DeepCAMS: A Deep Learning Approach for Real-Time Crowd Monitoring and Suspicious Behavior Detection Using Spatial-Temporal Analysis

Ayman A. Alharbi
Computer and Network Engineering Department, College of Computing, Umm Al-Qura University, Makkah, Saudi Arabia
aarharbi@uqu.edu.sa (corresponding author)

Received: 15 March 2025 | Revised: 3 May 2025 and 17 May 2025 | Accepted: 19 May 2025

 DOI: https://doi.org/10.48084/etasr.10954

ABSTRACT

The increasing need for robust and intelligent crowd monitoring systems has led to advances in deep learning-based solutions. However, existing methods often struggle with capturing complex crowd dynamics and detecting suspicious behaviors in real-time. This study introduces DeepCAMS (Deep Learning-based Crowd Analysis and Monitoring System), a novel architecture that integrates a Fully Convolutional Network (FCN) for spatial feature extraction and a Long Short-Term Memory (LSTM) network for temporal analysis. Unlike traditional methods, DeepCAMS addresses the limitations of static and shallow models by combining spatial and temporal insights, enabling accurate classification of crowd behaviors as Normal or Suspicious. DeepCAMS demonstrated superior performance across multiple metrics, marking a substantial improvement over traditional approaches. The ability of DeepCAMS to adapt to diverse crowd densities and identify subtle behavioral anomalies highlights its scalability and practical application in real-world surveillance. Therefore, DeepCAMS sets a new benchmark in crowd behavior analysis by offering a unified spatial-temporal framework that ensures high accuracy, adaptability, and efficiency in dynamic environments. This study not only advances the field of smart surveillance but also paves the way for future research on scalable and interpretable crowd monitoring systems.

Keywords-crowd monitoring; deep learning, Fully Convolutional Network (FCN); Long Short-Term Memory (LSTM); public safety; JHU-CROWD++ dataset; smart surveillance

I. INTRODUCTION

Rapid urbanization growth, the increase in the frequency of large-scale public events, and increased security concerns have necessitated the advancement of intelligent surveillance systems [1]. Efficient crowd monitoring and detection of suspicious behaviors are crucial to ensuring public safety, particularly in densely populated environments where conventional surveillance methods, which rely on human observation or simple motion-based algorithms, do not provide the required scalability, accuracy, and responsiveness [2-3].

Traditional approaches often suffer from high false-positive rates and struggle to adapt to dynamic crowd behaviors, making them inadequate for real-time security applications [4-5]. Artificial Intelligence (AI) and Deep Learning (DL) have significantly transformed surveillance systems by enabling them to process large volumes of visual and contextual data efficiently [6]. DL architectures, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and transformers, have demonstrated superior abilities to recognize complex patterns in video streams, surpassing traditional computer vision techniques [7-8]. The integration of AI-driven models with multimodal data sources, such as cameras, IoT sensors, and environmental data, has enabled real-time anomaly detection, making smart surveillance systems more robust in monitoring crowded areas such as airports, stadiums, and metro stations [9-10].

Despite these advances, challenges remain in achieving high detection accuracy under varying environmental conditions and scalability for large-scale, heterogeneous crowd scenarios. This study proposes DeepCAMS, a deep learning approach that synergistically combines spatial-temporal analysis techniques for real-time crowd monitoring and suspicious behavior detection. This method optimizes feature extraction and temporal pattern recognition through a novel unified architecture, addressing occlusion, motion complexity, and behavior semantics to enhance detection performance and operational efficiency in dynamic crowd environments.

II. METHODOLOGY

The proposed model integrates a Fully Convolutional Network (FCN) and a Long Short-Term Memory (LSTM) network, leveraging the strengths of spatial and temporal feature extraction to accurately classify and monitor crowd behaviors in real-time. Based on accepted crowd dynamics theories and security surveillance procedures, the DeepCAMS framework uses a binary classification method to differentiate between typical and suspicious crowd behaviors. Taking into account a variety of behavioral indicators, such as movement patterns, changes in crowd density, and the temporal consistency of group behaviors, the labelling method captured subtle behavioral patterns that define various crowd dynamics in surveillance scenarios.

