resolution and computational efficiency, and normalization by scaling pixels intensities to the [0, 1] range to stabilize gradient updates. The processed images are then cached in a compressed .npz file to speed up subsequent training and experimentation.

C. Data Augmentation

To address the limited training data, augmentation is applied to normal skeleton images using TensorFlow's ImageDataGenerator. This augmentation strategy improves generalization by introducing varied versions of normal poses. As a result, the model becomes more robust in detecting unseen anomalies during testing. Normal (non-anomalous) samples are augmented using the following transformations:

- Random rotation within ±15°
- Width and height shift up to 10%
- Zoom up to 15%
- Shear transformations to simulate pose variability
- Brightness adjustments in the range (0.8–1.2)
- Horizontal reflection disabled to preserve anatomical orientation

D. Skeleton-to-Image Transformation

The optimized model uses a symmetric CAE architecture, consisting of an encoder for feature extraction and a decoder for image reconstruction. This design prioritizes efficiency, spatial feature learning, and overall training stability.

The encoder is built using three layers, as illustrated in Figure 3, with SeparableConv2D layers having filter sizes of 32, 64, 128, and 256, each employing a 3 × 3 kernel. Batch normalization follows each convolutional layer to enhance stability and support faster convergence. Additionally, MaxPooling2D layers are used throughout the encoder to progressively reduce spatial dimensions from 128 to 64, then 32, and finally 16, as shown in Figure 3.

The decoder reconstructs the input using Conv2DTranspose layers that progressively restore spatial dimensions from 16 to 32, then 64, and finally 128. Batch normalization is applied to maintain stable gradient flow, and a sigmoid activation function is used in the output layer to produce a normalized grayscale reconstruction, as illustrated in Figure 3.

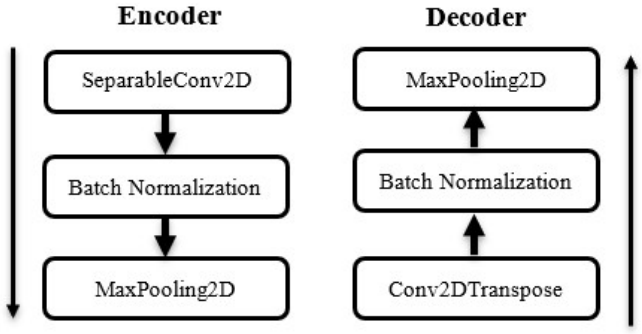

Fig. 3. Encoder and decoder layers.

E. Optimized Convolutional Neural Network Autoencoder (Training Strategy)

The training process is structured to maximize accuracy while minimizing overfitting. The model is trained exclusively on normal samples, following a semi-supervised anomaly detection approach. Training incorporates Early Stopping with a patience of 10 epochs to halt training when validation loss no longer improves, and ReduceLROnPlateau adjusts the learning rate by a factor of 0.5 if validation loss does not improve for five consecutive epochs. Model checkpointing is used to automatically save the best weights (best_skeleton_autoencoder.h5). Overall, the model is trained for up to 50 epochs with a batch size of 32, using real-time augmented image batches.

F. Reconstruction Error Computation (Anomaly Detection and Thresholding)

After training, anomalies are identified by measuring the reconstruction error between each input image and its reconstructed output. This is done by computing the pixel-wise Mean Squared Error (MSE) for each test image.

To distinguish normal from abnormal poses, a dynamic threshold is derived from the 95th percentile of reconstruction errors in normal validation data. Any image with an error above this threshold is flagged as anomalous. This adaptive approach enhances the model's ability to detect subtle deviations.

The adaptive thresholding method replaces the static fixed-value approach used in previous work, improving the model's ability to detect anomalies and significantly increasing recall. Although this results in a slight decrease in precision, it effectively addresses concerns regarding low recall.

G. Model Quantization for Deployment

To support the model's lightweight design and enable edge deployment, TensorFlow Lite quantization is applied after training. The resulting quantized model provides:

- 4× size reduction (e.g., from 12 MB to 3 MB)
- ≈ 60% faster inference time on CPU and embedded devices
- Minimal loss in reconstruction quality

This demonstrates that the optimized CNN autoencoder achieves strong accuracy while remaining highly efficient, making it suitable for real-world anomaly detection on resource-constrained devices.

