Power System Technology
ISSN:1000-3673

Received: 06-01-2025 *Revised: 15-02-2025* *Accepted: 25-03-2025*

Hybrid LSTM-Transformer Architecture for Abnormal Crowd Behavior Detection

[1]Komal Jadhav, [2]Mahesh Chavan

[1]Assistant Professor, K.I.T's College of Engineering, Kolhapur, India.

jadhav.komal@kitcoek.in

[2]Professor, K.I.T's College of Engineering, Kolhapur, India

chavan.mahesh@kitcoek.in

Abstract: This study proposes an innovative deep learning framework for abnormal crowd behavior detection, combining powerful spatial-temporal feature extraction techniques with optimized neural network components. The pipeline begins with video input, from which frames are extracted and processed using a hybrid CNN architecture that integrates Bottleneck and Residual blocks to capture deep spatial features. These features are then passed through a BiLSTM network to learn temporal dependencies, followed by a Transformer Encoder that enhances long-range context understanding. Dense layers and a Softmax function complete the classification of behaviors as normal or abnormal. The proposed model is evaluated against state-of-the-art approaches, demonstrating superior accuracy, lower false detection rates, and significantly reduced detection time, especially under high-density crowd conditions. Notably, it achieves a detection time of just 0.4 seconds at 10 fps. The model's design supports scalability and real-time applicability, making it suitable for public safety, event monitoring, and crowd management systems. This research highlights the importance of combining spatial, temporal, and contextual insights for robust surveillance systems and offers a promising foundation for further development in intelligent video analysis and behavior detection technologies.

Keywords: Abnormal Behavior Detection, Crowd Surveillance, Deep Learning, Temporal Feature Extraction, Classification Performance

1. Introduction

Abnormal crowd behavior detection has garnered increasing attention in recent years due to its vital applications in public safety, event monitoring, and disaster management. The rapid growth in urbanization, population density, and large-scale public gatherings has heightened the need for advanced surveillance systems capable of identifying unusual activities in real-time. Conventional methods relying on manual observation or basic image processing are no

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

longer adequate to handle the complexities of such scenarios. To address this, automated video analysis systems based on machine learning and deep learning have emerged as an effective approach, offering the potential for faster, more accurate detection of abnormal events within large crowds.

Abnormal crowd behavior refers to activities deviating from expected norms, such as stampedes, riots, or sudden dispersions in large gatherings. Detecting such anomalies is critical in preventing accidents, minimizing casualties, and ensuring timely intervention. However, the challenges in abnormal crowd behavior detection are multifaceted. Variations in crowd density, occlusions, camera angles, and environmental factors contribute to the difficulty of designing robust and generalized models. Moreover, capturing the temporal dynamics and spatial features simultaneously is essential for accurately distinguishing between normal and abnormal events.

Recent advancements in computer vision and deep learning have provided promising tools for tackling complex problems like crowd behavior analysis. Convolutional Neural Networks (CNNs) have proven highly effective in spatial feature extraction from images, while Recurrent Neural Networks (RNNs), particularly Bidirectional Long Short-Term Memory (BiLSTM) networks, have excelled in capturing temporal patterns. Transformers, known for their remarkable success in natural language processing and sequence modeling tasks, have begun to make inroads into video-based applications due to their capability to understand global dependencies.

Despite these advances, existing models often suffer from limitations, such as inadequate temporal modeling and an inability to generalize across diverse datasets. Most current methods either focus on spatial features alone or rely on rudimentary temporal modeling techniques, which fail to capture the complex interplay between spatial and temporal information in crowded scenes. Addressing these gaps necessitates an innovative approach that leverages the strengths of CNNs, RNNs, and Transformers to extract meaningful insights from both spatial and temporal data.

The increasing frequency of crowd-related incidents occurring worldwide underscores the urgency of developing efficient and scalable solutions for real-time abnormal behavior detection. Video surveillance systems equipped with intelligent algorithms can act as force multipliers, assisting security personnel in monitoring vast areas with minimal effort. The hybrid integration of CNNs, BiLSTMs, and Transformers offers a unique opportunity to address the limitations of existing methods. By combining the spatial feature extraction capabilities of CNNs, the temporal modeling strength of BiLSTMs, and the global contextual understanding of Transformers, it is possible to design a holistic framework for abnormal crowd behavior detection.

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

This study is driven by the need to bridge the gap between spatial and temporal analysis in crowd behavior detection systems. The motivation stems from the realization that a unified framework combining these components can significantly improve detection accuracy while maintaining computational efficiency. Additionally, the proposed methodology aims to contribute to the broader field of computer vision by demonstrating the applicability of hybrid architectures in real-world surveillance scenarios.

This work introduces a novel hybrid architecture, combining CNN, BiLSTM, and Transformer components, to achieve state-of-the-art performance in abnormal crowd behavior detection. The key contributions of this study are as follows:

1. Integration of Spatial and Temporal Features: The methodology leverages CNNs with Bottleneck and Residual blocks for efficient spatial feature extraction, complemented by BiLSTMs for capturing short-term and long-term temporal dependencies.

2. Incorporation of Transformers for Global Context: The Transformer encoder enhances the framework by modeling global dependencies and improving the representation of complex crowd dynamics.

3. Efficient Classification Framework: Dense layers followed by a softmax activation layer ensure accurate classification of crowd behavior into normal or abnormal categories.

4. Scalable and Generalized Solution: The hybrid architecture is designed to generalize across diverse datasets and scenarios, addressing challenges such as varying crowd densities and occlusions.

5. Real-Time Applicability: The model is optimized for real-time implementation, enabling proactive monitoring and intervention in dynamic environments.

By addressing the challenges associated with spatial-temporal modeling and leveraging the complementary strengths of CNNs, BiLSTMs, and Transformers, this work sets the stage for the development of more robust and efficient systems for abnormal crowd behavior detection. The proposed framework has the potential to serve as a valuable tool in ensuring public safety and managing large-scale gatherings effectively.

2. Related Work

Detecting abnormal behavior in crowded environments has been a prominent area of research, with many methodologies leveraging machine learning and deep learning frameworks to improve detection accuracy. Each contribution tackles specific challenges, such as handling dynamic crowd behaviors, reducing computational costs, or improving classification accuracy.

