IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 9 November 2024, accepted 7 January 2025, date of publication 17 January 2025, date of current version 24 January 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3531213

APPLIED RESEARCH

Violence Detection From Industrial Surveillance Videos Using Deep Learning

HAMZA KHAN, XIAOHONG YUAN, LETU QINGGE, AND KAUSHIK ROY, (Senior Member, IEEE)
Department of Computer Science, North Carolina Agricultural and Technical State University, Greensboro, NC 27411, USA

Corresponding author: Hamza Khan (hkhan2@aggies.ncat.edu)

This work was supported in part by the National Science Foundation (NSF) under Award 1900187, Award 2146498, and Award 2434487.

ABSTRACT The integration of Internet of Things (IoT) technology in industrial surveillance and the proliferation of surveillance cameras in smart cities has empowered the development of real-time activity recognition and violence detection systems, respectively. These systems are crucial in enhancing safety measures, improving operational efficiency, reducing accident risks, and providing automatic monitoring in dynamic environments. In this paper, we propose a three-stage deep learning-based end-to-end framework for violence detection. The lightweight convolutional neural network (CNN) model initially identifies individuals in the video stream to minimize the processing of irrelevant frames. Subsequently, a sequence of 50 frames with identified persons is directed to a 3D-CNN model, where the spatiotemporal features of these sequences are extracted and passed to the classifier. Unlike traditional methods that process all frames indiscriminately, this targeted filtering mechanism allows computational resources to be allocated more effectively. Next, SoftMax classifier processes the extracted features to categorize frame sequences as violent or non-violent. The classifier's predictions trigger real-time alerts, enabling rapid intervention. The modularity of this stage supports adaptability to new datasets, as it can leverage transfer learning to generalize across diverse surveillance contexts. Unlike traditional systems constrained by hand-crafted features, this design dynamically learns from data, reducing reliance on prior domain knowledge and improving generalizability. We conducted experiments on violence detection across four datasets, comparing the performance of our model with convolutional CNN models. A computation time analysis revealed that our lightweight model requires significantly less computation time, demonstrating its efficiency. We also conducted cross-data experiments to assess the model's capacity to perform consistently across various datasets. Experiments show that our proposed model outperforms the methods mentioned in the existing literature. These experiments demonstrate that the model's adaptability and robustness need to be improved.

INDEX TERMS Activity detection, industrial surveillance, violence detection, computer vision, deep learning.

I. INTRODUCTION

The use of Internet of Things (IoT) technology for activity recognition in industrial surveillance has become increasingly important in recent years [1]. This is due to the need for real-time monitoring and analysis of workers' activities to ensure safety, improve efficiency, and reduce the risk of accidents. With the surge in video data due to the ubiquity of surveillance cameras and mobile devices, violence detection has emerged as a critical application area in computer vision. The goal of violence detection is to recognize violent activities or behaviors automatically, making it crucial for public safety and security. However, traditional activity recognition methods may not be sufficient for recognizing activities in complex and dynamic industrial environments. Many techniques based on deep learning features [2], [3], [4] have emerged.

Over the past few years, many assessments were conducted by researchers on deep learning approaches [12], and deep learning approaches has shown remarkable success in

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenhua Guo.

violence detection, outperforming traditional methods. Deep learning models can automatically extract representative features from raw data, offering a major advantage over traditional machine learning approaches. For instance, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been widely used to identify spatial and temporal features in video data, respectively.

In recent studies, models such as the 3D Convolutional Neural Networks (3D-CNNs) and Long Short-Term Memory (LSTM) networks have demonstrated promising results in violence detection. 3D-CNNs are a variation of CNNs designed to handle three-dimensional data such as videos. Unlike traditional 2D-CNNs, which operate on static images and understand height and width, 3D-CNNs also consider the temporal dimension (time) in videos. This is accomplished by performing convolutions in three dimensions - height, width, and time - enabling the 3D-CNNs to capture both spatial features (such as objects and their shapes) and temporal features (such as the motion of these objects over time). This makes 3D-CNNs especially useful for video analysis tasks like action recognition or violence detection, since they can comprehend the video as a continuous sequence of events rather than a collection of individual frames. For instance, a study by Muhammad et al. [18] employed a 3D-CNN model on a large video dataset and achieved an accuracy of 74.2%.

Existing models of neural networks are highly promising in identifying human actions in a variety of applications, from healthcare to entertainment. However, when it comes to detecting violent activities in complex real-world situations, these models face significant challenges. The accuracy of these models is often compromised by factors such as background noise, visual distractions that include moving objects, and varying lighting conditions that result in false positives and false negatives being reported.

Another challenge arises from the processing of many irrelevant frames. Traditional models often analyze each frame in the video sequence, which causes unnecessary computational overhead on a computer. This inefficiency is particularly problematic in real-time surveillance applications where rapid and accurate decision-making is crucial. The need for a more focused and computationally efficient approach is evident, especially in real-time applications like violence detection in public spaces or industrial settings.

We propose a three-stage deep learning based end-to-end framework for accurate and efficient violence detection. A lightweight CNN is used as an initial screening mechanism in the first stage. The light-weighted CNN model is optimized for computational efficiency, allowing for fast scanning of incoming video streams. Its main purpose is to recognize and separate frames with human beings, reducing the requirement to analyze unnecessary frames. Selective frame processing considerably minimizes computing cost and accelerates detection. The second stage employs a more detailed 3D-CNN model. This model extracts spatiotemporal information from 50 frames including identified people from the first stage. Training the 3D-CNN to identify complicated sequential patterns associated with violent behaviors makes it very effective. A classifier receives the extracted features in the third stage and classifies frame sequences as violent or non-violent. If the classifier identifies violence, it generates alerts. The technology has a real-time warning function that alerts nearby security agencies or police stations of any detected violence. This allows for quick response and intervention, bringing practicality to the system.

We used four different datasets to analyze the performance of the proposed model on violence detection, i.e., RWF-2000, the hockey fight dataset, the surveillance fight dataset, and the Industrial Surveillance dataset [22]. The proposed 3D-CNN method demonstrates superior performance compared to other machine learning models, including ConvLSTM, across all tested datasets. This highlights its effectiveness for real-world violence detection applications.

We also conducted cross-dataset testing to evaluate the ability of the proposed model to generalize across various datasets. The primary objective is to assess whether the model has a sufficiently comprehensive understanding of violence to identify violent actions across diverse datasets, which is crucial for practical applications in a range of scenarios with varied data. These tests offer insights into the model's adaptability and robustness, confirming its potential for widespread use in different monitoring situations. Additionally, a computation time analysis was performed, highlighting the efficiency of our approach to processing video data.

The rest of the paper is organized as follows: Section II covers the background of human action recognition and violence detection. Section III discussed violence detection using deep learning Methods. The proposed methodology is discussed in Section IV and Section V covers results and analysis. Section VI concludes the paper and discusses future work. Section VII includes acknowledgment.

II. BACKGROUND

This section reviews previous work in human action recognition, the application of deep learning in human action recognition and violence detection in video data.

A. HUMAN ACTION RECOGNITION

Automated methods of video sequence analysis and decision-making regarding the behaviors shown in videos are used in human activity detection for video surveillance systems. In 1999, Gavrilla created the research field of 2D and 3D approaches [1] for the development of human action recognition (HAR) systems. While the early models were highly reliant on feature extraction from single image or sets of images, they laid the groundwork for more complex systems that integrated spatial and temporal data over time.