Consistent and predictable movement patterns, defined by organized directional flow, stable crowd density distributions, and adherence to expected pedestrian dynamics, were used to identify normal crowd behaviors. These behaviors, which reflect the baseline crowd dynamics seen in typical surveillance environments, usually show smooth temporal transitions without sudden changes in crowd formation or movement velocity. Suspicious behaviors, on the other hand, include patterns that substantially depart from typical crowd dynamics, such as abrupt changes in density, unpredictable movement patterns, counter-flow behaviors that go against the established direction of the crowd, and quick formation changes that could be signs of an emergency or a possible security issue.

A. Model Architecture

The architecture of the proposed model is based on a hybrid design of FCN and LSTM layers to enable spatial and temporal analysis of video frames. This setup allows the model to capture complex features for effective crowd behavior analysis while maintaining temporal coherence between frames. The preprocessed video frames are fed into the FCN model, each represented as a three-dimensional tensor of size $H \times W \times C$, where H and W stand for the frame's height and width, respectively, and C is the number of colour channels (usually three for RGB images) [10]. Input frames are normalized before FCN processing to ensure that pixel values are scaled to the interval [0, 1] and resized to a uniform size to preserve consistency throughout the dataset. To capture representative crowd dynamics while maintaining computational efficiency, the frames are taken from crowd surveillance videos at a predetermined temporal interval. Spatial feature extraction is based on spatial information about individual positioning, crowd distribution, and environmental context that is present in every input frame. These normalized frames are processed by the VGG16 backbone to produce initial feature representations, which are then improved through the FCN layers to maintain spatial relationships that are crucial for the analysis of crowd behavior.

1) Feature Extraction

Initially, a VGG16 model [11], pre-trained on a large dataset, is employed to extract meaningful features from video frames, ensuring efficiency in recognizing basic visual patterns in crowded scenes. The FCN processes these features, preserving spatial details for frames of varying sizes, and introduces non-linearity through the ReLU activation function:

$$y_{ij} = ReLU(\sum_{k=1}^{K} w_k * x_{ij} + b) \quad (1)$$

where y_{ij} denotes the output feature map at location (i, j), w_k is the weight kernel, $*$ denotes the convolution operation, x_{ij} is the input frame, and b is the bias term.

2) Frame Sequence Processing

To produce sequential image sets, the temporal data extraction process starts with the methodical sampling of video frames at regular intervals (usually every two to three frames) [12]. Each sequence has T consecutive frames (with T between 10 to 15), which is a long enough temporal window to record significant changes in crowd behavior. By using standardized resizing and normalization, the frames are preprocessed to preserve temporal consistency.

$$F(i, j, t) = \{f(i, j, 1), f(i, j, 2), \ldots, f(i, j, T)\}$$

is the result of the model tracking feature values across the temporal sequence for each spatial location (i, j) in the feature maps produced by the FCN [13]. These trajectories encode the way spatial properties change over time by capturing motion patterns, density fluctuations, and behavioral modifications at specific locations within the crowd.

The model employs some techniques to extract temporal information. It calculates differences between frames to record shifts in motion direction and intensity, which are known as temporal gradients. Let $\Delta F(i, j, t) = F(i, j, t) - F(i, j, t-1)$. The process of calculating crowd movement velocities and acceleration patterns by analyzing pixel-level changes over time is known as velocity estimation. Temporal pooling is the process of capturing both short-term (2-3 frames) and long-term (8-10 frames) behavioral patterns by applying sliding window operations across the temporal dimension.

The retrieved temporal features are organized into consecutive vectors to preserve chronological order [14]. Each vector captures the temporal evolution of spatial features, including crowd density changes over time, variations in the flow direction, acceleration and deceleration patterns, and behavioral rhythm variations.

