(IJIEST) 2026, Vol. No. 12, Issue 1, Jan-Dec **e-ISSN: 2454-9584; p-ISSN: 2454-8111**

# A Review: Campus Violence Detection Using Deep Learning Models

**Liqaa M. Shoohi**
Baghdad University, College of Physical Education and Sports Sciences for Women, Baghdad, Iraq.

*DOI:10.37648/ijiest.v12i01.007*

[1]*Received: 06 January 2026; Accepted: 08 February 2026; Published: 02 March 2026*

**Abstract**

This paper offers a systemic review of the deep learning methods to detect violence on campus, which is a critical issue in intelligent surveillance to improve the student safety and prompt cut off of violent accidents. The review reviews studies published 2018-2025, concentrating on model structure to detect fights, bullying, vandalism, and aggressive behavior on problematic campuses due to occlusion and light variations and complicated human interactions. The research design includes a comparative study of different deep learning networks, such as CNNs, RNNs, 3D CNNs, attention-based networks, transformers, graph neural networks, neuro-fuzzy, and multimodal systems and federated learning methods. The paper also assesses benchmark datasets frequently utilized, performance measures, and even real-time deployment considerations. Findings show that CNN models of light weight can fit well into real-time use but are not capable of time modeling but hybrid CNN-RNN and attention based models may provide better accuracy at increased computing cost. Transformer and multimodal models have shown promising performance, but are computationally expensive to e.g. deploy to edges. The review presents important research gaps, such as inadequate datasets to the specific campus, insufficient multimodal integration, privacy issues, and the necessity of explainable and lightweight implementation. This work can guide further research on viable solutions, effective, and privacy-conscious violence detection systems in a learning setting.

**Keywords:** *Violence detection; Campus surveillance; deep learning; CNN; Transformer; Video analysis; Multimodal learning; federated learning; Computer vision.*

## 1. Introduction

Over the past few years, automated violence detection in surveillance videos has become much better as a result of the quick evolution of deep learning techniques [1]. Such systems could especially be essential in campus settings, where, by identifying act of violence like fights, bullying, and aggressive behavior early, the systems could be used to improve the security of students and minimize the time required to respond. In spite of the advancement in overall violence detection, campus specific violence detection is problematic because of complex human interaction, crowd situations, occlusion and because violent and non-violent events are similarity. Current research does not cover campus settings in depth, but the general surveillance data, hence, a research gap exists in context-specific violence detection systems in educational settings. Hence, the given review seeks to offer a systematic review of deep learning methods

[1] How to cite the article: Shoohi L.M..; March 2026; A Review: Campus Violence Detection Using Deep Learning Models; *International Journal of Inventions in Engineering and Science Technology;* Vol 12 Issue 1, 47-63, DOI: http://doi.org/10.37648/ijiest.v12i01.007

in campus violence detection such as convolutional neural networks, recurrent neural networks, transformers, hybrid networks, and neuro-fuzzy systems [2]. Moreover, the review examines popular datasets, metrics of evaluation, obstacles, and further studies in this area. [3]. this review has made the following key contributions:

- An overview of deep learning models employed in the surveillance video detection of violence.
- Comparison of various model structures and their advantage.
- An overview of popular datasets and metrics of evaluation.
- Detection of major challenges with campus violence detection systems.
- Discussion Future research directions (i.e., multimodal learning, federated learning, and transformers based architectures).

## 2. Related Work

The recent few years have witnessed an enormous progress in automated violence-detection like the one undertaken by the surveillance in campuses. These issues of complex human relations, crowd interactions and occlusions coupled with lighting patterns have been proposed to be tackled using the various deep learning models including convolutional neural networks (CNNs), recurrent neural networks (RNNs), transformer-based and hybrid models [4]. The section will summarize and briefly review the most relevant articles published within the period of 202025 in terms of their methodology, data sets and key findings. [5]. Table 1 shows a thorough review of past literature that demonstrates the development of models architecture and the way they can perform in various environments.

**Table 1.** Previous Studies on Campus/Surveillance Violence Detection (2020–2025)

| Author | Year | Model / Approach | Environment | Dataset | Key Findings / Summary |
|---|---|---|---|---|---|
| M. Cheng, K. Cai, and M. Li[6] | 2020 | Lightweight CNN | Surveillance | RWF-2000 | A 94% accuracy with low computational complexity; it can be implemented in real time. |
| A. Traoré and M. A. Akhloufi [7] | 2020 | CNN + GRU | Surveillance | RLVS | 95% accuracy; enhanced time sequence modeling of violence sequences. |
| S. Sharma, B. Sudharsan, S. Naraharisetti, V. Trehan, and K. Jayavel. [8] | 2021 | CNN + LSTM | Sports | Hockey Fight | Records motion patterns effectively, slower processing does not allow real time use. |
| R. Maqsood, U. I. Bajwa, G. Saleem, R. H. | 2021 | 3D CNN | Surveillance | UCF-Crime | Powerful spatiotemporal learning; expensive |