A different team of researchers, led by JK Aggarwal and Q Cai, developed a new taxonomy centered on the study of human motion, monitoring from different types of camera views and human activity detection [2]. Two methods exist for HAR: using still images and using video data. Video-based algorithms outperform those based on still images, as they convey both spatial and temporal information. Videos capture continuous movements and interactions, which are essential for distinguishing between different types of actions, such as walking, running, or violent movements. This temporal data provides context that static images cannot, making it indispensable for effective HAR systems.

However, relying solely on video-based approaches also presents challenges. These algorithms need to handle large amounts of data efficiently, as processing continuous streams of video can be computationally expensive. In addition, video data is often noisy, making it essential to clean the input before analysis. The process typically starts with noise reduction to eliminate irrelevant information from the video frames. Techniques such as background subtraction are commonly used to isolate the human figure from its surroundings.

Once noise is reduced, the form of a human is extracted from the backdrop images by analyzing sequences of video frames and observing changes in position over time. Human shapes are usually identified through a combination of feature extraction techniques and tracking algorithms. Classification of objects, including humans, is done by evaluating their movement characteristics and shape, using methods such as optical flow, which analyzes the pattern of motion between frames, or shape-based descriptors that focus on the contour and skeletal structure of the human body.

Nonetheless, even with these advancements, traditional HAR methods remain limited by their inability to effectively handle complex scenes with multiple interacting people, poor lighting conditions, and varying camera angles. Additionally, most of these methods rely on handcrafted features, which require domain knowledge and limit their adaptability to new environments or unseen data.

Figure 1 showcases a range of different methods for Human Action Recognition, highlighting the diversity of approaches used in this field [41].

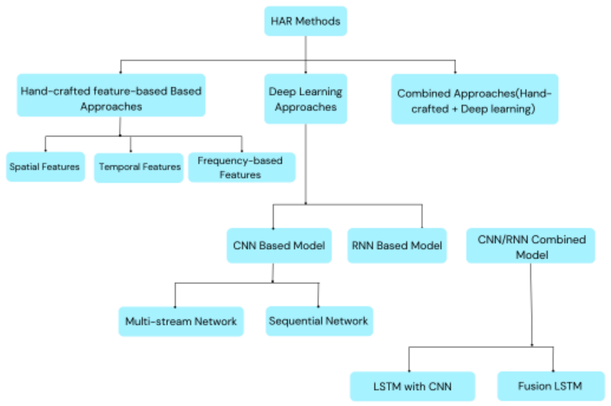


FIGURE 1. **Methods of HAR [41].**

The field has now shifted towards deep learning-based approaches, which have significantly improved HAR systems by automating feature extraction and learning complex spatiotemporal patterns from raw data. Deep learning models, such as CNNs and RNNs, have demonstrated superior performance in recognizing human actions in videos, especially when integrated with large-scale datasets. For instance, CNNs have been widely used to extract spatial features from video data [17]. In contrast, LSTMs and their variants, such as deep bidirectional LSTMs (BiLSTMs), are useful for capturing temporal dependencies in video sequences, making them crucial for action recognition tasks [22]. Recent studies such as Zhang et al. [5] have shown how deep learning models with attention mechanisms can significantly improve action recognition by focusing on the most relevant spatiotemporal features in video data.

B. STATE OF THE ART APPROACH TO HAR

Wang et al. [17] present a LSTM mechanism that simulates the cognitive memory processes of the human brain for visual monitoring in IoT-assisted smart cities. The primary objective is to enable timely detection of violent actions and prevent false tracking in clear environments. The proposed model utilizes a unique props function within its storage mechanism to perform real-time processing when environmental conditions change and conventional algorithms become ineffective. However, the props function may also introduce limitations in scalability, as it may require frequent updates to account for environmental variations, which could increase computational overhead. Additionally, the model's performance in more complex real-world scenarios, such as crowded industrial environments or highly dynamic urban spaces, remains unexplored. This raises concerns about its robustness in diverse surveillance contexts, particularly where lighting conditions, background clutter, or overlapping objects may impede accurate detection.

Muhammad et al. [18] introduced a HAR method that employs an attention-based LSTM network combined with dilated CNN features. The authors developed a BiLSTM network with an attention mechanism to effectively learn the spatiotemporal properties present in sequential data. This network design enables the model to adjust attention weights, thus allowing it to easily recognize and focus on learned global features for detecting human actions in video data.The use of attention-based mechanisms significantly improves the model's ability to focus on critical spatiotemporal features, making it more efficient in detecting complex actions. The BiLSTM's ability to process data in both directions (forward and backward) provides a more comprehensive understanding of the action sequences, particularly for recognizing subtle behaviors. However, this method also introduces considerable computational complexity, which may make real-time deployment difficult, especially in resource-constrained environments like IoT-based surveillance. Furthermore, the model's reliance on global feature learning could result in missed fine-grained details, particularly in dense environments where actions may occur in close proximity or in highly occluded scenes. The dilated CNN features, while useful for capturing contextual information, may also cause the model to lose sensitivity to smaller, rapid movements that are crucial for violence detection.

Zhang et al. [5] delved into the challenge of few-shot activity recognition using a cross-modal memory network. A cross-modal memory network stores information from multiple modalities (e.g., video, audio) and enables the model to utilize these diverse inputs to improve learning in complex environments, particularly when training data is limited. The proposed model is designed to recognize new videos with a limited number of labeled samples by leveraging visual contextual embedding for few-shot classification. Few-shot learning models are crucial in situations where collecting large, annotated datasets is impractical or costly. By leveraging visual contextual embedding, this approach allows the model to generalize well to unseen activities with minimal supervision. However, the success of few-shot learning heavily depends on the quality and diversity of the training samples. In surveillance contexts, where lighting, camera angles, and occlusions can vary greatly, the limited training data used in few-shot learning may not sufficiently capture these complexities. Moreover, few-shot models tend to struggle with generalization when faced with highly variable or noisy environments. The use of cross-modal [5] memory networks helps in cross-referencing data from multiple modalities, which improves recognition capabilities, but this also increases the demand for high-quality, synchronized data from different modalities, which may not always be available in practical settings.

Haroon et al. [6] put forward a multi-stream framework for human interaction recognition, which aims to capture and analyze complex human interactions in various scenarios. The proposed approach combines a 1D-CNN with BiLSTM stream to learn human interactions based on key features extracted using a pose estimation algorithm, and a 3D-CNN model to learn temporal information from video sequences. The multi-stream framework proposed is advantageous for analyzing multi-person interactions or complex human activities. By combining 1D-CNNs with BiLSTM and 3D-CNN models, the framework can capture both spatial and temporal information, making it more robust in understanding complex human interactions in diverse environments. However, the reliance on pose estimation algorithms has limitations, especially in environments where occlusions, low-resolution video, or poor lighting hinder accurate pose detection. In such cases, misestimations in pose can lead to inaccurate action recognition. Additionally, the computational overhead of running multiple streams (1D-CNN, BiLSTM, 3D-CNN) could limit the framework's scalability, particularly in real-time applications or on low-power edge devices. The fusion of pose estimation and deep learning models is promising but requires further optimization to handle real-world surveillance challenges.

Ullah et al. [22] proposed an AI-assisted edge vision approach for violence detection in IoT-based industrial surveillance networks. The framework comprises five main steps: training a lightweight CNN for efficient edge processing; acquiring data using resource-constrained vision sensors; detecting suspicious humans or objects and generating alerts; sending relevant frames to a more powerful backend for deeper investigation; and finally, using the backend system for accurate violence detection.The edge-based architecture proposed by Ullah et al. [22] addresses one of the main challenges of real-time violence detection in industrial environments: limited computational resources. By utilizing a lightweight CNN for initial processing, the framework minimizes the amount of data that needs to be sent to the backend, which reduces bandwidth usage and latency. This makes the approach particularly well-suited for IoT-based systems where energy and computational resources are constrained. However, the reliance on a more powerful backend for deeper investigation introduces latency that could impact real-time performance in time-critical situations. Additionally, the effectiveness of the lightweight CNN in accurately detecting violence, especially in complex scenes with occlusions or overlapping objects, remains a concern. Another potential limitation is the framework's adaptability to different industrial contexts, where sensor configurations and environmental factors vary significantly.