III. EXPERIMENTAL RESULTS

All experiments were conducted on a workstation featuring an Intel Core i7-11800H CPU and 16 GB of RAM, running Python 3.10 and TensorFlow 2.x on Windows 11. Mixed-precision training was enabled to accelerate computations, and random seeds for NumPy, TensorFlow, and Python were fixed to ensure reproducibility. Dataset preprocessing was also cached to maintain consistent results across runs.

The dataset contained approximately 1,510 skeleton images derived from human-pose estimation outputs. These were split into two sets: SKELE_TRAIN, with about 1,080 images for learning normal patterns, and SKELE_TEST, with around 430 images for evaluation. The data included two normal classes ("Tree Pose" and "Warrior II") and three anomalous classes ("Dog Pose," "Goddess Pose," and "Plank Pose"). A stratified 70/30 training–testing split was used to preserve class balance. Only normal samples were used to train the autoencoder, whereas the test set included both normal and anomalous images to assess generalization.

The model's training history illustrates the convergence of training and validation losses over 20 epochs, demonstrating stable learning behavior, as illustrated in Figure 4(a). The error distribution and threshold demonstrate the model's ability to distinguish anomalies from normal data, as illustrated in Figure 4(b).

A. *Performance Evaluation Metrics*

The proposed classification model was evaluated using several performance metrics, including accuracy, precision, recall, and F1-score. In addition, the Receiver Operating Characteristic–Area Under the Curve (ROC-AUC) metric was employed to distinguish between normal and abnormal cases, as illustrated in Table I. The ROC curve for anomaly detection plots the classifier's performance at various thresholds, showing an AUC of 0.540, as illustrated in Figure 5(a).

TABLE I. PERFORMANCE EVALUATION METRICS

Metric	Accuracy	Precision	Recall	F1-score	ROC-AUC
Value	59.6%	61.0%	95.5%	74.5%	0.54

The evaluated metrics demonstrate high recall, identifying nearly all anomalous poses, while maintaining a reasonable level of precision—an appropriate balance for anomaly detection scenarios where missing an anomaly is more critical than raising a false alarm. Expanding the dataset to approximately 1.5K images improved the model's stability and helped mitigate overfitting compared to earlier experiments using smaller subsets. However, the resulting ROC curve produced an AUC of 0.540, indicating limited ability of the model to distinguish between normal and abnormal cases. The precision–recall curve illustrates the model's performance, showing an Average Precision (AP) of 0.659, as illustrated in Figure 5(b), which reflects a reasonably balanced trade-off between precision and recall.

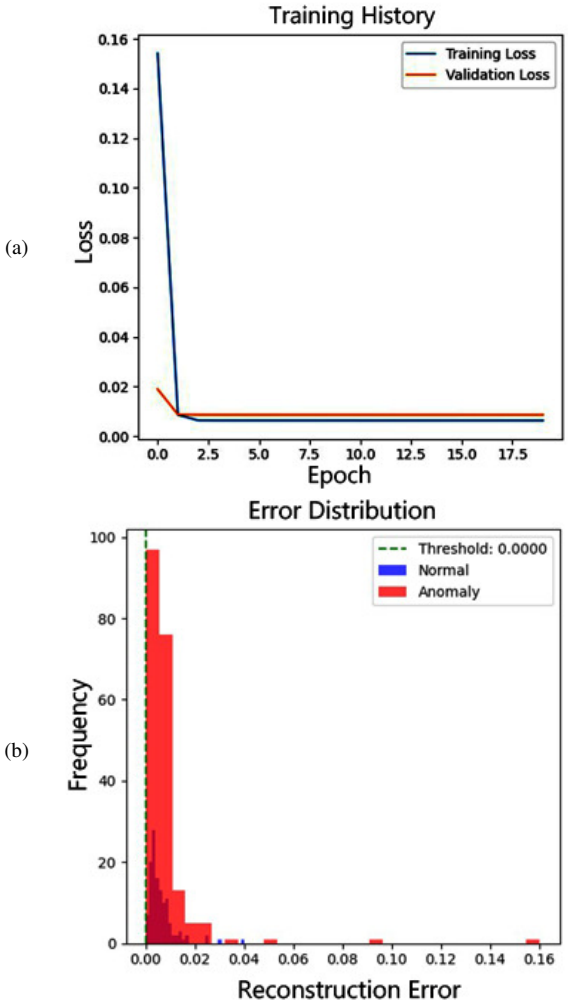

Fig. 4. Training performance analysis: (a) convergence of training and validation losses over 20 epochs, (b) reconstruction error distribution showing separation between normal and anomalous samples using the decision threshold.