| Author | Year | Model / Approach | Environment | Dataset | Key Findings / Summary |
|---|---|---|---|---|---|
| Raza, and M. W. Anwar. [9] | | | | | computations; non-real-time. |
| F. J. Rendón-Segador, F. Enríquez y O. Deniz.[10] | 2021 | Attention CNN | Surveillance | RWF-2000 | Concentrated on significant areas; 96% correctness; responsive to loss. |
| A. Arnab, M. Dehghani, G. Heigold, C. Sun, and C. Schmid.[11] | 2021 | Video Transformer | Surveillance | RLVS | Global temporal modeling; very high computation, had an accuracy of 98%. |
| K. B. Koushik, K. Raihan, and M. M. Khan. [12] | 2022 | CNN + BiLSTM | Crowd | Violent Flow | Two-way time details enhance identification of complicated relationships. |
| R. Vijeikis. [13] | 2022 | Hybrid CNN + Attention | Surveillance | RWF-2000 | The accuracy and speed are balanced; tuning on hyper parameters is needed. |
| N. Talpur, S. J. Abdulkadir, H. Alhussian, M. H. Hasan, N. Aziz, and A. Bamhdi.[14] | 2023 | Neuro-Fuzzy + DL | Multi-environment | Multiple | Deals with ambiguity in the human behavior; intricate training is necessitated. |
| W. Ullah, T. Hussain, and S. W. Baik.[15] | 2023 | Vision Transformer | Surveillance | UCF-Crime | Good long-range time-wise modelling; computationally expensive. |
| J. Zhang, Z. Mohd Yunos, and H. Haron.[16] | 2023 | Graph Neural Network | Campus / Crowd | RLVS | Gives a good representation of human interaction; graph construction is difficult. |
| L. Zhang, W. Cui, B. Li, Z. H. Chen, M. | 2023 | Federated Learning + CNN | Multi-source | Multi-source | Privacy preserving strategy; a bit less |

| Author | Year | Model / Approach | Environment | Dataset | Key Findings / Summary |
|---|---|---|---|---|---|
| Wu, T. S. Gee.[17] | | | | | accurate; overhead communications. |
| M. B. Shaikh, D. Chai, S. M. S. Islam, and N. Akhtar.[18] | 2024 | Multimodal CNN + Audio | Surveillance | RWF-2000 | Better context recognition with audio; accuracy, 98% full multimodal data. |
| M. Qaraqe, Y. D. Yang, E. B. Varghese, E. Basaran and A. Elzein.[19] | 2024 | Lightweight Transformer | Surveillance | Custom | Real-time transformer; 97% accuracy; slightly lower than heavy transformers. |
| K. Alomar, H. I. Aysel, and X. Cai.[20] | 2025 | Hybrid CNN + Transformer | Multi-environment | Multiple | Highest accuracy (99%); high computational cost; not real-time. |

## 3. Research Methodology

The review is conducted according to the systematic literature review approach in order to be transparent, reproducible, and cover all the most recent developments in the field of campus violence detection based on deep learning techniques. The review procedure was performed based on the organized guidelines, which were inspired by PRISMA (Preferred Reporting Items in Systematic Reviews and Meta-Analyses) framework [21].

### 3.1 Data Sources and Search Strategy

The major scientific databases, such as Scopus, IEEE Xplore, Science Direct, Springer Link, and Google Scholar, were analysed to find the relevant studies. The search strategy was thoroughly developed based on the following combinations of keywords: violence detection, campus violence detection, video violence recognition, deep learning based on surveillance, CNN, RNN, transformer based video analysis, and multimodal violence detection. Only studies published in 2018-2025 were considered to reflect the latest developments in deep learning approaches [ 21].

### 3.2 Inclusion and Exclusion Criteria

Inclusion and exclusion criteria were used to make sure that the chosen studies were of quality and relevance. Articles or conference papers were considered to include in the studies provided they were peer-reviewed articles or conference papers on video-based violence detection based on deep learning or hybrid models presenting validated experimental results on benchmark datasets. Research papers were not included in the review where they were not in English, were image-only, could not be validated through experimentation, or were duplicates or irrelevant [22].

### 3.3 Study Selection Process

First, about 135 studies were found in the database searches in Scopus, IEEE Xplore, Science Direct, Springer Link, and Google Scholar. Following the elimination of duplicated records, there were 102 papers left. Subsequently, the selection was done on the basis of the titles and abstracts, and 82 papers were eventually picked as full-text papers. Lastly, it included 58 quality studies in the scrupulous analysis meticulously on its relevance, quality of methodology, and application of data and experimental validation [23]. Figure 2: The PRISMA Flow Diagram.