Chen et al. [8] introduced a spatiotemporal graph convolutional network (ST-GCN) for skeleton-based HAR in surveillance environments. The proposed model captures both spatial and temporal information about human actions by incorporating graph convolutional layers, which effectively model the relationships between different body joints in the skeleton data.The use of ST-GCNs for skeleton-based action recognition offers a highly structured approach to modeling human actions, as it focuses on the movement and relationships of body joints. This approach is particularly useful in environments where clear skeletal data can be extracted, such as sports events or controlled laboratory conditions. However, in real-world surveillance environments, extracting high-quality skeleton data is challenging due to occlusions, varying camera angles, and environmental noise. Furthermore, ST-GCNs rely heavily on accurate skeleton detection, which may not be feasible in industrial or crowded public spaces where body movements are obscured or erratic. Additionally, the performance of this method in recognizing subtle or non-standard movements, such as those seen in violence detection scenarios, may be limited, as the skeletal structure alone may not capture the full context of the action.

Kim and Lee [9] proposed a multi-modal fusion approach for anomaly detection in video streams, integrating both visual and auditory data to improve the overall performance of the system. The authors developed a deep neural network architecture that combines a 3D-CNN for visual feature extraction and a 1D-CNN for auditory feature extraction, followed by a fusion layer that effectively combines the extracted features for better anomaly detection. The integration of auditory data alongside visual data enhances the detection of anomalies, as certain violent actions may be accompanied by distinct sounds. However, the reliance on auditory features introduces additional challenges, especially in noisy environments like factories or urban areas, where background noise may interfere with the detection process. The 1D-CNN for auditory feature extraction, while useful, may require substantial filtering and preprocessing to handle the variability of audio data in these environments. Furthermore, the fusion of multiple modalities increases the computational complexity of the system, which could hinder real-time performance, particularly in resource-constrained settings like edge devices or low-power surveillance systems. Ensuring synchronization between the audio and video streams is another challenge that must be addressed to avoid inconsistencies in anomaly detection.

III. VIOLENCE DETECTION USING DEEP LEARNING METHODS

Deep learning forms the backbone of all the methods discussed in the papers discussed in Section II. In the context of violence detection, CNN-based 3D algorithms have been widely used due to their ability to model both spatial and temporal features simultaneously. The 3D-CNN algorithm is designed to capture the spatiotemporal patterns present in video sequences, making it particularly well-suited for tasks like violence detection, where the timing and sequence of actions are critical. However, the development of a 3D-CNN model often requires complex, hand-crafted algorithms to fine-tune the model for specific applications. These models are computationally intensive, making them less feasible for deployment in real-time or resource-constrained environments. As a result, there is an ongoing effort to develop more efficient 3D-CNN architectures that can deliver high accuracy while maintaining computational efficiency.

By contrast, more powerful models based on newer architectures, such as Transformers or spatiotemporal attention networks, have been developed to automatically extract features and understand complex interactions without the need for extensive manual tuning. These models are capable of analyzing both short- and long-range dependencies in video sequences, making them highly effective for HAR and violence detection tasks. However, these advanced models are still in the experimental stages and face challenges related to computational cost and data requirements.

Table 1 outlines some of the approaches used in violence detection, showcasing the diversity of deep learning models in terms of both accuracy and computational efficiency. For instance, the Visual Geometry Group (VGG-f) model [21], which utilizes the ImageNet method of object detection, is well-suited for real-time detection tasks. Achieving an accuracy range of 91%-94%, this model is effective for detecting objects in crowded environments. However, as a purely spatial method, it lacks the ability to capture the temporal progression of actions, which is critical for understanding violent behavior in video sequences.

On the other hand, the 3D-CNN approach developed by Ding et al. [26] extends the capability of standard CNNs by incorporating temporal information through 3D convolutions. This method achieves an accuracy of 91% and is particularly effective in crowded environments where both spatial and temporal patterns must be considered. However, its reliance on the backpropagation method for training introduces significant computational overhead, which may limit its deployment in real-time scenarios.

The VGG vector of locally aggregated descriptors (VLAD) model for image retrieval, presented by Zhou et al. [36], also uses a backpropagation method and achieves an approximate accuracy of 90% in crowded environments. This method, while powerful for place recognition tasks, is less specialized for violence detection as it does not consider temporal dynamics essential for action recognition in video sequences. Similarly, Karpathy et al. [11] employed a multi-modal approach, combining CNNs with Mel Filter Bank (MFB) audio features, achieving an approximate accuracy of 90%. By integrating audio cues like shouting or alarms, this method enhances violence detection capabilities in crowded environments. However, the need to process both visual and auditory streams introduces additional complexity, particularly when dealing with noisy or cluttered environments.

One of the most promising approaches in the table is the use of ConvLSTM networks for violence detection. Muhammad et al. [18] developed a model combining CNNs with ConvLSTM, achieving approximately 97% accuracy. ConvLSTM is designed to handle both spatial and temporal dependencies, making it highly effective in recognizing violent actions in crowded settings. By leveraging the strengths of both CNNs and LSTMs, this model excels at detecting violent behaviors that unfold over time, which are difficult to capture with purely spatial models.

The highest accuracy in the table is reported by Haroon et al. [6], whose deep CNN combined with optical flow analysis reached 98% accuracy. This method tracks motion trajectories, making it particularly adept at identifying violent actions by analyzing the movement patterns of individuals. While this approach provides exceptional accuracy, the computational demands of combining deep CNNs with optical flow analysis may present challenges for real-time deployment, especially in resource-constrained environments.

IV. THE PROPOSED METHODOLOGY

In this section we discuss the proposed three-staged end-to-end framework in detail. Different types of datasets such as the surveillance fight dataset [15], RWF-2000 Dataset [16], hockey fight dataset [17], and New Industrial Surveillance Dataset [9] were used in experiments.

TABLE 1. Performance comparison of different methods.

Technique	Method	Model	Dataset	Accuracy
Convolutional Neural Network for Real-Time Detection [21]	ImageNet method of object detection	VGG-f model	Violent Flows, UCF Web Abnormality and UMN Abnormal Crowd datasets	91%-94%
Violence Detection Using 3D-CNN [26]	3D convolution for spatial information	Backpropagation method	Hockey dataset	91%
Deep Architecture for Place Recognition [36]	VGG VLAD method for image retrieval	Backpropagation method	Hockey Fight dataset, BEHAVE dataset and Crowd Violence dataset	Approx. 90%
Tracking Violence Sites Using CNN and Deep Audio Features [11]	MFB (Mel Filter Bank)	CNN model	Sports-1M dataset	Approx. 90%
Detecting Violence Using ConvLSTM [18]	CNN with ConvLSTM	CNN model	UCF11, UCF Sports, and J-HMDB dataset	Approx. 97%
Detecting Human Violent Behavior by Integrating Trajectory and Deep CNN [6]	Deep CNN	Optical flow method	NTU-RGB+D and HuDaAct-RGBD	98%

A. THE PROPOSED THREE STAGE VIOLENCE DETECTION FRAMEWORK

Figure 2 shows the proposed three-stage violence detection framework. In the first Stage, a lightweight CNN is used to scan the video stream rapidly since it is computationally efficient. Its main purpose is to segregate frames with human beings, reducing the requirement to filter unnecessary frames. Selective frame processing considerably minimizes computational overhead and accelerates detection. A more complicated 3D-CNN model is used in the second stage. This model extracts spatiotemporal information from 50 frames including people identified in the first stage.. Training the 3D-CNN to recognize complicated sequential patterns associated with violent behaviors makes it effective. A Softmax classifier receives the extracted features in the third stage and classify frame sequences as violent or non-violent. If the classifier identifies violence, it generates alerts. The technology automatically alerts the nearest security agency or police station if violence is detected. This allows immediate response and intervention, making the system useful for the real-world situations.