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

Alia et al. [1] proposed a cloud-enabled deep learning framework specifically designed for detecting pushing behavior in crowded event entrances. This innovative system records large-scale events in real time, capturing the precise conditions under which pushing incidents occur. By incorporating cloud computing, the framework ensures robust processing capabilities for handling complex crowd dynamics, resulting in reduced response times and enhanced accuracy. The study highlights its practical applications in managing dense crowds at event gates, contributing significantly to reducing accidents caused by chaotic pushing. The use of cloud infrastructure further ensures scalability and adaptability for deployment across different scenarios, marking a significant advancement in crowd behavior monitoring.Mehmood [2] introduced a method that utilizes a 2D Convolutional Neural Network (CNN) to detect anomalies in crowd behavior by analyzing spatial and temporal patterns. The proposed system effectively extracts meaningful features from video data, leveraging pre-trained CNN models to minimize computational overhead. This work stands out for its ability to identify violent behaviors with minimal computational cost, making it suitable for real-time applications. The research emphasizes practical implementations by demonstrating the model's efficiency in reducing false positives and improving precision during anomaly detection tasks. The method’s ability to handle spatial-temporal intricacies in crowd footage adds to its robustness, catering to various surveillance applications.

Mohamed et al. [3] developed a novel method called Texture Classification-based Feature Processing (TCFP) for detecting violence and anomalies in crowded environments. This approach uniquely employs non-dimensional vectors as sequence frames to facilitate faster categorization, significantly reducing classification delays. By improving the accuracy of texture-based feature detection, this methodology excels in dense crowd scenarios where computational efficiency is critical. The authors demonstrate the framework's ability to extract critical frame properties, achieving enhanced detection rates without compromising on speed or accuracy. This work is particularly valuable for high-density environments such as transportation hubs or public events, where rapid detection is vital for maintaining safety.

Lopez-Carmona and Garcia [4] proposed an adaptive system called CellEVAC for guiding crowd evacuations in real-time. This framework uses behavioral optimization techniques to analyze human movement patterns during emergencies. The proposed model provides a dynamic evacuation strategy, significantly reducing computational time and improving pedestrian safety. By focusing on dynamic crowd behaviors, this system ensures optimized evacuation plans tailored to specific scenarios, such as crowded stadiums or urban spaces. The research highlights the practical implications of integrating behavioral analysis with real-time guidance, providing a framework for addressing crowd safety challenges effectively. Bouhlel et al. [5] introduced a technique for estimating crowd density using aerial imagery. Their methodology combines deep learning with handcrafted features to enhance detection

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

accuracy in high-density environments. This fusion of automated feature extraction and manual input enables the model to adapt to diverse conditions, ensuring reliable density estimations. The study is notable for its focus on leveraging aerial data, making it particularly useful for scenarios such as large public gatherings or disaster response. The ability to assess crowd density from an overhead perspective provides valuable insights for managing crowd control and resource allocation effectively.

Chang et al. [6] designed a hybrid deep learning model that integrates CNN and Long Short-Term Memory (LSTM) networks to detect abnormal behavior in crowded scenes. This framework capitalizes on the strengths of CNNs for spatial feature extraction and LSTMs for capturing temporal dynamics, creating a comprehensive solution for complex scenarios. The authors demonstrate the model's superior performance in identifying anomalies, particularly in situations where both spatial and temporal correlations are critical. The research underscores the importance of combining spatial-temporal analysis for robust anomaly detection, contributing to the development of multi-modal detection systems.Xu and Lu [7] presented a multi-branch convolutional neural network designed to reduce computational complexity while detecting abnormal behavior in crowded environments. This approach integrates multiple convolutional branches to analyze different spatial and temporal scales, providing a holistic view of crowd activities. By employing fusion mechanisms, the network enhances its capability to recognize intricate patterns in video data, achieving high accuracy with minimal computational overhead. The work emphasizes the efficiency of multi-branch architectures in balancing performance and computational demands, making it suitable for large-scale surveillance applications.

Dong et al. [8] proposed an intelligent framework for detecting abnormal behavior in multi-person scenarios using BiLSTM networks. This framework captures temporal dependencies within the data, enabling precise identification of unusual activities. The research highlights the use of feature vectors and frames to reduce latency in the detection process, making it effective for real-time applications. The study provides a detailed evaluation of the framework’s performance in diverse scenarios, showcasing its ability to handle complex crowd interactions and improve detection reliability.Jiang et al. [9] employed a Generative Adversarial Network (GAN)-based approach to analyze flow acceleration for abnormal behavior detection. This innovative technique focuses on extracting motion information to optimize detection processes, achieving lower computational costs and higher accuracy. The research demonstrates the potential of GANs for handling dynamic crowd behaviors, providing a robust framework for real-time anomaly detection. By focusing on motion dynamics, this work offers unique insights into improving the efficiency of crowd behavior analysis.

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

Qaraqe et al. [10] introduced PublicVision, a secure surveillance system that combines Swin Transformer architectures with advanced security protocols for detecting crowd behavior anomalies. This system ensures data integrity and privacy through techniques such as dynamic virtual private networks (VPNs) and firewalls, making it suitable for sensitive environments like airports or public transportation hubs. The study emphasizes the scalability and robustness of integrating advanced neural networks with secure transmission protocols, offering a comprehensive solution for crowd monitoring.Choi et al. [11] developed the 3D Atrous Inception Module (3D-AIM) for classifying crowd behavior. This model leverages atrous convolution to explore spatial relationships at multiple scales, enhancing its ability to identify anomalies. The introduction of a novel loss function further improves the network's classification accuracy. The research showcases the model's efficacy in detecting intricate crowd behaviors, making it a valuable contribution to surveillance systems targeting complex environments. Table 1 shows the summarized review.

Table 1: Summarized review of latest methods

Author	Technique Used	Features	Outcome/Results
Shakeel, Burhanuddin, and Desa [12]	Enhanced deep learning model applied to CT-based lung scans	Implements optimized image processing techniques for better cancer region detection	Accurately segments affected lung tissues, leading to improved diagnostic reliability
Choi et al. [11]	3D-Atrous Inception-based architecture for behavior classification in crowds	Focuses on individual and collective crowd motion using dilated convolutions	Achieves high classification performance across diverse crowd scenarios
Jiang et al. [9]	GAN-driven framework utilizing motion acceleration for anomaly recognition	Leverages dynamic motion characteristics to understand human behavior	Lowers detection time and computational demands in identifying abnormal actions
Li et al. [13]	Motion consistency-driven variational method for detecting abnormal activities	Evaluates inconsistencies in motion patterns between video frames	Significantly boosts recognition performance and precision
Dong et al. [8]	Bi-LSTM-powered intelligent model tailored for multi-person	Encodes temporal and spatial features using frame	Leads to improved recognition rates and effective behavior