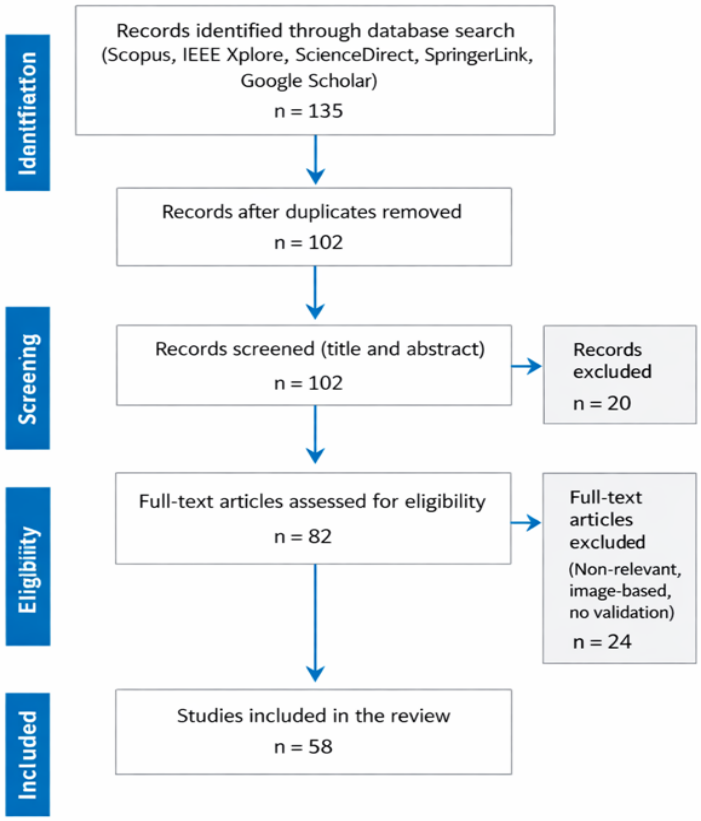


**Figure 1.** PRISMA flow diagram

### 3.4 Data Extraction and Analysis

As part of the chosen research, desirable data were retrieved, such as the model architecture, dataset involved, temporal modeling strategy, evaluation data and compute complexity, and applicability in real-time. The identified articles were then divided into various categories, depending on the type of model, e.g., CNN-based models, RNN-based models, transformer-based architectures, graph-based models, and hybrid. They were compared and contrasted against each other considering their strengths and weaknesses based on performance, efficiency, scalability and their compatibility with campus environments [21].

## 4. Deep Learning Approaches for Campus Violence Detection

Techniques of deep learning are now recognized as the new paradigm of video-based violence detection, since they can learn about complex spatial and temporal representations of the raw data automatically. In campus settings, it is a busy place with scenes and behavior that are subtle, therefore, a strong representation of both appearance and movement is required. In this section, the key types of deep learning models applicable to violence detection systems are reviewed [24]. Figure 2: Deep learning taxonomy of campus violence detection.

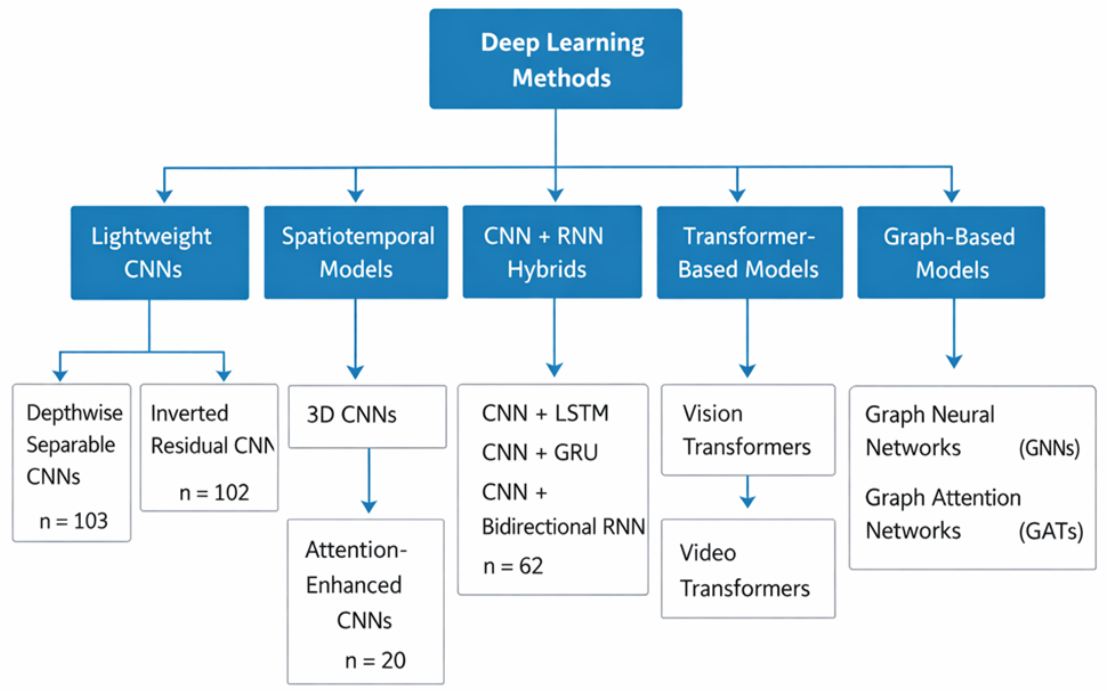


**Figure 2.** Taxonomy of Deep Learning Methods

### 4.1 Lightweight Convolutional Neural Networks (CNNs)

Lightweight CNNs have been extensively utilized in real-time violence detectors in resource-constrained deployment systems including campus monitors. Depth wise separable convolution-based and inverted residual block models like those found in Mobile Net variants can greatly lower the computational complexity required, but still achieve competitive accuracy. These models are especially ideal in edge deployment, and power consumption and latency are major constraints. Although lightweight CNNs are simple, they can be used to obtain a high accuracy on benchmark data sets, since they are effective in capturing spatial features related with violent interactions [25].