After completing the training process, the model was converted and quantized using TensorFlow Lite, reducing computational complexity and making it suitable for deployment on resource-constrained devices.

Table II shows the impact of the quantization process, where the TensorFlow Lite model reduced the size by approximately four times and improved inference speed by approximately 60%, supporting the model's lightweight designation and demonstrating its suitability for edge-device deployment without a substantial loss in accuracy, as illustrated in Figure 6.

(a)

(b)

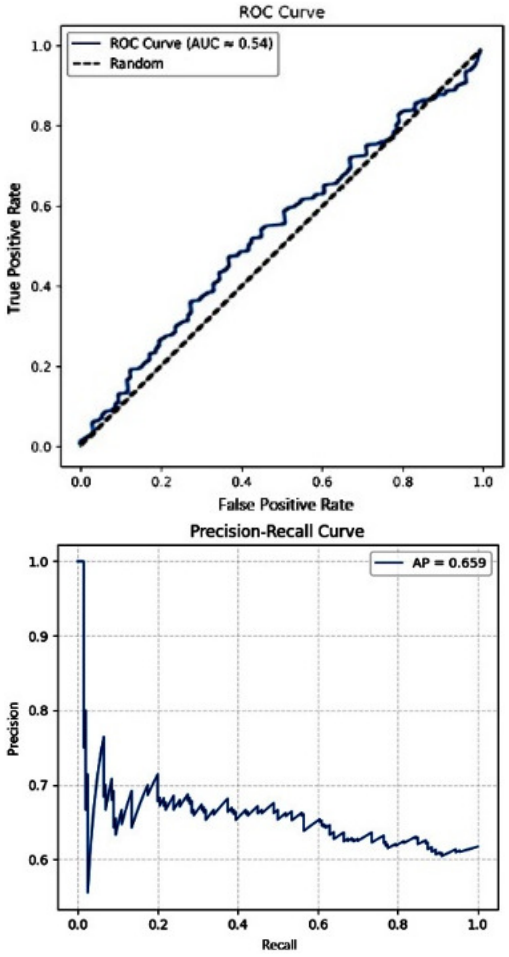

Fig. 5. Performance evaluation curves for anomaly detection: (a) ROC curve showing performance variations of the anomaly detection classifier across different classification thresholds, (b) precision–recall curve with AP summarizing the model's overall effectiveness.

TABLE II. PERFORMANCE COMPARISON BETWEEN ORIGINAL FP32 AND TENSORFLOW LITE

Features	Original FP32	TensorFlow Lite
Size (MB)	12.4	3.1
Inference time (ms/image)	41.2	17.3
Compression ratio	No compression	≈ 4×
Accuracy change	No change	≈ < 1% drop

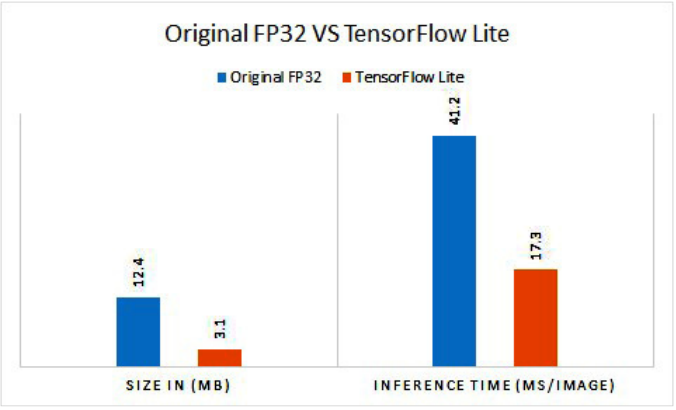

Fig. 6. Performance comparison between Original FP32 and TensorFlow Lite models.

The proposed optimized CNN autoencoder was compared against the LSTM autoencoder and the fully connected autoencoder approaches, as illustrated in Table III.