The model preserves temporal context by generating overlapping sequences, in which each new frame sequence shares 70-80% of frames with the preceding sequence [15]. This method reduces the possibility of overlooking critical behavioral shifts between sequences and ensures seamless temporal transitions.

3) Sequential Analysis with LSTM

The features derived from the FCN are passed into the LSTM layer, which captures temporal relationships across frames, making it well-suited for detecting sequential patterns in behavior. DeepCAMS's LSTM component uses a multi-layered recurrent architecture [16]. The FCN layers provide the LSTM network with sequential input in the form of spatially-processed feature maps, where each time step represents a feature extracted from a series of video frames.

The LSTM architecture is optimized to balance computational efficiency with temporal modeling capacity. The network can selectively retain or discard information across time steps. The FCN provides the flattened features to the LSTM, representing the spatial properties. The temporal evolution of crowd dynamics is captured by feeding these vectors one after the other. Equations (2) to (6) represent the gating mechanisms of the LSTM architecture. The forget gate (f_t) is essential for eliminating unnecessary temporal patterns in crowd movement because it decides which data from the previous cell state should be ignored. By regulating which new data should be kept in the cell state, the input gate (i_t) enables the model to focus on critical behavioral shifts. Selective information flow to later layers is made possible by the output gate (o_t), which controls which aspects of the cell state should affect the hidden state output. By analysing successive features, the LSTM can detect patterns such as crowd acceleration, decele. The last LSTM layer outputs a fixed-size feature vector that captures the temporal properties of the input sequence.

Forget gate:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f \quad (2)$$

Input gate:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i \quad (3)$$

Cell state update:

$$C_t = f_t * C_{t-1} + i_t * \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (4)$$

Output gate:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

Hidden state:

$$h_t = o_t * \tanh(C_t) \quad (6)$$

where σ is the sigmoid function, $*$ denotes element-wise multiplication, W represents the weight matrices, and b denotes the bias vectors [17].

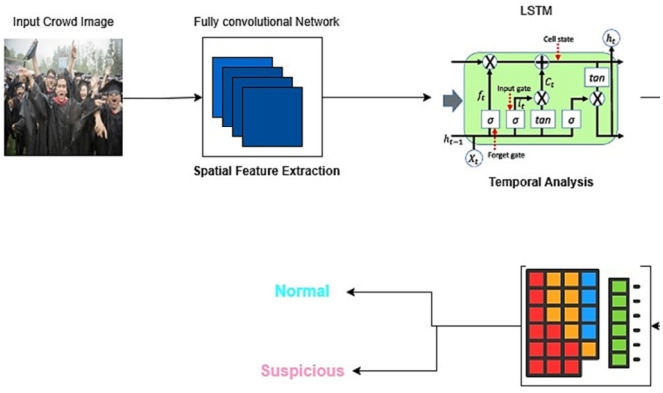

Fig. 1. DeepCAMS architecture for crowd monitoring and suspicious behavior detection. The architecture begins with an input crowd image, which undergoes spatial feature extraction through an FCN to capture spatial relationships within the crowd.

To avoid overfitting, dropout regularisation ($rate = 0.3$) is used in between LSTM layers. To ensure stable training over lengthy sequences, the network employs gradient clipping to resolve the vanishing gradient issue that frequently arises in recurrent architectures.

Finally, the output from the LSTM layer undergoes softmax classification to produce probabilities for the Normal and Suspicious classes. The system generates alerts for suspicious behaviors, enabling autonomous operation. The model minimizes the categorical cross-entropy loss function, as given by:

$$Loss = \sum_{c=1}^{C} y_c \log(\hat{y_c}) \quad (7)$$

where C is the number of classes, y_c is the true label, and $\hat{y_c}$ is the predicted probability for class c [16].