### 4.2 Spatiotemporal Models and 3D CNNs

In contrast to the traditional 2D CNN models, spatiotemporal models like 3D CNN models do not learn on a frame by frame basis but instead learn on video clips to learn both spatial and temporal features. The models do not employ any dynamic features in capturing motion patterns thus being very effective in identifying dynamic activities like fights and aggressive behavior. Nonetheless, 3D CNNs are computationally intensive and they need to be trained on a large scale to be applicable in real time campus setting [26].

### 4.3 Recurrent Neural Networks (RNNs) and Hybrid Models

There is a common use of recurrent neural networks together with CNNs, such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), which reproduce temporal dependencies within video frames. In these hybrid systems, CNNs are used to extract spatial features, and such features are given to RNNs to take care of sequence modeling. Bidirectional variants (BiLSTM and BiGRU) represent another improvement, as they take into account the past and future temporal context. These models have proven to be very effective in identifying multifaceted human interactions when it comes to surveillance conditions [20].

### 4.4 Attention-Enhanced and Explainable Models

Attention mechanisms have greatly enhanced the quality of deep learning models as they come to focus on the most informative parts within video frames. Attention modules are useful in the context of detecting violence by pointing out significant spatiotemporal areas that are related to aggression [27]. Also, explainable AI methods are also added to enhance the transparency of models. As a form of interpretation, key frame selection and attention visualization are principles that offer interpretable results, which is crucial in real-world applications where a sensitive environment is concerned, such as campuses [28].

### 4.5 Transformer-Based Architectures

Transformer models, self-attention-based, have now emerged to be the Billy sticks of video understanding [29]. Transformers can also discover the long-range relationship and global pools among video sequences, unlike CNN-based systems. Video Transformers (ViTs) and other Video Transformers have been demonstrated to be highly successful especially in difficult conditions involving many actors and interactions. However, they have high computation cost and hence large computation cost to real-time applications [30].

### 4.6 Graph-Based and Relational Models

Graph based the objects or individuals in the scene are represented in a graph (i.e. as nodes) and the relationships amongst them as edges [31]. GNNs and Graph Attention Networks (GATs) have been shown to be especially useful when it comes to modeling relational dynamics of crowded spaces. The theories are most applicable in the campus scenarios where many individual's interaction is vital in accurate violence detection [32].

### 4.7 Neuro-Fuzzy and Hybrid Intelligent Systems

The learning behavior of the neural network, and reasoning of the fuzzy logic are fitted to the neuro-fuzzy systems. The systems assist in grappling with uncertainty and ambiguity of human behavior that normally is a typical occurrence in the process of detecting violence [33]. Such hybrid models allow robustness and understandability of the decisions through a combination of deep feature representations and fuzzy inference mechanisms [34].

### 4.8 Multimodal and Federated Learning Approaches

The latest studies have investigated multimodal learning, in which visual information is integrated with other sources (audio cues and human pose detection) [35] [36]. This combination develops enhanced contextualization and increases detection accuracy. Federated learning has received attention as well as one of the privacy-saving solutions since it allows joint model training among distributed devices without exchanging raw data [37]. This is especially of great importance in academic setting where the issues of privacy are crucial.

## 5. Datasets and Benchmark Practices

Whether deep learning models are able to detect violence or not greatly relies on the availability of good datasets with annotations [6]. Benchmark datasets have a standard evaluation setup that can reasonably compare different algorithms [38]. In the discipline of violence detection, surveillance videos, movies, or sports footage are used as the baseline of data mainly because of the challenges in gathering actual campus violence data due to privacy and ethical reasons [39]. Current data sets differ in the environment and amount of videos, type of annotation, and types of violence [40]. Crowd violence datasets concentrate on the violence of crowds, but single fight or abnormal behavior detection datasets aren't as many. The choice of suitable dataset is thus crucial in entrusting performance of the model and ability to generalize [41]. The most popular datasets used in the research on violence detection are summarized in Table 2.

**Table 2.** Common Datasets Used for Violence Detection Research

| Dataset | Year | Environment | Classes | Number of Videos | Duration | Application | Notes |
|---|---|---|---|---|---|---|---|
| RLVS (Real-Life Violent Scenes) | 2011 | Real surveillance | Violence / Non-violence | 2000+ | Short clips | Violence detection | Real-world scenarios |
| RWF-2000 | 2020 | Surveillance | Fight / Non-fight | 2000 | Short clips | Violence detection | Widely used benchmark |
| Hockey Fight Dataset | 2012 | Sports | Fight / Non-fight | 1000 | Short clips | Violence detection | Controlled environment |
| Violent Flow | 2015 | Crowd scenes | Violent / Non-violent crowd | 246 | Video clips | Crowd violence | Crowd behavior analysis |
| UCF-Crime | 2018 | Surveillance | Crime 13 classes | 1900 | Long videos | Anomaly detection | Real surveillance videos |
| Surveillance Camera Fight Dataset | 2019 | Surveillance | Fight / Non-fight | Various | Video clips | Violence detection | Real camera footage |