TABLE III. COMPARISON OF THE PROPOSED APPROACH WITH EXISTING APPROACHES

Feature	LSTM autoencoder [15]	Fully connected autoencoder [15]	Proposed CNN autoencoder
Input type	1-D joint coordinates	Flattened vectors	Skeleton images (128 × 128)
Architecture	64-unit LSTM encoder–decoder	3-layer dense autoencoder	SeparableConv2D + Conv2DTranspose
Recall	0.48	0.55	0.955
F1-score	0.61	0.68	0.75

The proposed CNN autoencoder significantly outperforms earlier LSTM-based and fully connected approaches, achieving a recall of 95.5% and an F1-score of 74.5%. This performance gain is driven by the model's ability to encode joint positions spatially as images, improving geometric awareness, the use of adaptive thresholding to enhance anomaly sensitivity, and data augmentation that increases intra-class variability, as illustrated in Figure 7. While accuracy and precision remain moderate, the model's ability to detect nearly all anomalies underscores the effectiveness of spatial representations for skeleton-based anomaly detection.

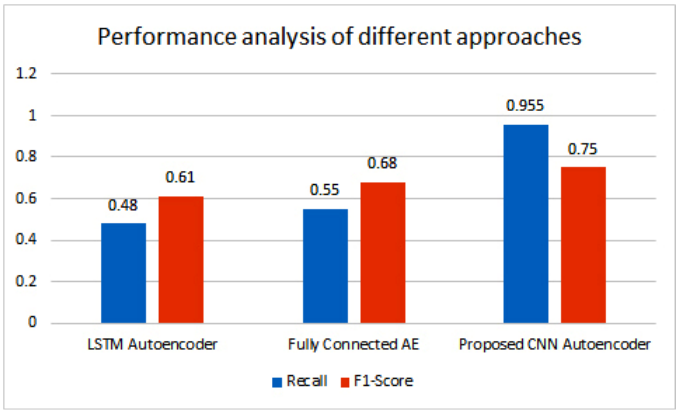

Fig. 7. Performance comparison analysis across different modeling approaches.

B. *Limitations*

Although the proposed approach demonstrates promising performance, it also presents several inherent limitations that should be acknowledged:

- Projecting skeletons into 2D removes essential depth cues, causing actions that are clear in 3D to become ambiguous, which remains a major source of errors in complex scenes.
- The method is limited to single-person skeletons and does not yet handle the multi-person, occluded, and interaction-heavy scenarios common in real-world surveillance.
- The dataset remains relatively small and lacks scene and subject diversity, indicating the need for validation on larger public skeleton-based benchmarks.

- Frame-wise reconstruction captures many anomalies, but subtle motion deviations such as gait irregularities require temporal context, and full spatiotemporal evaluation is left for future work.
- The model's robustness to adversarial attacks and severe sensor noise remains untested, leaving its performance under degraded conditions uncertain.
- The dynamic thresholding approach yields high recall but only moderate precision, indicating the need for more adaptive or learned thresholding methods.

IV. CONCLUSION AND FUTURE WORK

In this study, the proposed framework enhances sample diversity through data augmentation, supports real-time deployment via quantization, and captures both local and global spatial structures in a compact skeleton-based architecture. The dataset comprised approximately 1,510 skeleton images from human-pose estimation, split 70/30 into training (approximately 1,080 normal images) and testing (approximately 430 images including both normal and anomalous poses). The autoencoder was trained solely on normal samples, and evaluation on the test set assessed its ability to generalize to both normal and anomalous cases.

The proposed approach employs Convolutional Neural Networks (CNNs) to capture spatial hierarchies by converting skeleton coordinates into heatmaps or binary images, enabling the network to directly learn inter-joint relationships. Convolutional Autoencoders (CAEs) enable unsupervised feature learning and anomaly detection through reconstruction, without requiring labeled anomaly data.

The optimized CNN-based autoencoder achieved 59.6% accuracy, 61.0% precision, 95.5% recall, and a 74.5% F1-score, indicating strong anomaly sensitivity and effective spatial feature learning. Quantization reduced the model size by approximately four times and improved inference speed by around 60%, confirming its lightweight and deployable design. These results highlight the potential of spatial skeleton encoding for real-time anomaly detection and suggest that incorporating temporal modeling could further enhance precision.

The proposed approach can be extended in future work by integrating state-of-the-art 3D pose estimation techniques to reduce depth ambiguities and evaluating the resulting 3D representations for anomaly detection. We also plan to develop lightweight temporal fusion strategies to efficiently capture sequential patterns, incorporate multi-person tracking and composite skeleton frames for group-level anomaly detection, and conduct comprehensive evaluations on public benchmarks to assess generalizability. Additionally, we aim to investigate uncertainty-aware thresholding and the model's robustness under corrupted or shifted inputs to improve reliability in real-world scenarios.

DECLARATION OF COMPETING INTERESTS

Not applicable to this work.