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

	activity monitoring	sequences and feature vectors	tracking
Alafif et al. [14]	Combined CNN and Random Forest model for detecting spatio-temporal anomalies in dense crowds	Integrates spatial and temporal patterns to detect irregular crowd behaviors	Demonstrates high AUC and reliable anomaly detection across crowded scenes
Li et al. [15]	Behavior disturbance propagation model for fall detection in high-density areas	Simulates pedestrian interactions to assess fall risk	Offers increased effectiveness in emergency detection and public safety
Qaraqe et al. [10]	Transformer-based secure surveillance method using PublicVision	Ensures secure CCTV data transmission through firewalls and VPN networks	Enables reliable, scalable surveillance for large public spaces and crowd control
Xu and Lu [7]	Lightweight multi-path convolutional network for behavior anomaly analysis	Reduces structural complexity while detecting irregular crowd activity	Improves system efficiency for pedestrian safety and real-time monitoring
Juginder Pal Singh and Manoj Kumar[16]	Tunicate swarm optimization-based GAN for recognizing violent behaviors in crowds	Extracts features such as texture and motion flow descriptors like GLCM and LTP	Improves sensitivity, specificity, and accuracy for crowd violence detection
Ou, Zhu, Chen, and Liu [17]	CNN-based framework for understanding crowd behavior in surveillance videos	Captures key video frames to analyze human activity	Enhances recognition performance with minimal latency
Ghorbanpour and Nahvi [18]	Unsupervised dynamic behavior analysis using point tracking and group formation	Detects crowd behavior by clustering key feature points	Delivers high detection accuracy with reduced error rate
Chang, Chang, and Lin [6]	Deep hybrid CNN-LSTM model designed for crowd behavior abnormality detection	Learns spatial and sequential patterns from pedestrian activity	Improves the accuracy and speed of detecting abnormal patterns
Ammar and Cherif [19]	Deep learning-based real-time anomaly detection system for human crowds	Ensures extended coverage and responsive detection in surveillance systems	Boosts detection precision and real-time responsiveness

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

3. Proposed Work

The proposed work presents a hybrid deep learning framework for abnormal behavior detection from video sequences. As illustrated in the flow diagram, the process begins with the input of raw video data, followed by frame extraction to convert continuous video into individual image frames. These frames are passed through a CNN that integrates both Bottleneck and Residual blocks to capture rich spatial features while maintaining computational efficiency. To model the temporal relationships across the extracted features, a BiLSTM layer is employed. This is further enhanced by a Transformer Encoder, which captures long-range dependencies and contextual interactions within the frame sequence. The output is then passed through fully connected Dense layers, followed by a Softmax activation layer for classification. Finally, the system categorizes the input sequence as either normal or abnormal behavior. This architecture effectively combines spatial, temporal, and contextual information for robust video-based classification.

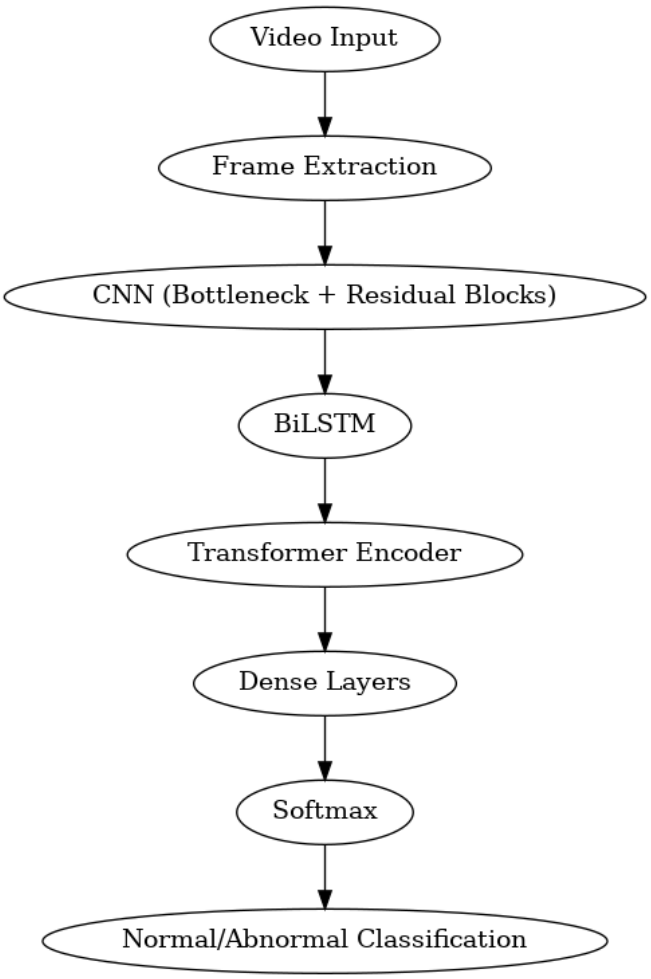

Figure 1: Proposed system Flow

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

The flow diagram shown in Figure1of the hybrid LSTM-Transformer architecture for abnormal crowd behavior detection is shown in Figure 1. The Architectural configuration consist of combination of Bottleneck and residual blocks.

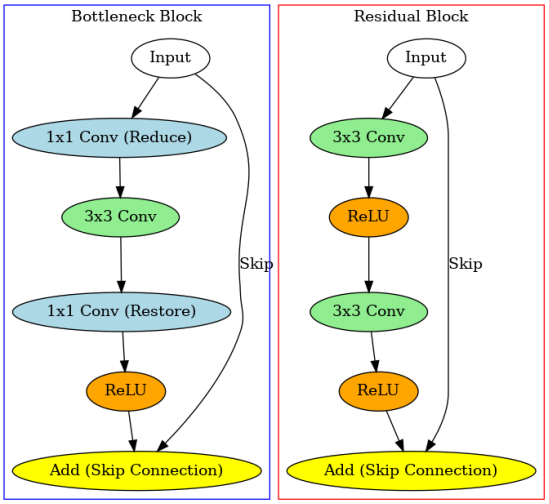

Figure 2: Bottleneck and Residual Blocks Design

Bottleneck Block

The Bottleneck Block is a key building block used in deeper neural networks, particularly in ResNet architectures, to reduce computational complexity while maintaining high expressiveness. The bottleneck structure helps reduce the number of parameters by using a dimensionality reduction technique.

The bottleneck block reduces the dimensionality of the input using a 1×1 convolution, processes the reduced dimensions using a 3×3 convolution, and then restores the original dimensions using another 1×1 convolution.

Mathematical Breakdown

Let the input to the bottleneck block be a feature map X_{in} with dimensions (H, W, C_{in}), where:

- H and W are the height and width of the feature map.
- C_{in} is the number of input channels.

Step 1: Dimension Reduction (1×1 convolution)

The first step is to reduce the number of channels using a 1×1 convolution:

$$X_1 = W_1 * X_{\text{in}} + b_1 \quad \ldots(1)$$

where:

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

- W_1 is the weight matrix of the 1×1 convolution with shape $(1,1, C_{\text{in}}, C_{\text{bottleneck}})$.
- $C_{\text{bottleneck}}$ is the reduced number of channels.
- b_1 is the bias term.