### 5.1 Dataset Analysis

The prevalent datasets used in the study of violence detection are summarized in Table 2. RWF- 2000 and RLVS are the most common ones in all those datasets as they are the most realistic in terms of the surveillance scenario and balanced in-violent/ non- violent proportions. Experiments with the Hockey Fight dataset commonly have a controlled environment, whereas the Violent Flow dataset deals with crowd violence cases. UCF-Crime is a massive dataset that has been created to detect anomalies (i.e. not pure violence detection) yet is commonly utilized to test spatiotemporal

deep learning models. Although such datasets are available, the majority of them is not specifically created in the campus setting, and this puts a disconnect between research experiments and real-world implementation in a campus. Hence, creating campus-specific datasets is also a key avenue of research in the future.

### 5.2 Dataset Challenges

Campus violence detection has several issues in its dataset collection process such as privacy concerns [38], difficulty in data annotation, class imbalance, and accessibility of real-world violent incidents. These complications ensure that it is not easily possible to train deep learning models that generalize across campuses.

## 6. Comparative Analysis of Methods

In this section, we give a comparative analysis between deep learning in violence detection through surveillance and campus setting. Their comparison will center on model architecture, how the two approaches handle time features, computational complexity, real-time, and the strengths of each approach. The literature has proposed various deep learning models, such as CNN-based models, recurrent neural networks [42], hybrid CNN-RNN-based models [38], transformer-based models [39], neuro-fuzzy systems [24], and federated learning models [37]. All the approaches possess their strengths and weaknesses based on the situation of application and computational constraints. A comparison of the most common approaches used in research on violence detection summarizes Table 3. Figure 3 Comparison of models used to detect violence in campuses.