ACKNOWLEDGMENT

We would like to express sincere gratitude to Eng. Alaa Ayman Mohamed Hussein, Department of Computer Science, Faculty of Computers and Information, Suez University, for her valuable cooperation in the preparation of this research paper. Her contributions and dedication greatly improved the quality and completion of this work.

DATA AVAILABILITY

The utilized benchmark datasets include NTU RGB+D 120 [11, 12] and the ShanghaiTech Campus Skeleton Subset, [13, 14].

REFERENCES

[1] B. Ren, M. Liu, R. Ding, and H. Liu, "A Survey on 3D Skeleton-Based Action Recognition Using Learning Method," *Cyborg and Bionic Systems*, vol. 5, May 2024, Art. no. 0100, https://doi.org/10.34133/cbsystems.0100.

[2] P. K. Mishra, A. Mihailidis, and S. S. Khan, "Skeletal Video Anomaly Detection Using Deep Learning: Survey, Challenges, and Future Directions," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 8, no. 2, pp. 1073–1085, Apr. 2024, https://doi.org/10.1109/TETCI.2024.3358103.

[3] J. Liu, "Algorithm for Skeleton Action Recognition by Integrating Attention Mechanism and Convolutional Neural Networks," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, pp. 604–613, Aug. 2023, https://doi.org/10.14569/IJACSA.2023.0140867.

[4] A. Alqahtani, "An optimized multi-scale convolutional autoencoder for efficient abnormal event detection using rgb, depth and optical flow data," *Multimedia Tools and Applications*, vol. 84, no. 28, pp. 34401–34435, Aug. 2025, https://doi.org/10.1007/s11042-025-20608-5.

[5] Z. Liu, X. Wu, J. Wu, X. Wang, and L. Yang, "Language-guided Open-world Video Anomaly Detection under Weak Supervision." arXiv, Mar. 17, 2025, https://doi.org/10.48550/arXiv.2503.13160.

[6] I.-C. Hwang and H.-S. Kang, "Anomaly Detection Based on a 3D Convolutional Neural Network Combining Convolutional Block Attention Module Using Merged Frames," *Sensors*, vol. 23, no. 23, Dec. 2023, Art. no. 9616, https://doi.org/10.3390/s23239616.

[7] W. Pang, Q. He, Y. Li, and N. Ahmed, "Detecting video anomalies by jointly utilizing appearance and skeleton information," *Expert Systems with Applications*, vol. 246, July 2024, Art. no. 123135, https://doi.org/10.1016/j.eswa.2023.123135.

[8] C. Xin, S. Kim, Y. Cho, and K. S. Park, "Enhancing Human Action Recognition with 3D Skeleton Data: A Comprehensive Study of Deep Learning and Data Augmentation," *Electronics*, vol. 13, no. 4, Feb. 2024, Art. no. 747, https://doi.org/10.3390/electronics13040747.

[9] R. Wu *et al.*, "DA-Flow: Dual Attention Normalizing Flow for Skeleton-Based Video Anomaly Detection," *IEEE Transactions on Multimedia*, vol. 27, pp. 8847–8858, 2025, https://doi.org/10.1109/TMM.2025.3607708.

[10] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction," in *21th International Conference on Artificial Neural Networks*, Espoo, Finland, 2011, pp. 52–59, https://doi.org/10.1007/978-3-642-21735-7_7.

[11] "Action Recognition Datasets: 'NTU RGB+D' Dataset and 'NTU RGB+D 120' Dataset." Rapid-Rich Object Search Lab. [Online]. Available: https://rose1.ntu.edu.sg/dataset/actionRecognition/.

[12] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020, https://doi.org/10.1109/TPAMI.2019.2916873.

[13] W. Liu, "StevenLiuWen/ano_pred_cvpr2018." Mar. 12, 2026, [Online]. Available: https://github.com/StevenLiuWen/ano_pred_cvpr2018.

[14] W. Liu, W. Luo, D. Lian, and S. Gao, "Future Frame Prediction for Anomaly Detection - A New Baseline," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 6536–6545, https://doi.org/10.1109/CVPR.2018.00684.

[15] D. Manju *et al.*, "Early Anomalus Action Detection in Surveillance Video Using MRCNN-LSTM Classification," *Engineering, Technology & Applied Science Research*, vol. 15, no. 4, pp. 25668–25676, Aug. 2025, https://doi.org/10.48084/etasr.10656.