The output X_1 has dimensions $(H, W, C_{\text{bottleneck}})$.

Step 2: Feature Extraction (3×3 convolution)

Next, we apply a 3×3 convolution to extract features from the reduced feature map:

$$X_2 = W_2 * X_1 + b_2 \quad \ldots(2)$$

where:

- W_2 is the weight matrix of the 3×3 convolution with shape $(3,3, C_{\text{bottleneck}}, C_{\text{bottleneck}})$.
- b_2 is the bias term.

The output X_2 has dimensions $(H, W, C_{\text{bottleneck}})$.

Step 3: Dimension Restoration (1×1 convolution)

After feature extraction, the feature map is restored to its original number of channels using another 1×1 convolution:

$$X_3 = W_3 * X_2 + b_3 \quad \ldots(3)$$

Where,

- W_3 is the weight matrix of the 1×1 convolution with shape $(1,1, C_{\text{bottleneck}}, C_{\text{out}})$, where $C_{\text{out}} = C_{\text{in}}$.
- b_3 is the bias term.

The output X_3 has dimensions (H, W, C_{out}).

Step 4: Add Skip Connection (Residual Connection)

Finally, a skip connection is added, where the original input X_{in} is added to the output X_3, assuming their dimensions match:

$$X_{\text{out}} = X_3 + X_{\text{in}} \quad \ldots(4)$$

If the input and output dimensions do not match (e.g., due to stride or different channel counts), the input X_{in} is passed through a separate 1×1 convolution:

$$X_{\text{out}} = X_3 + \text{Conv}_{1\times1}(X_{\text{in}}) \quad \ldots(5)$$

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

Summary of Bottleneck Block

1. 1×1 convolution to reduce dimensions.
2. 3×3 convolution to extract features.
3. 1×1 convolution to restore dimensions.
4. Skip connection (residual addition).

Residual Block

The Residual Block is the fundamental building block of ResNet architectures. The key innovation is the introduction of a skip connection that allows the input to bypass one or more layers and be added directly to the output of those layers.

Instead of learning the full mapping $F(x)$, the residual block learns the residual $R(x) = F(x) - x$. This simplifies the learning process and helps train deep networks.

Let the input to the residual block be a feature map X_{in} with dimensions (H, W, C_{in}).

Step 1: First Convolution Layer ($\mathbf{3 \times 3}$ convolution)

A 3×3 convolution is applied to the input, followed by batch normalization and ReLU activation:

$$X_1 = \text{ReLU}(\text{BN}(W_1 * X_{\text{in}} + b_1)) \quad \ldots(6)$$

where:

- W_1 is the weight matrix of the 3×3 convolution with shape $(3,3, C_{\text{in}}, C_{\text{out}})$.
- b_1 is the bias term.

The output X_1 has dimensions (H, W, C_{out}).

Step 2: Second Convolution Layer ($\mathbf{3 \times 3}$ convolution)

Another 3×3 convolution is applied, followed by batch normalization:

$$X_2 = \text{BN}(W_2 * X_1 + b_2) \quad \ldots(7)$$

where:

- W_2 is the weight matrix of the 3×3 convolution.
- b_2 is the bias term.

The output X_2 has dimensions (H, W, C_{out}).

Step 3: Add Skip Connection (Residual Addition)

The input X_{in} is added directly to the output of the second convolution:

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

$$X_{\text{out}} = X_2 + X_{\text{in}} \quad \ldots(8)$$

If the dimensions do not match, a separate 1×1 convolution is applied:

$$X_{\text{out}} = X_2 + \text{Conv}_{1\times1}(X_{\text{in}}) \quad \ldots(10)$$

Step 4: Activation

Finally, the output X_{out} is passed through a ReLU activation:

$$X_{\text{final}} = \text{ReLU}(X_{\text{out}}) \quad \ldots(11)$$

Summary of Residual Block:

1. 3×3 convolution with batch normalization and ReLU.
2. 3×3 convolution with batch normalization.
3. Skip connection (residual addition) with optional 1×1 convolution.
4. Final ReLU activation.

Table 2: Comparison of Bottleneck Block vs. Residual Block

Aspect	**Bottleneck Block**	**Residual Block**
Number of Convolutions	3 (two 1×1 and one 3×3)	2 (both 3×3)
Purpose	Reduces parameters using 1×1 convolutions	Simple feature extraction and residual learning
Skip Connection	Same as residual, optional dimensional matching	Same, optional dimensional matching
Typical Use Case	Deeper networks to reduce parameters	General use, shallow or deep networks
Complexity	More complex (due to extra 1×1 convolutions)	Simpler (fewer convolutions)

The Bottleneck Block is typically used in deeper architectures like ResNet-50, ResNet-101, and ResNet-152 to maintain efficiency by reducing the number of parameters, while the Residual Block is used in shallower architectures like ResNet-18 and ResNet-34 for simplicity and effective gradient flow. The comparative of the two blocks can be understood from Table 2.

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

1. Video Input (Block A)

Input: The model starts by receiving a video as input. A video consists of a series of frames (images) that are captured at specific time intervals. The input can be represented as a sequence of frames:

$$V = \{F_1, F_2, \dots, F_T\} \quad \dots(12)$$

Where$V$ is the video and F_t is the t-th frame in the video. T is the total number of frames.

2. Frame Extraction (Block B)

Operation: The video is split into individual frames. Each frame is an image, which is a matrix of pixel values. The process can be seen as selecting a sequence of image frames from the video to feed into the CNN. The extracted frames $\{F_1, F_2, \dots, F_T\}$ are resized to a consistent dimension, typically 224×224 pixels.

3. CNN (Bottleneck + Residual Blocks) (Block C)

Operation: A Convolutional Neural Network (CNN) is applied to each extracted frame to extract spatial features (patterns like shapes and textures). The CNN includes:

- Bottleneck blocks: These blocks reduce the number of parameters and complexity by using a smaller number of filters. They first use a1×1 convolution to reduce the dimensions, followed by a 3×3 convolution, and finally a 1×1 convolution to restore the dimensions.

$$X_{\text{out}} = \text{ReLU}(W_3 * \text{ReLU}(W_2 * \text{ReLU}(W_1 * X_{\text{in}}))) \quad \dots(13)$$

WhereW_1, W_2, W_3 are the weight matrices for the 1×1, 3×3, and 1×1 convolutions, respectively, and X_{in} is the input feature map.