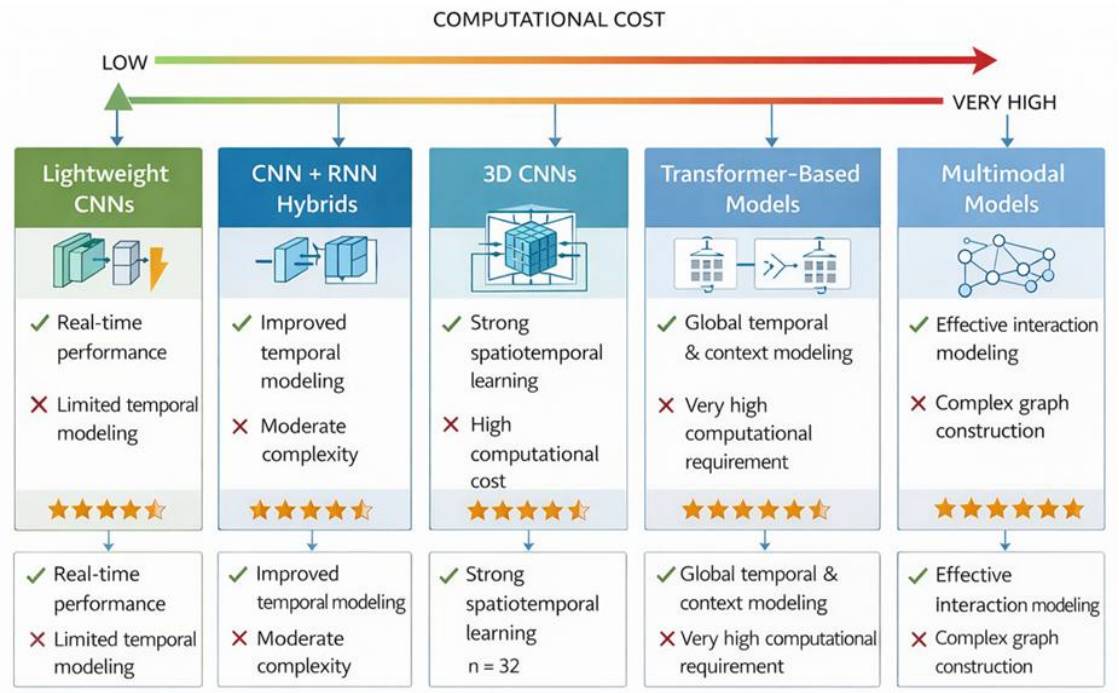


**Figure 3.** Comparison of models for campus violence detection

**Table 3.** Comparison of Deep Learning Methods for Violence Detection

| Author | Year | Model | Dataset | Accuracy | Real-Time | Key Strength | Limitation |
|---|---|---|---|---|---|---|---|
| M. Cheng, K. Cai, and M. Li. | 2020 | Lightweight CNN | RWF-2000 | 94% | Yes | Low computational cost | Limited temporal features |
| A. Traoré and M. A. Akhloufi. | 2020 | CNN + GRU | RLVS | 95% | Yes | Good temporal modeling | Moderate complexity |
| S. Sharma, B. Sudharsan, S. Naraharisetti, V. Trehan, and K. Jayavel. | 2020 | CNN + LSTM | Hockey Fight | 96% | No | Captures motion patterns | Slow processing |
| R. Maqsood, U. I. Bajwa, G. Saleem, R. H. Raza, and M. W. Anwar. | 2021 | 3D CNN | UCF-Crime | 97% | No | Strong spatiotemporal learning | High computational cost |
| F. J. Rendón-Segador, F. Enríquez y O. Deniz. | 2021 | Attention CNN | RWF-2000 | 96% | Yes | Focus on important regions | Sensitive to occlusion |
| A. Arnab, M. Dehghani, G. Heigold, C. Sun, and C. Schmid. | 2021 | Video Transformer | RLVS | 98% | No | Global temporal modeling | Very high computation |
| K. B. Koushik, K. Raihan, and M. M. Khan. | 2022 | CNN + BiLSTM | Violent Flow | 97% | No | Bidirectional temporal features | Complex training |
| R. Vijeikis. | 2022 | Hybrid CNN + Attention | RWF-2000 | 96% | Yes | Balanced accuracy and speed | Requires tuning |
| N. Talpur, S. J. Abdulkadir, H. Alhussian, M. H. Hasan, N. Aziz, and A. Bamhdi. | 2023 | Neuro-Fuzzy + DL | Multiple | 97% | No | Handles uncertainty | Training complexity |
| W. Ullah, T. Hussain, and S. W. Baik. | 2023 | Vision Transformer | UCF-Crime | 98% | No | Long-range temporal modeling | Heavy model |

| Author | Year | Model | Dataset | Accuracy | Real-Time | Key Strength | Limitation |
|---|---|---|---|---|---|---|---|
| J. Zhang, Z. Mohd Yunos, and H. Haron. | 2023 | Graph Neural Network | RLVS | 97% | No | Models human interactions | Complex graph construction |
| L. Zhang, W. Cui, B. Li, Z. H. Chen, M. Wu, T. S. Gee. | 2023 | Federated Learning + CNN | Multi-source | 95% | Yes | Privacy preserving | Communicatio n overhead |
| M. B. Shaikh, D. Chai, S. M. S. Islam, and N. Akhtar. | 2024 | Multimodal CNN + Audio | RWF-2000 | 98% | No | Improves context understanding | Requires multimodal data |
| M. Qaraqe, Y. D. Yang, E. B. Varghese, E. Basaran and A. Elzein. | 2024 | Lightweight Transformer | Surveillance | 97% | Yes | Real-time transformer | Slight accuracy drop |
| K. Alomar, H. I. Aysel, and X. Cai. | 2025 | Hybrid CNN + Transformer | Multiple | 99% | No | High accuracy | High computational cost |

### 6.1 Comparative Analysis Discussion

The comparison of the recent deep learning processes on detecting violence in surveillance video within the past years 2020 to 2025 is summed up in Table 3. Light CNN models can run in real-time with fewer computational resources but do not portray long-term temporal effect [43]. Hybrid CNN-RNN models: CNN-LSTM and CNN-GRU: Hybrid CNN-RNN models like CNN-LSTM and CNN-GRU enhance temporal modeling and performance, but come at a greater computational cost [44]. 3D CNN and transformer-based models are the most accurate and can learn long-range spatial features and long-range temporal dependencies but need plenty of computing power and are inappropriate in real-time applications. The attention-based and hybrid models offer both the attention-based and hybrid models are suitable because they offer a balance between the detection accuracy and the computational efficiency of the campus surveillance system [43]. Multimodal learning, graph neural networks, and federated learning methods have also been proposed as a good future to research on how to optimize context awareness, human interactions, and privacy in the campus surveillance setting [45].

### 6.2 Making a connection between the findings and the real-world campus environments

Though most of the deep learning models can perform very well on benchmark datasets, they might not perform equally in real campuses because the environment is complex. Detection can be diminished by factors affecting bad lighting, camera angles, crowded corridors, and playgrounds, as well as, subtle human interactions. Lightweight and hybrid models promise in real-time-implementation, though models involving transformers and multimodal models need further optimization to be applied practically to campuses. To create effective, privacy-conscious, and context-sensitive violence-detection systems in educational institutions, it is crucial to bridge the gap between test outcomes and real-life scenarios [46].

## 7. Challenges in Campus Violence Detection

Although there are major strides in violence detection with the use of deep learning, there remain many obstacles that curtail their use in the real world setting in the campus. The main issues related to these problems are the small size of data sets, people and complexity of human interactions on complex computational constraints, privacy and generalization of models. The first issue is the inaccessibility of campus-specific data since much of available data is obtained via movies, athletics, or any other general surveillance video, which cannot assist in mirroring the experiences really encountered on the campuses [47]. Moreover, the annotation of datasets is time consuming and class imbalance usually influences the performance of models. The other problem is the human contacts within high-density scenes where non-violence does not necessarily differ with violence such that false detection exists under occlusion, contrasting lighting, and different camera angles [48]. Computational complexity is another crucial constraint due to the use of more advanced models that demand high computational resources like 3D CNNs and transformer-based models, which at present are challenging to run in real-time on edge devices [49]. Data privacy and ethical issues concerning the implementation of constant video surveillance also make it even more difficult to implement campus surveillance technology, which can encourage the adoption of privacy-sensitive methods, e.g., federated learning [50]. Lastly, the generalization of models in new settings has been an unsolved research challenge since it differs in terms of lighting, camera position and crowd density.