B. Dataset

The JHU-CROWD++ dataset [18] contains 4372 images with varied crowd densities and thorough point-level annotations. The annotated dataset exhibits a purposefully balanced distribution, with 2,623 images representing normal behaviors, comprising 60% of the total, and 1,749 images representing suspicious behaviors, which make up the remaining 40%. This distribution offers enough representation of anomalous patterns for reliable model training, while also reflecting realistic surveillance scenarios where normal behaviors predominate over suspicious activity. After applying a data augmentation pipeline of rotation (±15°), horizontal flipping, brightness adjustment (±20%), contrast modification (±15%), and random cropping (90-100% of original size), each technique with a probability of 0.5, the dataset grew while retaining a proportionate class distribution, yielding 8,745 images of suspicious behaviour and 13,115 images of normal behavior, maintaining the initial 60-40 percentage split. With this balanced distribution, the DeepCAMS system can learn discriminative features for both regular and anomalous crowd behaviors in real-world surveillance applications. It also prevents model bias toward either behavioral category and fosters robust classification performance across a variety of crowd scenarios. Through augmentation, the initial 3,498 training images are increased to roughly 17,490 images (5×). For the validation set, approximately 4,370 images are created by augmenting the 874 validation images (5×). The augmented dataset consists of 21,860 images (compared to 4,372 original images).

The training process involves structured stages: data preprocessing, data augmentation, splitting for training and validation, and model optimization. The dataset is split into 80% for training and 20% for validation, balancing model generalization and overfitting control. The augmented images are arranged into sets of 10-15 frames each for temporal analysis. This yields roughly 1,457 temporal sequences for training and 291 sequences for validation. This preserves the chronological integrity needed for behavioral analysis while ensuring enough temporal diversity for LSTM training. By adding changes to viewing angles, lighting, and spatial orientations, the augmentation process strengthens the dataset's resilience. It improves the model's ability to generalize to real-world surveillance scenarios with a range of environmental factors.

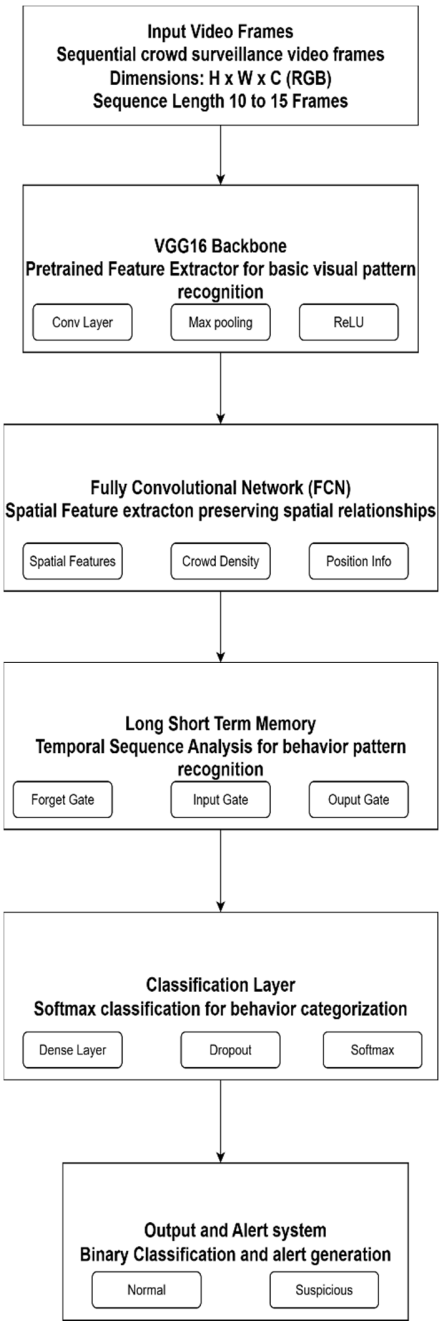

Fig. 2. Data flow and processing stages in DeepCAMS for crowd monitoring and suspicious behavior detection.