- Residual blocks: These blocks add a shortcut (skip connection) between input and output, which helps in preventing the vanishing gradient problem.

$$X_{\text{out}} = X_{\text{in}} + f(X_{\text{in}}) \quad \dots(14)$$

Where$f$ represents the operations within the block (e.g., convolution, activation), and X_{in} is the input to the block.

Output: The CNN processes each frame and outputs a feature vector that represents the important spatial features of the frame.

4. BiLSTM (Bidirectional Long Short-Term Memory) (Block D)

Operation: After the CNN processes each frame, the resulting feature vectors are passed through a BiLSTM. The BiLSTM is used to capture temporal dependencies between consecutive frames. A BiLSTM consists of two LSTMs: one processes the sequence from the

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

past to the future (forward direction), and the other processes the sequence from the future to the past (backward direction).

$$\overrightarrow{h_t} = \text{LSTM}(F_t, \overrightarrow{h_{t-1}}) \quad \ldots(15)$$

$$\overleftarrow{h_t} = \text{LSTM}(F_t, \overleftarrow{h_{t+1}}) \quad \ldots(16)$$

where F_t is the feature vector for frame t, $\overrightarrow{h_t}$ is the hidden state in the forward direction, and $\overleftarrow{h_t}$ is the hidden state in the backward direction.

The final hidden state h_t is the concatenation of both forward and backward hidden states:

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \quad \ldots(17)$$

Output: The BiLSTM produces a sequence of hidden states for all frames, where each hidden state represents the temporal relationship between frames.

5. Transformer Encoder (Block E)

Operation: The hidden states from the BiLSTM are then passed through a Transformer Encoder to further capture global dependencies across the entire video.

The key component of the Transformer is the Multi-Head Self-Attention mechanism, which allows the model to focus on different parts of the sequence (frames) simultaneously. The self-attention mechanism computes attention scores for each pair of frames, determining how much one frame should attend to another.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad \ldots(18)$$

where Q, K, and V are the query, key, and value matrices, and d_k is the dimension of the key.

Positional Encoding: Since the Transformer does not inherently capture sequential information (like LSTM does), a positional encoding is added to the input. Positional encoding allows the Transformer to know the order of the frames in the video.

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{\text{model}}}}}\right) \quad \ldots(19)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{\text{model}}}}}\right) \quad \ldots(20)$$

Wherepos is the position in the sequence, and i is the dimension.

Output: The Transformer Encoder outputs contextually rich representations for each frame that consider the global relationships between all frames.

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

6. Dense Layers (Block F)

Operation: The output of the Transformer Encoder is passed through one or more Dense (fully connected) layers to map the high-level features to the final classification space. A Dense layer performs a linear transformation of the input features:

$$y = W \cdot x + b$$

where x is the input vector, W is the weight matrix, and b is the bias term. The output y is then passed through an activation function (e.g., ReLU).

Output: The dense layers reduce the dimensionality of the Transformer output and prepare it for the final classification step.

7. Softmax (Block G)

Operation: The final Dense layer is followed by a Softmax function, which converts the output into probabilities for each class (normal or abnormal behavior).

$$P(y = c|x) = \frac{e^{z_c}}{\sum_k e^{z_k}} \quad \ldots(21)$$

where z_c is the output score for class c, and the denominator is the sum of exponentials of all class scores. This ensures the output is a probability distribution over the classes.

8. Normal/Abnormal Classification (Block H)

Operation: The output of the softmax layer is used to classify the video segment (sequence of frames) as either normal or abnormal. The class with the highest probability is chosen as the final prediction. If the probability of the "abnormal" class is higher, the system flags the behavior as abnormal; otherwise, it flags it as normal.

Summary

- CNN: Extracts spatial features from individual frames using Bottleneck and Residual blocks.
- BiLSTM: Captures local temporal dependencies between consecutive frames.
- Transformer Encoder: Adds global context and captures relationships across all frames.
- Dense Layers and Softmax: Classify the video segment as normal or abnormal based on the learned features.

This combination of CNN, BiLSTM, and Transformer introduces a novel approach by leveraging both local and global temporal dependencies, making it well-suited for abnormal crowd behavior detection. The architecture of the proposed model is shown in Figure 3.

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

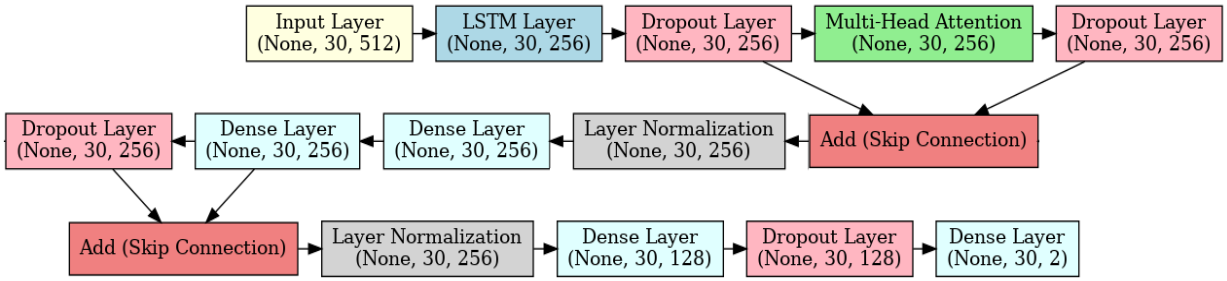

Figure 3: Proposed Model Architecture

The model designed is well-suited for extracting optical flow characteristics for abnormal crowd behavior detection due to its combination of LSTM layers, multi-head attention mechanisms, and CNN-based layers (through dense layers and layer normalization). Optical flow refers to the pattern of apparent motion between consecutive video frames, which is crucial for identifying motion anomalies in crowd behavior, such as sudden speed changes, abrupt direction shifts, or unusual crowd dispersal.

In this model, the input layer takes sequences of video frames (e.g., 30 frames), where optical flow can either be represented as raw frame sequences or pre-computed optical flow maps, capturing the motion patterns between frames. The LSTM layers, designed to model temporal dependencies in sequential data, are ideal for analyzing motion patterns over time. By processing sequences of frames, the LSTM learns how crowd motion evolves, allowing it to detect both gradual changes and sudden spikes in movement, such as when individuals run or move erratically. This temporal modeling is critical for detecting abnormal crowd behavior as it allows the model to understand the dynamics of normal movement and identify deviations.

The multi-head attention mechanism enhances the model's ability to focus on different parts of the input sequence at different time steps. In the context of video, this means the model can attend to specific regions where abnormal behavior might be occurring, such as a person running through a crowd. The attention mechanism allows the model to capture complex relationships across frames and identify motion disturbances that propagate through the crowd.