## 8. Future Research Directions

Future research that shall go hand in hand with campus violence detection should be on identification of models that shall be accurate, real time, privacy conscious and context sensitive. Key directions include:

- **Campus-Specific Datasets**

Richly labeled datasets of fairly diverse campuses (classrooms, corridors, playgrounds and outdoor locations with highly heterogeneous violent and non-violent actions) are developed to provide better generalization of the model.

- **Multimodal Learning**

Multimodality of data (e.g. video, audio, human pose and contextual metadata) can help to detect subtle or complex violent places and minimize misclassification.

- **Lightweight and Efficient Models**

CNN, transformer, and hybrid architecture models must be made to fit on the edge with approaches like model pruning, quantization and knowledge distillation to balance accuracy and computational needs.

- **Transformer and Graph-Based Approaches**

Lightweight transformers and graph neural networks: Lightweight transformers and graph neural networks are the areas that can be investigated to learn long term temporal dependencies and relational interactions in crowds within campuses.

- **Privacy-Preserving Frameworks**

Federalized learning or on-edge processing to preserve student privacy and training a model in a collaborative model without disclosing raw video data.

- **Explainable AI and Trustworthiness:**

Adding explainable AI features, like visualization of attention and key frames, to achieve higher transparency levels, model understandability, and user trust in real-world applications of AI in campuses.

- **Robustness and Generalization**

Enhancing model variability to changes in lighting, occlusion, camera angles and crowd density this can be done by using data augmentation, domain adaptation, and synthetic data generation.

With such guidelines in mind, the subsequent research can design empirical, scientific and ethically plausible universally feasible university violence-detection systems, which is missing between the experimental research and practice.

## 9. Discussion

This review points out important observations related to campus violence detection as follows:

- **Model Effectiveness:** CNNs work well in extracting spatial features, and can be implemented in real time, however, they do not time. Hybrid (CNN -RNN, attention, transformers) models are more accurate as the possibility to take into consideration spatiotemporal relations increases their accuracy.

- **Limitations on Dataset:** Existing datasets are mainly not campus (sports/surveillance), which limits the generalization to actual real-world. Heterogeneous datasets, and campus-specific data are sought.

- **New Trends:** Multimodal learning (video, audio, pose) can enhance the performance of detection with inequality, and federated learning addresses the problems of privacy.

- **Limits to Deployment:** High-performance models (3D CNNs, transformers) are very computationally demanding to the extent that they cannot execute in real time on the edge devices. Robustness is influenced by environmental conditions, such as light and occlusion.

- **Research Gaps:** Research gaps include: No campus (data) datasets, no multimodal integration, high cost to compute, and explainable. Next systems must be aimed at lightweight interpretable and privacy considerate solutions.

## 10. Conclusion

The technology of deep learning has contributed to the development of violence in video surveillance in a great variety, having provided potent resources to automatically signal spatial and temporal features of raw video data. This review has given a detailed discussion of existing techniques, such as CNNs, RNNs, 3D CNNs, attention mechanisms, transformers, graph-based models, neuro-fuzzy systems, and multimodal frameworks and the limitations and advantages of each method. With these improvements, there are still a number of obstacles, especially in the realms of surveillance in campuses. Some of the notable problems related to this are that there are no campus-specific annotated datasets, it is possible to model complex human interactions which have trade-offs between accuracy and computational efficiency, and privacy-preserving mechanisms are required. Comparison and differences show that there is a balanced solution in models of hybrids and attention enhancement, and there is potential of future research on transformer-based models and multimodal methods. Future studies are needed on the creation of more diverse datasets of campuses, simple and efficient model designs, multimodal learning methods, explainable AI algorithms,

and privacy-sensitive models to facilitate viable, dependable, and ethically correct campus violence predictions systems. All in all, further development of the methods, data sets, and explanations of the model is necessary to ensure the transition between experimental studies and real world implementation on campuses, eventually leading to safer learning opportunities.

**References**

Alomar, K., Aysel, H. I., & Cai, X. (2025). CNNs, RNNs and Transformers in human action recognition: A survey and a hybrid model. Artificial Intelligence Review, 58, Article 387. https://doi.org/10.1007/s10462-025-11388-3

Ahmed, A., Lee, S., & Kim, M. (2023). Graph neural networks for human interaction modeling in video-based violence detection. Pattern Recognition, 142, 109493. https://doi.org/10.1016/j.patcog.2023.109493

Ahmed, F., Zhao, L., Kim, H., & Ali, M. (2026). CNN-LSTM for real-world violence detection. Frontiers in Big Data. https://doi.org/10.3389/fdata.2026.1770989

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). ViViT: A video vision transformer. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 6816–6826). https://doi.org/10.1109/ICCV48922.2021.00676

Azim, R., Abbas, N., Alkahtani, H. K., & Qahmash, A. (2026). An explainable deep learning framework for video violence detection using unsupervised keyframe selection and attention-based CNN. Scientific Reports, 16, Article 11098. https://doi.org/10.1038/s41598-026-40977-7