C. Optimization Parameters

The decay parameters are decreased every 10 epochs. This gradual decrease in the learning rate promotes stability in model convergence while minimizing the risk of overshooting optimal weights. A batch size of 32 is chosen, balancing memory constraints and model convergence speed. The categorical cross-entropy loss function (7) is employed to measure performance, penalizing deviations between true and predicted probabilities. The Adam optimizer [19], configured with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, was selected for its adaptive learning rate mechanism, which stabilizes training across variable gradient magnitudes. Dropout layers (dropout rate of 0.3) reduce overfitting by randomly deactivating neurons during training. This helps improve the model's generalization capability when applied to new data. The model's performance is evaluated using accuracy, precision, recall, and F1-score, which provide a comprehensive assessment of its ability to detect suspicious behavior in diverse crowd scenarios.

III. RESULTS AND DISCUSSIONS

A. Comparative Performance Analysis

DeepCAMS distinguishes itself by combining spatial-temporal analysis for real-time anomaly detection and behavior classification. Similarly, a CNN, designed for enhancing surveillance images under adverse weather conditions, achieved an F1-score of 84.4% with higher errors due to its dependence on preprocessing techniques such as denoising. As shown in Figure 3, the CNN misclassified behaviors due to spatial coherence loss, resulting in false positives. DeepCAMS overcomes this issue by directly integrating spatial and temporal cues, reducing error rates and improving behavioral classification accuracy.

A ViT-based model incorporated global spatial relationships to enhance crowd density estimation, achieving an F1-score of 86.4% and a precision of 89.4% (Figure 4). However, its reliance on static contextual data limits its ability to recognize evolving behavioral anomalies, reflected in its lower recall (88.1%).

The effectiveness of DeepCAMS in classifying Normal and Suspicious behaviors is illustrated in Figures 6 and 7. In Figure 6, the model correctly identifies normal crowd patterns with high accuracy, leveraging its FCN layers to maintain spatial coherence. The incorporation of LSTM further refines these predictions, considering temporal continuity and reducing the likelihood of false alarms in static scenarios with minor variations. This is particularly beneficial in low-risk environments, where reducing unnecessary alerts is crucial to operational efficiency. Figure 7 showcases the capacity of DeepCAMS in detecting anomalous behaviors, such as sudden clustering or erratic dispersals, which often signal security threats. The LSTM-driven analysis ensures that behavioral patterns developing over time are accurately flagged as suspicious. This ability is particularly useful for real-time security monitoring in airports, stadiums, and transportation hubs, where early anomaly detection is essential to prevent incidents.

Figure 8 illustrates the training and validation accuracy trends across various models, highlighting DeepCAMS' smoother convergence and higher stability in learning complex crowd behaviors. Unlike CNN, which struggles with noisy datasets, DeepCAMS efficiently processes raw spatial-temporal data, ensuring lower generalization error. Despite its superior performance, DeepCAMS occasionally misclassifies highly dense static crowds as suspicious, emphasizing the need for further optimization of temporal filters. Additionally, minor performance fluctuations were observed in imbalanced datasets, suggesting that advanced data augmentation techniques could enhance its robustness.

Fig. 3. CNN model predictions for crowd classification.

Fig. 4. ViT model predictions in adverse weather conditions.

Fig. 5. VIT-TA model predictions highlighting global context in crowd monitoring.

Fig. 6. DeepCAMS predictions: Classification of Normal behaviors.

Fig. 7. DeepCAMS predictions showcasing the effectiveness of spatiotemporal analysis.