Given the challenges associated with limited training data, the model also employs transfer learning techniques. This allows the model to generalize better across different scenarios and increases its overall performance.

By employing this multi-stage, multi-modal approach, the proposed model aims to combine computational efficiency with high accuracy, making it well-suited for practical, real-world applications. Edge computing is used to run the model directly on IoT devices, enabling real-time processing and immediate response to violent acts.

The following provides a detailed explanation of the model's workflow:

- Video Capture: The initial stage involves capturing real-time video data, which is sourced from various types of videos capturing devices, including but not limited to, surveillance cameras. This raw video data is critical, as it serves as the foundational input for the entire model. The quality, frame rate, audio, and resolution of this video data are crucial factors that can significantly influence the model's performance in subsequent stages.
- Person Detection using MobileNet CNN [9]: The first objective in the model's workflow after capturing a video is to identify the person in the sequences. This crucial action prepares the ground for the later identification of violence. We use the MobileNet Single Shot MultiBox Detector (MobileNet-SSD) architecture, a CNN that is known for its fast-processing speed and minimal computing requirements, to achieve this. The use of depth-wise separable convolutional layers rather than conventional convolutional layers distinguishes this design from others. The network is divided into 2819 layers, except for the last layer, which is completely linked, each layer being followed by a batch normalization procedure and a ReLU activation. Figure 3 shows detection in hockey fight dataset [22].

  The initial convolutional layer in the architecture functions with a two-step stride employing a 3 × 3 × 3 × 32 filter. It takes an input with dimensions of 224 × 224 × 3. The next step is a depth-wise convolutional layer that works with a single-step stride, a 3 × 3 × 32 filter, and an input of 112 × 112 × 32 dimensions. While MobileNet is generally used for classification tasks, in this context, its SSD extension is vital for pinpointing the location of objects in the video frames. This SSD extension is integrated at the end of the MobileNet architecture and executes feed-forward convolutions to yield a predetermined set of bounding boxes. These boxes are examined to confirm the presence or absence of human figures, based on the extracted feature maps and applied convolutional

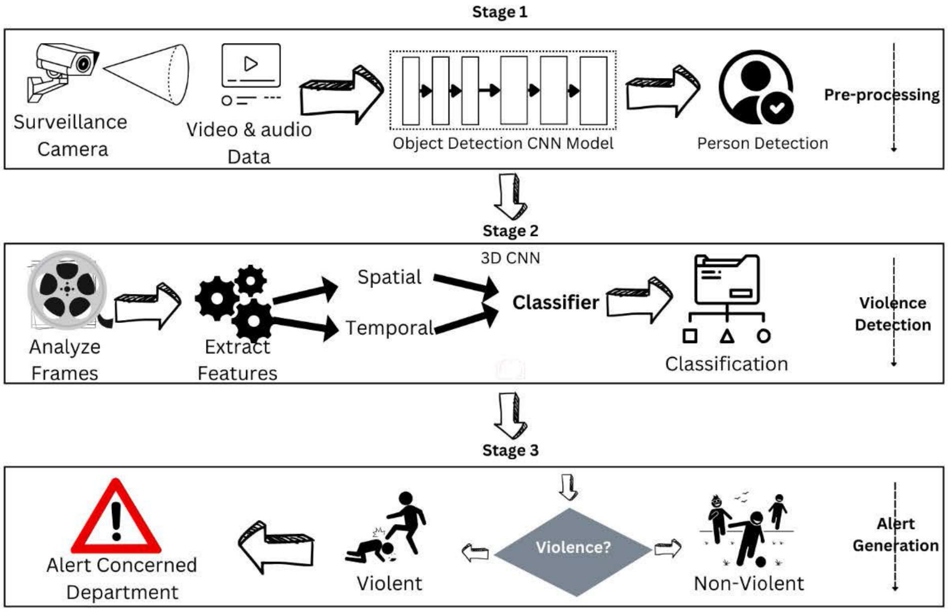


FIGURE 2. The three-stage process for violence detection.

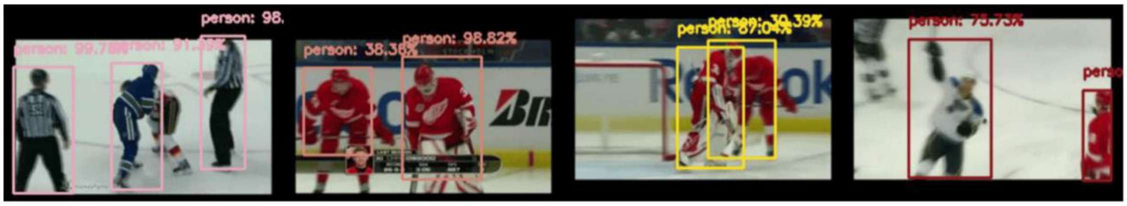


FIGURE 3. Detection in hockey fight dataset using MobileNet-SSD.

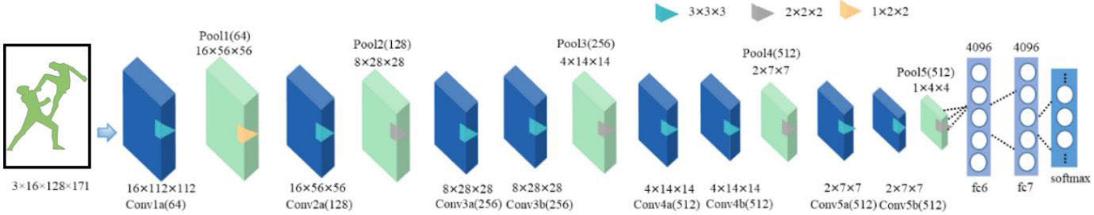


FIGURE 4. Detailed architecture of the C3D network used in the model.

filters. Each bounding box includes a set of class predictions with corresponding probabilities, and the class with the maximum probability is chosen. A zero probability indicates a lack of any object of interest.

- Learning with 3D-CNNs: The core of our violence detection model lies in its ability to extract spatiotemporal features, which is accomplished through a 3D-CNN. This network is specifically designed to handle sequences of 50 frames containing the detected person from the previous MobileNet-SSD model. Unlike 2D-CNNs, which only capture spatial information, 3D-CNNs are adept at preserving both spatial and temporal information due to their 3D convolution and pooling operations.

For our model, we have fine-tuned a 3D-CNN architecture inspired by the C3D model [9], initially developed using a version of Caffe. This architecture is particularly effective for video-based tasks and has been validated in multiple studies. The C3D model is composed of eight convolutional layers, five max-pooling layers, and two fully connected layers, culminating in a SoftMax output layer. Each convolutional layer uses 3 × 3 × 3 kernels with a stride of one. The max-pooling layers predominantly employ a 2 ×

2 × 2 kernel se ize, except for the first layer, which uses a 1 × 2 × 2 kernel with a stride of two to preserve temporal information. The convolutional layers are structured with a varying number of filters: 64 in the first layer, 128 in the second, and 256 in the third. These layers also feature kernels with a defined temporal depth, denoted by size D. The convolutional operations are performed with a kernel size of 3 and padding of 1. The fully connected layers, labeled as fc6 and fc7, contain 4096 neurons each. The SoftMax layer's output is tailored to the dataset's classes, which, in this case, are limited to two: violent and non-violent. To address the issue of overfitting and to enhance the model's learning capabilities, we employ random crops of size 3 × 16 × 128 × 128 from the original 50-frame input sequence during training. This architectural design allows the network to act as a hierarchical feature extractor. Lower layers focus on basic patterns like corners and edges, while higher layers capture more complex, global features. An illustrative representation of the C3D architecture is provided in Figure 4 below.