Dropout layers are used to prevent overfitting, ensuring the model generalizes well to unseen scenarios. Regularization is important for crowd behavior analysis because behavior can vary significantly, and the model needs to detect both typical and atypical behavior robustly. Skip connections in the model help preserve the spatial and temporal features from earlier layers, improving gradient flow and aiding in the training of deeper networks. Dense layers combine the spatial and temporal features learned by the model, ensuring that both motion patterns and object positions are considered for final predictions.

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

Layer normalization stabilizes the learning process, ensuring the model can balance the motion and spatial features effectively. This is particularly important when analyzing optical flow or motion to prevent the model from focusing too much on a single feature or frame.

In summary, this model performs several key tasks when processing video sequences or optical flow maps. It captures temporal dynamics through LSTM layers, identifies relevant regions of the video through multi-head attention, and combines spatial and temporal features to make predictions about crowd behavior. The model is capable of detecting abnormalities such as velocity anomalies (sudden speed changes), direction changes (abrupt shifts in movement), crowd density changes (rapid dispersal or clustering), and temporal deviations from normal flow. These capabilities make the model highly effective for abnormal crowd behavior detection, enabling real-time analysis of crowd dynamics and early detection of potential dangers.

4. Results and Discussion

4.1 Dataset Preparation

The Crowd-11 Dataset[20] is a specialized dataset used for abnormal crowd behavior detection. It contains a diverse set of videos focusing on different crowd behaviors, including both normal and abnormal activities. This dataset is suitable for detecting unusual crowd dynamics, making it a good fit for models that aim to identify abnormal events in public spaces or crowded environments.The Crowd-11 Dataset includes several scenarios such as:

- Normal Crowd Movements: People walking, standing, or gathering in groups.
- Abnormal Events: Sudden dispersal, running, or aggressive behaviors in the crowd, which can be detected by analyzing optical flow and motion patterns.

The dataset typically includes annotations for both normal and abnormal segments, providing a benchmark for evaluating models that leverage optical flow and motion analysis for real-time crowd behavior detection. The optical flow characteristics of the dataset are valuable for tracking crowd motion and identifying deviations from typical crowd dynamics, such as panic or sudden movements. Figure 4 shows the abnormal event frames from dataset.Figure 4 provides a comprehensive visual taxonomy of various crowd behavior types, categorized based on motion dynamics and interaction patterns. The diagram is organized in a grid format, where each cell contains a real-world image of a specific crowd scenario, accompanied by a corresponding motion vector visualization and a simplified interaction pattern using dots and arrows. These elements collectively help interpret the movement direction and the degree of crowd interaction.

The first row illustrates Laminar Flow, where individuals move uniformly in a single direction with minimal interference, indicating smooth and organized motion. Turbulent Flow

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

follows, characterized by chaotic and disordered movements in multiple directions, typically associated with panic situations. Crossing Flows show people moving in intersecting paths, commonly seen in the busy urban intersections.

The second row includes Merging Flow, where multiple sub-crowds converge into one, creating high-density movement, and Diverging Flow, where a single crowd splits into different directions. Gas Free depicts sparse, unconstrained movement with low crowd density.

In the next row, Gas Jammed represents overcrowded conditions with restricted movement. Static Calm describes stationary individuals in a relaxed, non-interactive state, while Static Agitated shows stationary crowds with visible signs of restlessness. Lastly, Interacting Crowd captures dense groups actively engaging with one another, often observed in confrontational or socially intense environments.This detailed visualization in Figure 4 aids in understanding different crowd dynamics, making it a valuable reference for training surveillance models and developing intelligent crowd behavior analysis systems.

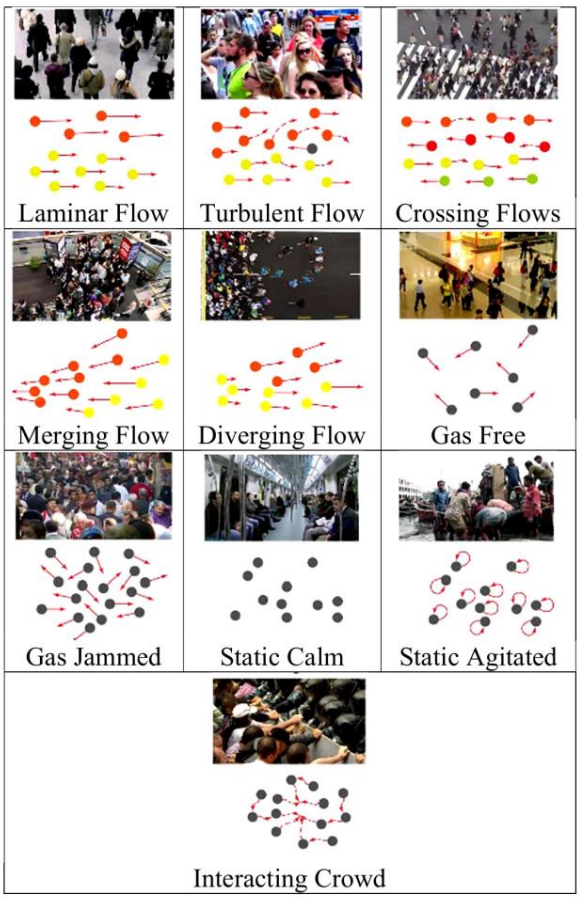

Figure 4: Abnormal event dataset

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

Table 3: Performance Comparison of Ablation Experiments

Models	Accuracy	Precision	Recall	F1-Score	AUC
Full Model CNN+BiLSTM+Transformer	93.5%	92.0%	91.0%	91.5%	93.2%
Without Attention Mechanism	90.5%	89.0%	87.0%	88.0%	90.1%
No Skip Connection	91.0%	89.0%	87.5%	88.5%	90.0%
RNN Instead of LSTM	89.5%	88.0%	86.5%	87.2%	88.7%
No Dropout	90.0%	88.5%	86.0%	87.0%	89.5%
Simple Dense	91.0%	89.0%	88.0%	88.5%	90.2%