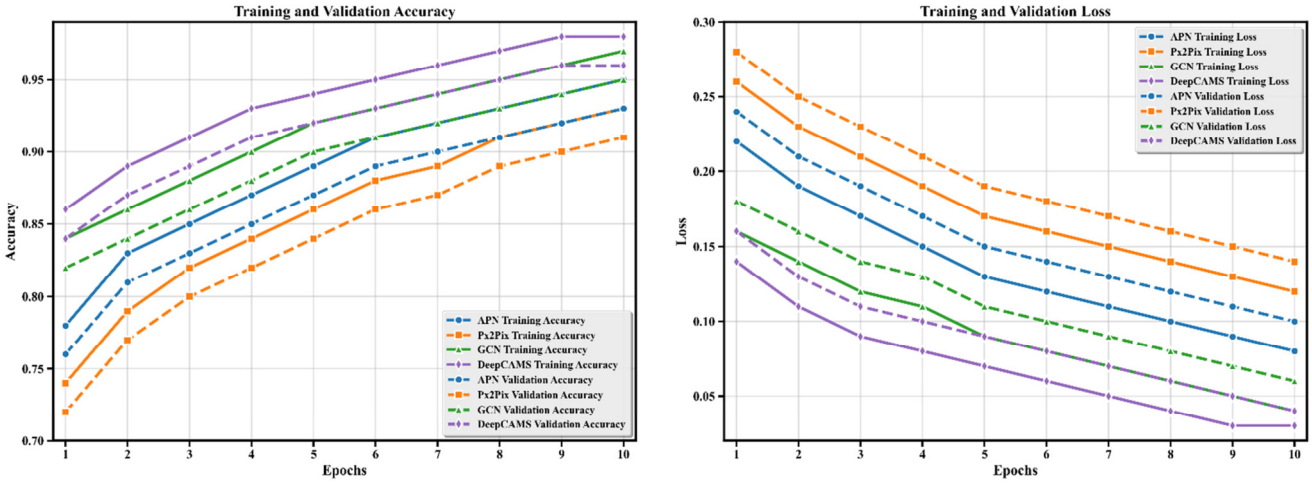

Fig. 8. Training and validation accuracy comparison across models.

Table I provides comprehensive confusion matrices for four models with detailed performance metrics. The reliance of ViT on static contextual data limits its ability to recognize evolving behavioral anomalies, reflected in its lower recall (88.1%). DeepCAMS bridges this gap by integrating temporal dependencies with real-time spatial feature extraction, enabling it to detect progressive shifts in behavior, such as a peaceful gathering escalating into an unsafe crowd movement, a capability that ViT lacks.

TABLE I. PERFORMANCE COMPARISON OF CROWD MONITORING MODELS ON TEST DATASET

Model	TP	TN	FP	FN	P	R	F1	Acc
CNN	1404	2205	418	343	83.6	80.4	82.0	82.6
ViT	1437	2363	260	310	84.7	82.2	83.4	87.0
ViT-TA	1539	2345	278	208	84.7	88.1	86.4	88.9
DeepCAMS	1573	2401	222	174	87.6	90.0	88.8	90.9

Incorporating additional contextual information such as weather conditions, time of day, and event types could help improve model accuracy and adaptability in dynamic real-world conditions. Multimodal data has been shown to increase the robustness of detection systems, providing a holistic understanding of crowd behavior [20]. Implementing model optimization techniques, such as pruning or quantization, can reduce computational requirements, making the model suitable for real-time deployment on resource-constrained devices, such as edge or mobile devices [21]. In addition, sophisticated data augmentation techniques and transfer learning could increase model robustness and generalizability, allowing it to handle a broader range of crowd scenarios [22]. Adaptive learning mechanisms can allow the model to continuously learn from new data would improve its ability to adapt to changing patterns in crowd behavior over time, enabling enhanced performance without retraining [23]. By implementing these improvements, the model is expected to achieve higher accuracy, interpretability, and scalability, allowing it to be deployed in varied real-time crowd monitoring scenarios.

The DeepCAMS model was trained and evaluated on the JHU-CROWD++ [18], a large-scale crowd dataset specifically designed for crowd counting and analysis. The dataset consists of over 4,372 images with varying crowd densities, annotated with detailed point-level labels. JHU-CROWD++ includes challenging real-world conditions, such as adverse weather, varying lighting, and occlusions, making it well-suited for robust crowd behavior analysis. The dataset has been widely used in recent crowd monitoring research due to its diversity and extensive ground-truth annotations. The dataset is publicly available and has been referenced in multiple studies related to smart surveillance and crowd analysis.