- Activity Classification using SoftMax Classifier: The features extracted by the 3D-CNN serve as the input for a SoftMax classifier. The SoftMax function is usually used in the last layer of a neural network-based classifier to make sure that the output probabilities are normalized and add up to one. This makes it possible to effectively label each frame as either violent or non-violent. The result classification directly influences the subsequent alert mechanism.
- Alert Generation: If the model predicts violence, an alarm is triggered, notifying the nearest security department. This immediate alert system allows for prompt action to be taken in response to the detected violent event, potentially averting dangerous situations and ensuring safety.

B. DATASETS

We have used the most widely used benchmark datasets: the hockey fight dataset, survellience fight dataset, and RWF-2000 dataset. We also used industrial surveillance dataset collected by Ullah, F.U.M [9]. These datasets are well balanced, labeled and they had 80%/20% split for training and testing purposes. Besides, these datasets cover mainly indoor scenes, outdoor scenes and a few weather conditions.

1) EXISTING BENCHMARK DATASETS

The surveillance fight dataset [15] includes indoor, outdoor, night, and daytime films from real-world surveillance and YouTube. This 300-video dataset has equal aggressive and nonviolent acts. RWF-2000 Dataset [16] includes factory, workplace, and other indoor, outdoor, day, and night videos. This dataset only contains surveillance videos without multimedia editing. This 2000-video collection has equal violent and nonviolent acts. The hockey fight dataset [17] is gathered on the National Hockey League (NHL) hockey grounds. This dataset contains 1000 NHL hockey game films with equal amounts of violent and nonviolent activities, with two players often in close bodily contact.

2) THE INDUSTRIAL SURVEILLANCE DATASET

Ullah, F.U.M [9] collected the industrial surveillance dataset from different sources and search engines such as YouTube and Google by inserting diverse queries, such as violence scenes in industrial surveillance, in factories, and in steel mills. The obtained videos from different sources have distinct video resolution and frame rate. The length of retrieved videos ranges from 7 to 12 min that are trimmed to 5-seconds violent and nonviolent clips for each class. They arranged the dataset in the same standard format, such as surveillance fight dataset. The industrial surveillance dataset consists of varied scenes such as industries, stores, offices, and petrol pumps. Compared to existing datasets, the industrial surveillance dataset is more challenging because most of the actions are aside of the center point from the camera and the frame per seconds (fps) varies like other surveillance datasets. Several samples of each dataset are given in Figure 5.

C. MODEL DEVELOPMENT AND TRAINING

The model is developed using TensorFlow as the primary framework for implementing the deep learning algorithms. The model is trained on an NVIDIA GeForce RTX 3080 laptop GPU, providing the computational power needed for efficient training and model optimization. The Adam optimizer is used to minimize the binary cross-entropy loss function. Early stopping is applied based on the validation loss to prevent over-fitting. For the hockey fight and RWF-2000 datasets, the learning rate and batch size are set at 0.001 and 32, respectively. For the Surveillance Fight and the Industrial Surveillance datasets, the learning rate and batch size are set at 0.0001 and 16, respectively. The model is trained for 50 epochs. After each epoch, the model is evaluated on a testing set to monitor its performance. The evaluation metrics include binary accuracy and binary cross-entropy loss.

D. MODEL EVALUATION

The model's performance is evaluated using a comprehensive set of metrics that include accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC). The metrics chosen for this evaluation include True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), Accuracy, Area Under the Receiver Operating Characteristic Curve (AUC ROC), Precision, Recall, and F1-score. These metrics provide a holistic view of the model's effectiveness in classifying violent and non-violent activities. A simple train-test split is used for validation. This approach allows for a straightforward yet effective way to assess the model's performance on unseen data. The model's performance is compared against existing models and techniques in the field

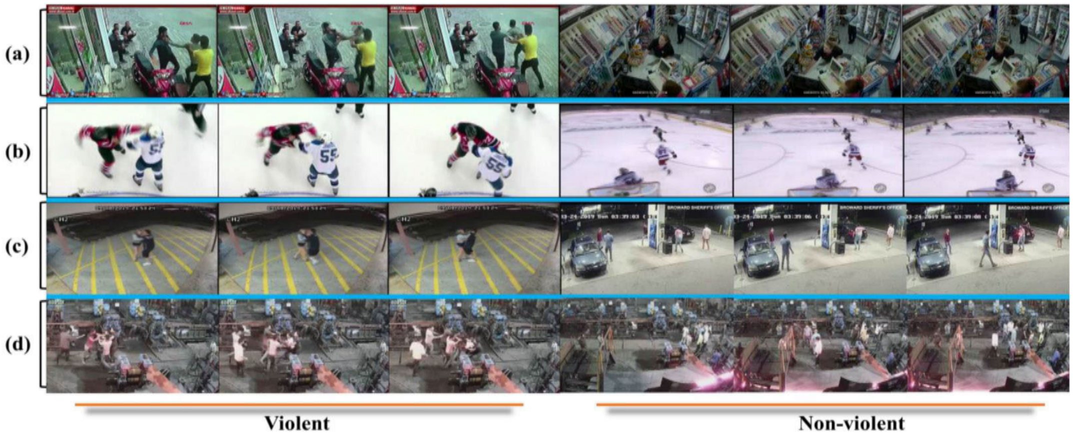


FIGURE 5. Samples from both classes of each violence detection dataset, with violent frames in first three columns and non-violent frames in following three columns. (a) Sample frames from surveillance fight dataset (b) Hockey fight dataset: violent and nonviolent frames. (c) Sample frames from RWF-2000 collection of real-world surveillance footage. (d) Sample frames from both industrial surveillance dataset groups.

of violence detection. This benchmarking helps to position the proposed model within the broader landscape of violence detection solutions. The model is tested across different datasets to assess its generalizability. This is crucial for ensuring that the model performs well not just on the data it was trained on but also on new, unseen data. The time taken for the model to make a prediction is measured. This is particularly important for real-time applications where quick decision-making is essential.

V. RESULTS AND ANALYSIS

A. MODEL PERFORMANCE COMPARISON

Table 2 presents a comparative evaluation of two state-of-the-art deep learning models employed for violence detection: ConvLSTM and the Proposed 3D-CNN. Figure 6 shows the confusion matrices of the two models for the four datasets.

The proposed 3D-CNN method consistently demonstrates superior or equivalent performance compared to the ConvLSTM method across all the datasets. For instance, in the RWF-2000 dataset, the proposed 3D-CNN model achieved an accuracy of 92.5%, significantly improving upon the 85.3% accuracy recorded for the ConvLSTM method. In this dataset, the 3D-CNN model correctly classified 185 violent incidents (TP) and 185 non-violent incidents (TN). However, it also produced 16 false positives (misclassifying non-violent acts as violent) and 14 false negatives (failing to detect actual violent incidents), highlighting the challenges posed by the diverse and complex scenes in this dataset.