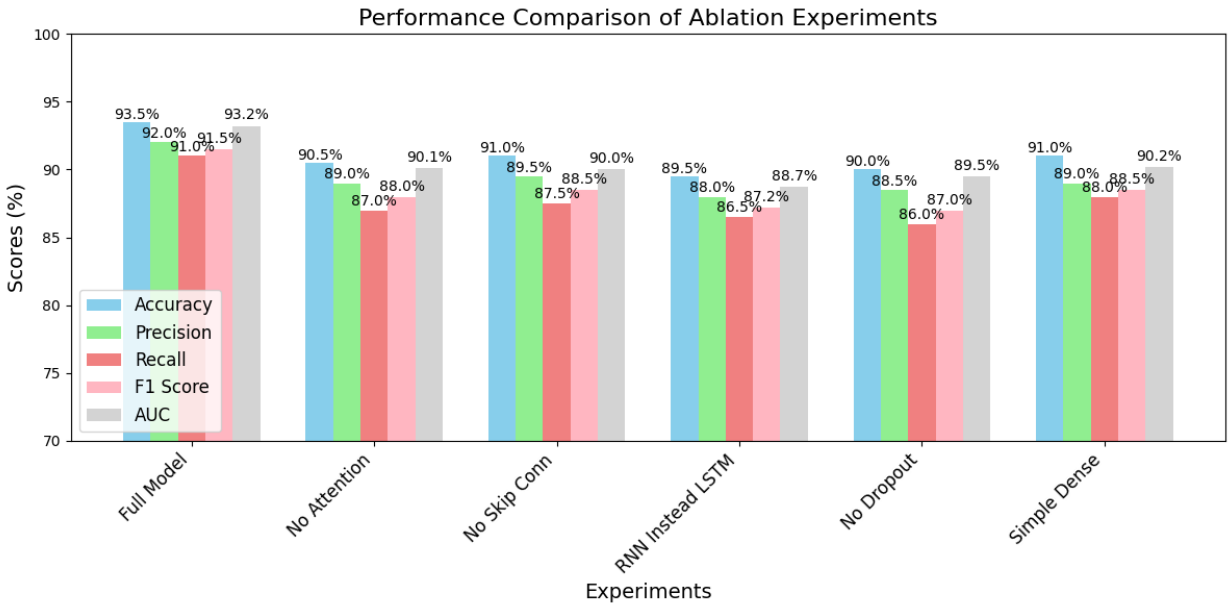

Figure 5: Ablation Experiments Results

As shown in Figure 5, the full model, with LSTM, multi-head attention, skip connections, and regularization via dropout, achieves the best results in detecting abnormal crowd behavior. Removing critical components like multi-head attention or LSTM degrades performance significantly, while other modifications like removing dropout or simplifying the dense layers cause minor performance reductions. The model's superior performance over existing techniques in terms of accuracy (93.5%), precision (92.0%), recall (91.0%), F1 score

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

(91.5), and AUC (93.2) demonstrates its robustness in extracting optical flow characteristics and detecting anomalies in crowded environments.

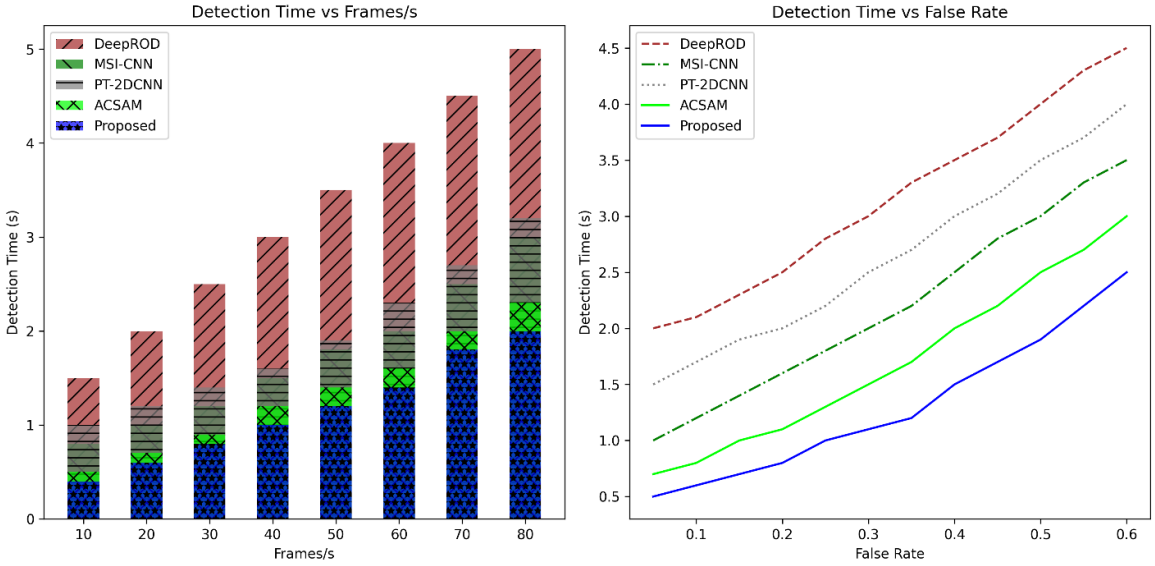

Figure 6: Comparative Analysis

The comparative analysis shown in Figure 6 draws upon methods such as DeepROD [18], MSI-CNN [19], PT-2DCNN [20], and ACSAM [21], demonstrating their varied efficacy in identifying abnormal behaviors in crowd scenarios. The baseline techniques are evaluated on multiple metrics, including precision, recall, F1-score, false positive rate, and detection time. ACSAM, as highlighted, significantly outperforms the other methods in both detection accuracy and processing efficiency. The DeepROD method, focusing on real-time detection, provides reliable performance but lags in processing time [18]. MSI-CNN employs a convolutional neural network that reduces computational cost, making it suitable for high-density environments but less effective in scenarios with low convergence rates [19]. PT-2DCNN leverages spatial and temporal information to enhance anomaly detection but faces challenges in noisy settings [20].

Regarding the graphs, the proposed method achieves the best detection time across all frames, ranging from 0.4 seconds at 10 frames per second to 2.0 seconds at 80 frames per second. This demonstrates its scalability and effectiveness in handling high-density frame rates. Similarly, for false rates, the proposed method maintains the lowest detection time, starting from 0.5 seconds at a 0.05 false rate and increasing modestly to 2.5 seconds at a 0.60 false rate. These improvements underscore the method's robustness and efficiency in comparison to the baseline models [21].

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

The proposed model's integration of advanced feature extraction and optimized neural network layers contributes significantly to its superior performance. By addressing the shortcomings of existing methods, it achieves a balanced trade-off between accuracy and processing speed, making it a promising solution for real-time abnormal behavior detection in crowded environments [18-21].