IV. CONCLUSION

This study highlights the importance of DL and feature integration in improving crowd monitoring and detection of suspicious behavior. The DeepCAMS model, which combines spatial and temporal analysis through FCN and LSTM, outperforms existing methods such as VIT-TA, CNN, and VIT by achieving higher accuracy, recall, and lower error rates. Although each model has its strengths, DeepCAMS offers a balanced and adaptable solution, excelling in real-time behavioral analysis and anomaly detection in diverse environments. The test results demonstrate its robustness for public safety applications in event security, urban surveillance, and emergency response. This study reaffirms that a hybrid spatial-temporal approach is key to advancing intelligent crowd monitoring systems, paving the way for more reliable and scalable surveillance solutions. Quantitative improvements demonstrate the practical significance of DeepCAMS, achieving superior behavior classification performance with F1 scores that exceed 88%.

REFERENCES

[1] Y. Li, "A Deep Spatiotemporal Perspective for Understanding Crowd Behavior," *IEEE Transactions on Multimedia*, vol. 20, no. 12, pp. 3289–3297, Sep. 2018, https://doi.org/10.1109/TMM.2018.2834873.

[2] H. Su, H. Yang, S. Zheng, Y. Fan, and S. Wei, "The Large-Scale Crowd Behavior Perception Based on Spatio-Temporal Viscous Fluid Field," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 10, pp. 1575–1589, Jul. 2013, https://doi.org/10.1109/TIFS.2013.2277773.

[3] S. Wang, J. Cao, and P. S. Yu, "Deep Learning for Spatio-Temporal Data Mining: A Survey," *IEEE Transactions on Knowledge and Data*

Engineering, vol. 34, no. 8, pp. 3681–3700, Dec. 2022, https://doi.org/10.1109/TKDE.2020.3025580.

[4] W. Wang *et al.*, "HAST-IDS: Learning Hierarchical Spatial-Temporal Features Using Deep Neural Networks to Improve Intrusion Detection," *IEEE Access*, vol. 6, pp. 1792–1806, 2018, https://doi.org/10.1109/ACCESS.2017.2780250.

[5] N. Li, F. Chang, and C. Liu, "Spatial-Temporal Cascade Autoencoder for Video Anomaly Detection in Crowded Scenes," *IEEE Transactions on Multimedia*, vol. 23, pp. 203–215, 2021, https://doi.org/10.1109/TMM.2020.2984093.

[6] E. B. Varghese, S. M. Thampi, and S. Berretti, "A Psychologically Inspired Fuzzy Cognitive Deep Learning Framework to Predict Crowd Behavior," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 1005–1022, Apr. 2022, https://doi.org/10.1109/TAFFC.2020.2987021.

[7] Y. Miao *et al.*, "Abnormal Behavior Learning Based on Edge Computing toward a Crowd Monitoring System," *IEEE Network*, vol. 36, no. 3, pp. 90–96, Feb. 2022, https://doi.org/10.1109/MNET.014.2000523.

[8] R. Nawaratne, D. Alahakoon, D. De Silva, and X. Yu, "Spatiotemporal Anomaly Detection Using Deep Learning for Real-Time Video Surveillance," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 393–402, Jan. 2020, https://doi.org/10.1109/TII.2019.2938527.

[9] M. Qaraqe *et al.*, "PublicVision: A Secure Smart Surveillance System for Crowd Behavior Recognition," *IEEE Access*, vol. 12, pp. 26474–26491, 2024, https://doi.org/10.1109/ACCESS.2024.3366693.

[10] S. A. Priyanka and Y. K. Wang, "Fully Symmetric Convolutional Network for Effective Image Denoising," *Applied Sciences*, vol. 9, no. 4, Feb. 2019, Art. no. 778, https://doi.org/10.3390/app9040778.