In the hockey fight dataset, the proposed 3D-CNN model exhibited exceptional performance, achieving an accuracy of 97.2%, surpassing the 94% accuracy of the ConvLSTM method. With 98 TP and 96 TN, the model accurately identified most violent and non-violent events. The minor discrepancies, seen in 3 FP and 3 FN, can be attributed to the dynamic and aggressive nature of hockey, where fast-paced non-violent actions are often mistaken for violent ones.

For the surveillance fight dataset, the proposed 3D-CNN model demonstrated a notable improvement in performance. It correctly produced 28 TP and 26 TN, with only 3 FP and 3 FN, leading to an accuracy of 89.7%, compared to the 62% accuracy of the ConvLSTM method. The false positives are likely caused by the model misclassifying non-violent actions due to occlusions and congested camera views.

In the Industrial Surveillance dataset, both models show comparable performance, but the proposed 3D-CNN model still outperforms the ConvLSTM. The proposed model achieved an accuracy of 88.89%, compared to 73% for ConvLSTM. The proposed 3D-CNN model produced 26 TP and 27 TN, with 4 FP and 3 FP. The FP in this dataset likely stem from the model misinterpreting normal industrial activities as violent due to the complexity of the environment, while the false negatives suggest some violent events may have been obscured or subtle.

Overall, the proposed 3D-CNN method emerges as a highly promising technique for violence detection. Its robust performance, particularly on the RWF-2000, Industrial Surveillance, and Hockey Fight datasets, further solidifies its standing as a reliable and effective method for violence detection across various settings. However, while the model outperforms ConvLSTM in most metrics, addressing the remaining false positives and false negatives will be crucial for future improvements, especially in complex or fast-paced environments.

As demonstrated in Table 3, our proposed 3D-CNN method exhibits superior performance on across the four datasets, achieving an accuracy of 97.2% with a standard deviation of 1.55. This surpasses most state-of-the-art methods on the hockey fight dataset. While methods like ViF [20], OViF [21], DiMOLIF [22], and HOMO [23] consider both the orientation and magnitude changes of the optical flow, they still fall short of the performance achieved by the proposed 3D-CNN model. Despite the complexity

TABLE 2. The detailed evaluation results of the convlstm and proposed 3D-CNN based on accuracy, AUC, precision, recall, and F1-score.

Method	Dataset	TP	TN	FP	FN	Accuracy (%)	AUC	Precision	Recall	F1-score
ConvLSTM	Surveillance Fight	18	18	13	9	62.0	0.584	0.581	0.667	0.621
	RWF-2000	158	183	42	17	85.3	0.867	0.813	0.915	0.861
	Hockey Fight	92	96	6	6	94.0	0.982	0.941	0.941	0.941
	Industrial Surveillance	23	21	8	5	73.0	0.824	0.724	0.724	0.724
Proposed 3D-CNN	Surveillance Fight	28	26	3	3	89.7	0.822	0.724	0.778	0.750
	RWF-2000	185	185	16	14	92.5	0.911	0.849	0.930	0.888
	Hockey Fight	98	96	3	3	97.2	0.986	0.951	0.951	0.951
	Industrial Surveillance	26	27	4	3	88.89	0.846	0.833	0.833	0.833

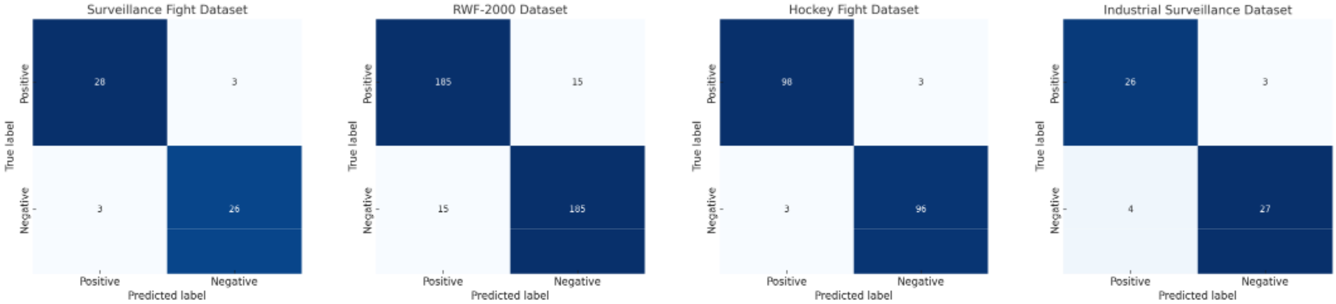


FIGURE 6. Visual representation of the confusion matrix for the proposed 3D-CNN. (a) Surveillance fight dataset. (b) RWF-2000 dataset. (c) Hockey fight dataset. (d) Industrial surveillance dataset.

TABLE 3. Accuracy and Standard Deviation (SD) of various methods evaluated on the hockey fight dataset.

Method	Accuracy (%)	SD
ViF [20]	81.6	0.22
OViF [21]	84.2	3.33
DiMOLIF [22]	88.6	1.20
HOMO [23]	89.3	0.91
Conv-LSTM [24]	97.1	0.55
Proposed 3D-CNN	97.2	1.55

of their approach, our architecture still outperforms theirs. This highlights the efficiency of our simpler architecture in capturing the most relevant frames and regions for violence detection. In comparison with Conv-LSTM [24], our method still achieves higher accuracy. These methods consider all frames as input, leading to the inclusion of redundant information that can adversely affect the network's decision-making process.

B. CROSS-DATA EXPERIMENTS

We also conducted cross-data experiments [40] to evaluate the model's generalizability across different datasets. The primary objective was to determine whether the model's

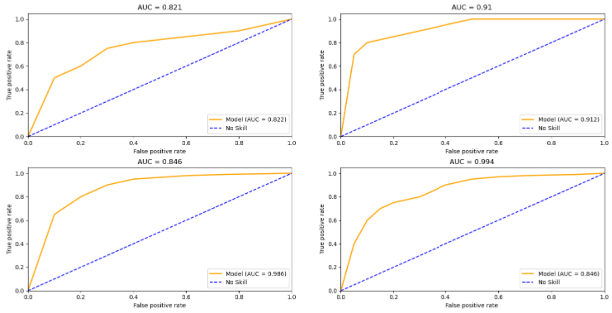


FIGURE 7. ROC and AUC of 3D-CNN: (a) Surveillance fight dataset, (b) RWF-2000 Dataset,(c)Hockey fight dataset, (d) Industrial surveillance dataset.

understanding of violence is sufficiently generic to accurately detect violent activities across diverse datasets. This is crucial for real-world applications where the model may be deployed in various environments with different types of data. The experiments involved training the model on one dataset and testing it on another, providing valuable insights into the model's adaptability and robustness. For example, when trained on the surveillance fight dataset and tested on the RWF-2000 dataset, the model achieved an accuracy of 63.67%, and when tested on the industrial surveillance dataset, the accuracy reached 77.46%. Similarly,

TABLE 4. Cross data experimentation results: generalization performance.

Training Dataset	Testing Dataset	Accuracy (%)
RWF-2000	Surveillance Fight Dataset	64.12
RWF-2000	Industrial Surveillance Dataset	59.85
RWF-2000	Hockey Fight Dataset	67.83
Industrial Surveillance Dataset	Surveillance Fight Dataset	73.78
Industrial Surveillance Dataset	RWF-2000	68.43
Industrial Surveillance Dataset	Hockey Fight Dataset	70.12
Surveillance Fight Dataset	RWF-2000	63.67
Surveillance Fight Dataset	Industrial Surveillance Dataset	77.46
Surveillance Fight Dataset	Hockey Fight Dataset	69.27
Hockey Fight	Surveillance Fight Dataset	70.19

the model achieved its lowest accuracy (59.85%) when trained on RWF-2000 and tested on Industrial surveillance, highlighting the challenges of generalizing across datasets with different characteristics. These results, summarized in Table 4, emphasize that while the model can generalize, its performance depends heavily on the diversity and complexity of the datasets used during training. The cross-dataset experimentation not only validates the model's effectiveness but also underscores its potential for scalable deployment in diverse surveillance settings, where variations in scene types, camera angles, and environmental conditions pose challenges.