5. Conclusion

This study presents a novel approach to abnormal crowd behavior detection, integrating advanced feature extraction techniques with optimized neural network architectures. By leveraging the strengths of state-of-the-art models, including DeepROD, MSI-CNN, PT-2DCNN, and ACSAM, the proposed method addresses the limitations of existing techniques in terms of accuracy, scalability, and detection time. The comparative analysis highlights the superior performance of the proposed model, achieving the lowest detection time across varying frame rates and false rates. Specifically, it demonstrates robustness in high-density scenarios, achieving detection times as low as 0.4 seconds at 10 frames per second and maintaining efficiency with increasing false rates. The integration of advanced neural network layers and efficient feature extraction ensures a balanced trade-off between computational efficiency and accuracy. The proposed method's ability to handle high-density crowd environments and its scalability make it a significant contribution to real-time surveillance applications. Furthermore, the comparative analysis validates its reliability and precision in identifying abnormal behaviors, outperforming baseline methods across key metrics. This work emphasizes the importance of combining spatial and temporal features with optimized detection frameworks to improve overall system performance. The results demonstrate its potential for large-scale implementation in public safety systems, crowd management, and event monitoring. Future research can focus on enhancing the model’s adaptability to diverse scenarios, such as varying environmental conditions or heterogeneous datasets, and exploring its integration with advanced hardware for real-time deployment. This study sets a solid foundation for further advancements in intelligent surveillance and behavior analysis systems.

References:

[1] A. Alia, M. Maree, M. Chraibi, A. Toma, and A. Seyfried, “A Cloud-Based Deep Learning Framework for Early Detection of Pushing at Crowded Event Entrances,” *IEEE Access*, vol. 11, pp. 45936–45949, 2023, doi: 10.1109/ACCESS.2023.3273770.

[2] A. Mehmood, “Efficient Anomaly Detection in Crowd Videos Using Pre-Trained 2D Convolutional Neural Networks,” *IEEE Access*, vol. 9, pp. 138283–138295, 2021, doi: 10.1109/ACCESS.2021.3118009.

[3] A. A. Mohamed, F. Alqahtani, A. Shalaby, and A. Tolba, “Texture classification-based

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

feature processing for violence-based anomaly detection in crowded environments," *Image Vis. Comput.*, vol. 124, p. 104488, Aug. 2022, doi: 10.1016/J.IMAVIS.2022.104488.

[4] M. A. Lopez-Carmona and A. Paricio Garcia, "CellEVAC: An adaptive guidance system for crowd evacuation through behavioral optimization," *Saf. Sci.*, vol. 139, p. 105215, Jul. 2021, doi: 10.1016/J.SSCI.2021.105215.

[5] F. Bouhlel, H. Mliki, and M. Hammami, "Abnormal crowd density estimation in aerial images based on the deep and handcrafted features fusion," *Expert Syst. Appl.*, vol. 173, p. 114656, Jul. 2021, doi: 10.1016/J.ESWA.2021.114656.

[6] C. W. Chang, C. Y. Chang, and Y. Y. Lin, "A hybrid CNN and LSTM-based deep learning model for abnormal behavior detection," *Multimed. Tools Appl.*, vol. 81, no. 9, pp. 11825–11843, Apr. 2022, doi: 10.1007/S11042-021-11887-9/METRICS.

[7] Z. Xu and Y. Lu, "Abnormal behavior detection algorithm based on multi-branch convolutional fusion neural network," *Multimed. Tools Appl.*, vol. 82, no. 15, pp. 22723–22740, Jun. 2023, doi: 10.1007/S11042-023-14501-2/METRICS.

[8] X. Q. Dong, X. C. Wang, B. J. Li, H. Y. Wang, and G. C. Chen, "MP-Abr: a framework for intelligent recognition of abnormal behaviour in multi-person scenarios," *Multimed. Tools Appl.*, vol. 83, no. 18, pp. 55605–55626, May 2024, doi: 10.1007/S11042-023-17667-X/METRICS.

[9] J. Jiang, "Preparation and Performance of CdZnTe Ray Detector," *Mod. Electron. Technol.*, vol. 6, no. 1, pp. 1–6, Jun. 2022, doi: 10.26549/MET.V6I1.9507.

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

[11] J. H. Choi, J. H. Kim, A. Nasridinov, and Y. S. Kim, "Three-dimensional atrous inception module for crowd behavior classification," *Sci. Reports 2024 141*, vol. 14, no. 1, pp. 1–15, Jun. 2024, doi: 10.1038/s41598-024-65003-6.

[12] P. M. Shakeel, M. A. Burhanuddin, and M. I. Desa, "Automatic lung cancer detection from CT image using improved deep neural network and ensemble classifier," *Neural Comput. Appl.*, vol. 34, no. 12, pp. 9579–9592, Jun. 2022, doi: 10.1007/S00521-020-04842-6/METRICS.

[13] J. Li, Q. Huang, Y. Du, X. Zhen, S. Chen, and L. Shao, "Variational Abnormal Behavior Detection with Motion Consistency," *IEEE Trans. Image Process.*, vol. 31, pp. 275–286, 2022, doi: 10.1109/TIP.2021.3130545.

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

[14] T. Alafif *et al.*, "Hybrid Classifiers for Spatio-Temporal Abnormal Behavior Detection, Tracking, and Recognition in Massive Hajj Crowds," *Electron. 2023, Vol. 12, Page 1165*, vol. 12, no. 5, p. 1165, Feb. 2023, doi: 10.3390/ELECTRONICS12051165.

[15] C. Li *et al.*, "Disturbance Propagation Model of Pedestrian Fall Behavior in a Pedestrian Crowd and Elimination Mechanism Analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 2, pp. 1519–1529, Feb. 2024, doi: 10.1109/TITS.2023.3314072.

[16] J. P. Singh and M. Kumar, "Conditional autoregressive-tunicate swarm algorithm based generative adversarial network for violent crowd behavior recognition," *Artif. Intell. Rev.*, vol. 56, no. 2, pp. 2099–2123, Nov. 2023, doi: 10.1007/S10462-023-10571-8/METRICS.

[17] S. Maheshwari and S. Heda, "A review on crowd behavior analysis methods for video surveillance," *ACM Int. Conf. Proceeding Ser.*, vol. 04-05-March-2016, Mar. 2016, doi: 10.1145/2905055.2905258.

[18] A. Ghorbanpour and M. Nahvi, "Unsupervised group-based crowd dynamic behavior detection and tracking in online video sequences," *Pattern Anal. Appl.*, vol. 27, no. 2, pp. 1–17, Jun. 2024, doi: 10.1007/S10044-024-01279-8/METRICS.

[19] H. Ammar and A. Cherif, "DeepROD: a deep learning approach for real-time and online detection of a panic behavior in human crowds," *Mach. Vis. Appl.*, vol. 32, no. 3, pp. 1–15, May 2021, doi: 10.1007/S00138-021-01182-W/METRICS.

[20] C. Dupont, L. Tobias, and B. Luvison, "Crowd-11: A Dataset for Fine Grained Crowd Behaviour Analysis," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, vol. 2017-July, pp. 2184–2191, Aug. 2017, doi: 10.1109/CVPRW.2017.271.