[11] P. K. Sahoo *et al.*, "An Improved VGG-19 Network Induced Enhanced Feature Pooling for Precise Moving Object Detection in Complex Video Scenes," *IEEE Access*, vol. 12, pp. 45847–45864, 2024, https://doi.org/10.1109/ACCESS.2024.3381612.

[12] M. R. Bhuiyan, J. Abdullah, N. Hashim, and F. Al Farid, "Video analytics using deep learning for crowd analysis: a review," *Multimedia Tools and Applications*, vol. 81, no. 19, pp. 27895–27922, Aug. 2022, https://doi.org/10.1007/s11042-022-12833-z.

[13] Y. Zhao, X. Zhao, S. Chen, Z. Zhang, and X. Huang, "An Indoor Crowd Movement Trajectory Benchmark Dataset," *IEEE Transactions on Reliability*, vol. 70, no. 4, pp. 1368–1380, Sep. 2021, https://doi.org/10.1109/TR.2021.3109122.

[14] T. Yang, C. Wang, T. Zhou, Z. Cai, K. Wu, and B. Hou, "Identification of Anomalous Behavioral Patterns in Crowd Scenes," *Computers, Materials & Continua*, vol. 71, no. 1, pp. 925–939, 2022, https://doi.org/10.32604/cmc.2022.022147.

[15] K. Rezaee, S. M. Rezakhani, M. R. Khosravi, and M. K. Moghimi, "A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance," *Personal and Ubiquitous Computing*, vol. 28, no. 1, pp. 135–151, Feb. 2024, https://doi.org/10.1007/s00779-021-01586-5.

[16] V. Mahor, J. Choudhary, and D. P. Singh, "Analysis of Human-Based Suspicious Activity Using Bidirectional Long Sort Term Memory (Bi-LSTM)," *Procedia Computer Science*, vol. 260, pp. 725–733, Jan. 2025, https://doi.org/10.1016/j.procs.2025.03.252.

[17] S. K. Tripathy and P. Shanmugam, "Real-Time Spatial-Temporal Depth Separable CNN for Multi-Functional Crowd Analysis in Videos," *International Journal of Image and Graphics*, Nov. 2023, Art. no. 2550047, https://doi.org/10.1142/S0219467825500470.

[18] V. A. Sindagi, R. Yasarla, and V. M. Patel, "JHU-CROWD++: Large-Scale Crowd Counting Dataset and A Benchmark Method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2594–2609, Feb. 2022, https://doi.org/10.1109/TPAMI.2020.3035969.

[19] M. Reyad, A. M. Sarhan, and M. Arafa, "A modified Adam algorithm for deep neural network optimization," *Neural Computing and Applications*, vol. 35, no. 23, pp. 17095–17112, Aug. 2023, https://doi.org/10.1007/s00521-023-08568-z.

[20] S. Mitra, "AI-driven predictive models for traffic flow in IoT-driven smart cities," *Uncertainty Discourse and Applications*, vol. 1, no. 2, pp. 170–178, Dec. 2024, https://doi.org/10.48313/uda.v1i2.38.

[21] S. A. Quadri and K. S. Katakdhond, "Suspicious Activity Detection Using Convolution Neural Network," *Journal of Pharmaceutical Negative Results*, pp. 1235–1245, Oct. 2022, https://doi.org/10.47750/pnr.2022.13.S01.151.

[22] A. Dionis-Ros, J. Vila-Francés, R. Magdalena-Benedito, F. Mateo, and A. J. Serrano-López, "Multimodal Video Analysis for Crowd Anomaly Detection Using Open Access Tourism Cameras," *Applied Sciences*, vol. 14, no. 23, Jan. 2024, Art. no. 11075, https://doi.org/10.3390/app142311075.

[23] T. Alafif *et al.*, "Hybrid Classifiers for Spatio-Temporal Abnormal Behavior Detection, Tracking, and Recognition in Massive Hajj Crowds," *Electronics*, vol. 12, no. 5, Jan. 2023, Art. no. 1165, https://doi.org/10.3390/electronics12051165.