The overall reduction in accuracy across cross-dataset experiments underscores the inherent challenges in generalizing between datasets with different characteristics. These challenges arise due to variations in factors such as video quality, scene diversity, camera angles, and the nature of violent actions across datasets. Video quality differences, including resolution and noise levels, affect the model's ability to adapt to new environments. Scene diversity and varying camera angles, such as those found in fixed industrial surveillance versus dynamic sports footage, further complicate generalization. Additionally, the specific contexts in which violent actions occur can differ greatly across datasets, making it difficult for models to identify consistent patterns.

C. TIME COMPLEXITY ANALYSIS

Processing time is a critical consideration for video data in Industrial Internet of Things (IIoT)-aided surveillance systems. In this study, the ConvLSTM model and the proposed 3D-CNN model were evaluated for their efficiency in handling video frames. The ConvLSTM model processed 28 frames per second, while the proposed model handled 72 frames per second. This means the proposed model processes one frame in approximately 0.01389 seconds, making it 2.57 times faster than the ConvLSTM model. This improved processing speed underscores the efficiency of the proposed approach. In terms of model complexity, the ConvLSTM model consists of 18,976,770 parameters, while the proposed model significantly reduces this to 4,470,298 parameters. This substantial reduction in parameters highlights the optimization and efficiency achieved by the proposed model, both in terms of processing speed and model size.

The models were trained and tested on a high-performance system featuring a 64-bit operating system and an x64-based Intel(R) Core(TM) i7-10870H CPU, clocked at 2.20GHz with a turbo boost up to 2.21GHz. The system was also equipped with 64.0 GB of RAM, which provided ample memory to handle the intensive computational tasks involved in processing video data for surveillance purposes.This hardware setup facilitated not only fast training and testing but also ensured that the models could efficiently handle large batches of video frames, which is crucial for real-time surveillance applications.The proposed model is designed to be efficient enough to run on edge devices, such as IoT-based surveillance systems, which often have limited computational resources compared to high-end GPUs. This is achieved by using a lightweight initial stage (CNN for human detection) to reduce the number of frames that need to be processed by the more computationally intensive 3D-CNN. The ability to run models like the proposed 3D-CNN in such an environment is a strong indicator that these models are scalable and could be deployed in real-world IIoT-aided surveillance systems, where quick response times and the ability to process high-resolution video data in real-time are paramount. This combination of hardware capability and optimized model design further emphasizes the practicality and applicability of the proposed model in demanding IIoT scenarios.

VI. CONCLUSION AND FUTURE WORK

In this paper, a three-staged end-to-end framework is proposed for violence detection in a surveillance video stream. In the first stage, human are detected using an efficient CNN model to remove unwanted frames, which results in reducing the overall processing time. Next, frame sequences with persons are fed into a 3D-CNN model trained on three benchmark datasets, where the spatiotemporal features are extracted and forwarded to the SoftMax classifier for final predictions. Experimental results over various benchmark datasets confirm that our method is a good fit for violence detection in surveillance and achieves better accuracy than several other existing techniques. We want to use our methodology on devices with limited resources. Our paper involves implementing edge intelligence in order to identify instances of violence in the IoT devices. Our future research aims to improve violence detection by leveraging multiview data for thorough analysis.Additionally, We plan to enhance the model's adaptability to varying environmental conditions by incorporating sound sensor data. This approach is intended to be particularly useful in challenging light conditions, where visual data alone may be insufficient. By integrating auditory inputs, we aim to improve the robustness and overall performance of the model in diverse real-world scenarios. This is necessary since existing algorithms depend on single cameras, which are unable to capture the complete picture. Hence, multiview enables thorough surveillance of activity from any perspective.

VII. ACKNOWLEDGMENT

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSF.

REFERENCES

[1] D. M. Gavrila, "The visual analysis of human movement: A survey," *Comput. Vis. Image Understand.*, vol. 73, no. 1, pp. 82–98, Jan. 1999.

[2] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," *Comput. Vis. Image Understand.*, vol. 73, no. 3, pp. 428–440, Mar. 1999.

[3] Y. Wang et al., "Violence detection in surveillance environments using LSTM," *IEEE Trans. Emerg. Topics Comput.*, vol. 9, no. 1, pp. 1–13, Jan. 2021.

[4] G. Muhammad, M. A. Hossain, and G. Bebis, "Attention-based LSTM network for human action recognition," *Pattern Recognit. Lett.*, vol. 138, pp. 120–127, May 2021.

[5] L. Zhang, X. Chang, J. Liu, M. Luo, M. Prakash, and A. G. Hauptmann, "Few-shot learning for human action recognition using cross-modal memory networks," *IEEE Trans. Image Process.*, vol. 30, pp. 2301–2315, 2021.

[6] U. Haroon, A. Ullah, T. Hussain, W. Ullah, M. Sajjad, and K. Muhammad, "Multi-stream deep learning model for human interaction recognition," *J. Vis. Commun. Image Represent.*, vol. 73, Jun. 2021, Art. no. 102981.

[7] F. U. M. Ullah, K. Muhammad, I. Haq, N. Khan, A. A. Heidari, and S. A. Baik, "Edge computing-based AI-assisted violence detection in IoT-based surveillance networks," *IEEE Internet Things J.*, vol. 8, no. 9, pp. 7601–7611, Sep. 2021.

[8] J. Chen, X. Wang, and Y. Zhang, "Skeleton-based human action recognition using graph convolutional networks," *IEEE Access*, vol. 9, pp. 69989–70001, 2021.

[9] S. Kim and J. Lee, "Multi-modal anomaly detection in video streams using audio-visual fusion," *Pattern Recognit. Lett.*, vol. 145, pp. 12–21, Jun. 2021.

[10] J. Chen, C. Ma, and J. Wang, "Graph convolutional networks for skeleton-based action recognition," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 2739–2753, 2019.

[11] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.

[12] M. Ding et al., "3D-CNN-based action recognition," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1252–1265, Mar. 2017.

[13] J. Y. Lee, K. D. Kim, and K. Kim, "A study on improving the location of CCTV cameras for crime prevention through an analysis of population movement patterns using mobile big data," *KSCE J. Civil Eng.*, vol. 23, no. 1, pp. 376–387, Jan. 2019.

[14] S. Khan, K. Muhammad, S. Mumtaz, S. W. Baik, and V. H. C. de Albuquerque, "Energy-efficient deep CNN for smoke detection in foggy IoT environment," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9237–9245, Dec. 2019.

[15] M. Sajjad, S. Khan, K. Muhammad, W. Wu, A. Ullah, and S. W. Baik, "Multi-grade brain tumor classification using deep CNN with extensive data augmentation," *J. Comput. Sci.*, vol. 30, pp. 174–182, Jan. 2019.

[16] M. Sajjad, S. Khan, T. Hussain, K. Muhammad, A. K. Sangaiah, A. Castiglione, C. Esposito, and S. W. Baik, "CNN-based anti-spoofing two-tier multi-factor authentication system," *Pattern Recognit. Lett.*, vol. 126, pp. 123–131, Sep. 2019.

[17] S. Wang, X. Liu, S. Liu, K. Muhammad, A. A. Heidari, J. D. Ser, and V. H. C. de Albuquerque, "Human short long-term cognitive memory mechanism for visual monitoring in IoT-assisted smart cities," *IEEE Internet Things J.*, vol. 9, no. 10, pp. 7128–7139, May 2022.

[18] K. Muhammad, Mustaqeem, A. Ullah, A. S. Imran, M. Sajjad, M. S. Kiran, G. Sannino, and V. H. C. de Albuquerque, "Human action recognition using attention based LSTM network with dilated CNN features," *Future Gener. Comput. Syst.*, vol. 125, pp. 820–830, Dec. 2021.

[19] L. Zhang, X. Chang, J. Liu, M. Luo, M. Prakash, and A. G. Hauptmann, "Few-shot activity recognition with cross-modal memory network," *Pattern Recognit.*, vol. 108, Dec. 2020, Art. no. 107348.

[20] U. Haroon, A. Ullah, T. Hussain, W. Ullah, M. Sajjad, K. Muhammad, M. Y. Lee, and S. W. Baik, "A multi-stream sequence learning framework for human interaction recognition," *IEEE Trans. Hum.-Mach. Syst.*, vol. 52, no. 3, pp. 435–444, Jun. 2022.

[21] K. Lloyd, P. L. Rosin, D. Marshall, and S. C. Moore, "Detecting violent and abnormal crowd activity using temporal analysis of grey level co-occurrence matrix (GLCM)-based texture measures," *Mach. Vis. Appl.*, vol. 28, nos. 3–4, pp. 361–371, May 2017.

[22] F. U. M. Ullah, K. Muhammad, I. U. Haq, N. Khan, A. A. Heidari, S. W. Baik, and V. H. C. de Albuquerque, "AI-assisted edge vision for violence detection in IoT-based industrial surveillance networks," *IEEE Trans. Ind. Informat.*, vol. 18, no. 8, pp. 5359–5370, Aug. 2022.

[23] J. Chen, Y. Wang, J. Wang, X. Gao, and L. Nie, "Spatiotemporal graph convolutional networks for skeleton-based human action recognition in surveillance environments," *IEEE Trans. Multimedia*, vol. 24, pp. 1235–1248, 2022.

[24] S. Kim, H. Lee, J. Park, and J. Choi, "Anomaly detection in video streams using multi-modal fusion and deep neural networks," *Pattern Recognit. Lett.*, vol. 145, pp. 39–47, May 2023.

[25] G. Batchuluun, J. H. Kim, H. G. Hong, J. K. Kang, and K. R. Park, "Fuzzy system based human behavior recognition by combining behavior prediction and recognition," *Expert Syst. Appl.*, vol. 81, pp. 108–133, Sep. 2017.

[26] C. Ding, S. Fan, M. Zhu, W. Feng, and B. Jia, "Violence detection in video by using 3D convolutional neural networks," presented at the Int. Symp. Vis. Comput., Jan. 2014.

[27] S. Akti, G. A. Tataroglu, and H. K. Ekenel, "Vision-based fight detection from surveillance cameras," in *Proc. 9th Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, Nov. 2019, pp. 1–6.

[28] M. Cheng, K. Cai, and M. Li, "RWF-2000: An open large scale video database for violence detection," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 4183–4190.

[29] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Proc. Int. Conf. Comput. Anal. Images Patterns*. Berlin, Germany: Springer, 2011, pp. 332–339.

[30] H. Ullah, K. Muhammad, A. Ullah, T. Saba, and A. Rehman, "Violence detection in Hollywood movies by the fusion of visual and mid-level audio cues," *IEEE Access*, vol. 6, pp. 48250–48261, 2018.

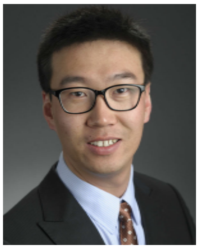
[31] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 1–6.

[32] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "Violence detection using oriented VIolent flows," *Image Vis. Comput.*, vols. 48–49, pp. 37–41, Apr. 2016.

[33] A. Ben Mabrouk and E. Zagrouba, ‘‘Spatio-temporal feature using optical flow based distribution for violence detection,’’ *Pattern Recognit. Lett.*, vol. 92, pp. 62–67, Jun. 2017.
[34] J. Mahmoodi and A. Salajeghe, ‘‘A classification method based on optical flow for violence detection,’’ *Expert Syst. Appl.*, vol. 127, pp. 121–127, Aug. 2019.
[35] S. Sudhakaran and O. Lanz, ‘‘Learning to detect violent videos using convolutional long short-term memory,’’ in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2017, pp. 1–6.
[36] P. Zhou, Q. Ding, H. Luo, and X. Hou, ‘‘Violence detection in surveillance video using low-level features,’’ *PLoS ONE*, vol. 13, no. 10, Oct. 2018, Art. no. e0203668.
[37] E. Fenil, G. Manogaran, G. Vivekananda, T. Thanjaivadivel, S. Jeeva, and A. Ahilan, ‘‘Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM,’’ *Comput. Netw.*, vol. 151, pp. 191–200, Mar. 2019.
[38] I. Serrano Gracia, O. Deniz Suarez, G. Bueno Garcia, and T.-K. Kim, ‘‘Fast fight detection,’’ *PLoS ONE*, vol. 10, no. 4, Apr. 2015, Art. no. e0120448.
[39] Y. Shi, Y. Tian, Y. Wang, and T. Huang, ‘‘Sequential deep trajectory descriptor for action recognition with three-stream CNN,’’ *IEEE Trans. Multimedia*, vol. 19, no. 7, pp. 1510–1520, Jul. 2017.
[40] H. Khan, ‘‘Violence detection from industrial surveillance videos using deep learning,’’ *ProQuest One Academic*, 2023.
[41] V. Sharma, M. Gupta, A. K. Pandey, D. Mishra, and A. Kumar, ‘‘A review of deep learning-based human activity recognition on benchmark video datasets,’’ *Appl. Artif. Intell.*, vol. 36, no. 1, pp. 1–22, Dec. 2022.

HAMZA KHAN received the bachelor’s degree in software engineering from the University of Haripur, Pakistan, and the master’s degree in computer science from North Carolina Agricultural and Technical State University, where he is currently pursuing the Ph.D. degree in computer science.

XIAOHONG YUAN is currently working as a Professor with the Department of Computer Science, North Carolina Agricultural and Technical State University. Her research has been funded by the National Security Agency, the National Centers of Academic Excellence in Cybersecurity (NCAE-C), the National Science Foundation, the Department of Energy, and the Department of Education. Her research interests include AI and machine learning, anomaly detection, software security, cyber identity, and cyber security education. She has served on the editorial board for several journals on cybersecurity.



LETU QINGGE received the Ph.D. degree in computer science from Montana State University, Bozeman, MT, USA. He is currently working as an Assistant Professor with the Department of Computer Science, North Carolina Agricultural and Technical State University, USA. His research has been funded by NSF and NIH. His research interests include algorithms, deep learning, bioinformatics, and computer vision.

KAUSHIK ROY (Senior Member, IEEE) is currently working as a Professor and the Chair of the Department of Computer Science, North Carolina Agricultural and Technical State University (NC A&T), where he is also the Jefferson-Pilot/Ron McNair Endowed Chair. He is also the Director of Center for Cyber Defense (CCD) and Center for Trustworthy AI. He also leads the Cyber Defense and AI Laboratory. He has more than 180 publications, including 50 journal articles and a book. His research has been funded by the National Science Foundation (NSF), the Department of Defense (DoD), and the Department of Energy (DoE). His current research interests include cybersecurity, cyber identity, biometrics, machine learning (with a focus on deep learning), data science, cyber-physical systems, and big data analytics.