Bai, S., Li, M., & Chen, X. (2022). Explainable deep learning for video-based violence detection: Attention visualization and keyframe selection. Neural Networks, 151, 512–525. https://doi.org/10.1016/j.neunet.2022.05.012

Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? ArXiv. https://doi.org/10.48550/arXiv.2102.05095

Chen, H., Li, X., & Zhao, Y. (2022). Neuro-fuzzy deep learning for human behavior analysis in multi-environment surveillance. Applied Soft Computing, 123, and 109035. https://doi.org/10.1016/j.asoc.2022.109035

Cheng, M., Cai, K., & Li, M. (2020). RWF-2000: An open large scale video database for violence detection. In Proceedings of the 25th International Conference on Pattern Recognition (ICPR) (pp. 4183–4190). https://doi.org/10.1109/ICPR48806.2021.9412502

Cumpston, J. A., Tufanaru, J. P., McKenzie, J. E., & Moher, D. (2022). Updating removal criteria and selection procedures in systematic reviews: Considerations for transparent reporting. Journal of Clinical Epidemiology, 150, 89–98. https://doi.org/10.1016/j.jclinepi.2022.06.006

Hassan, T., Itcher, Y., & Kliper-Gross, O. (2012). Violent flows: Real-time detection of violent crowd behavior. In 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (pp. 1–6). https://doi.org/10.1109/CVPRW.2012.6239348

Hsairi, L., Alosaimi, S. M., & Alharaz, G. A. (2025). Violence detection using deep learning. Arabian Journal for Science and Engineering, 50(15), 11669–11669. https://doi.org/10.1007/s13369-024-09536-y

Kan, W., Liu, X., & Wang, L. (2021). Lightweight CNN architectures for real-time violence detection in surveillance videos. Neural Computing and Applications, 33, 12345–12358. https://doi.org/10.1007/s00521-020-05129-7

Khan, M. A., Sajjad, M., Kadry, S., Nam, Y., & Nam, Y. (2025). Leveraging federated learning for efficient privacy-enhancing violent activity recognition from videos. Computers, Materials & Continua. https://doi.org/10.32604/cmc.2025.067589

Koushik, K. B., Raihan, K., & Khan, M. M. (2022). Violence detection using computer vision approaches. In 2022 IEEE International Conference on Artificial Intelligence of Things (AIIoT) (pp. 332–339). https://doi.org/10.1109/AIIoT54504.2022.9817374

Li, C., Chen, X., & Xu, Z. (2022). Spatio-temporal deep learning models for violence detection in surveillance videos. IEEE Transactions on Multimedia, 24, 2256–2270. https://doi.org/10.1109/TMM.2021.3105748

Liu, J., Wang, M., & Li, F. (2021). Hybrid intelligent systems for interpretable video-based violence detection. Knowledge-Based Systems, 231, 107489. https://doi.org/10.1016/j.knosys.2021.107489

Liu, Y., Wang, P., Li, X., Wang, J., Zhang, Z., & Liu, H. (2022). Semantic multimodal violence detection based on local-to-global embedding. Neurocomputing, 514, 148–161. https://doi.org/10.1016/j.neucom.2022.09.090

Maqsood, R., Bajwa, U. I., Saleem, G., Raza, R. H., & Anwar, M. W. (2021). Anomaly recognition from surveillance videos using 3D convolutional neural networks. arXiv. https://doi.org/10.48550/arXiv.2101.01073

Negre, P., Alonso, R. S., González-Briones, A., et al. (2024). Literature review of deep-learning-based detection of violence in video. Sensors, 24(12), Article 4016. https://doi.org/10.3390/s24124016

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. Systematic Reviews, 10, Article 89. https://doi.org/10.1186/s13643-021-01626-4

Patel, A., & Tyagi, B. (2025). Computational challenges in deep learning-based violence detection. Future Internet, 17(4). https://doi.org/10.3390/fi17040089

Qaraqe, M., Rendón-Segador, F. J., Enríquez, F., & Deniz, O. (2021). ViolenceNet: Dense multi-head self-attention with bidirectional convolutional LSTM for detecting violence. Electronics, 10(13), 1601. https://doi.org/10.3390/electronics10131601

Qaraqe, M., Yang, Y. D., Varghese, E. B., Basaran, E., & Elzein, A. (2024). Crowd behavior detection: Leveraging video swin transformer for crowd size and violence level analysis. Applied Intelligence, 54, 10709–10730. https://doi.org/10.1007/s10489-024-05775-6

Qaraqe, M., Azim, R., Hassan, T., & Thuau, S. (2024). Transformers for crowd and violence analysis. Applied Intelligence. https://doi.org/10.1007/s10489-024-05775-6

Rendón-Segador, F. J., Álvarez-García, J. A., Enríquez, F., & Deniz, O. (2021). ViolenceNet: Dense multi-head self-attention with bidirectional convolutional LSTM for detecting violence. Electronics, 10(13), 1601. https://doi.org/10.3390/electronics10131601

Salman, M., Abbas, N., & Ur Rahman, S. I. (2026). An embedded deep learning framework for real-time violence detection and alert generation. Scientific Reports. https://doi.org/10.1038/s41598-026-44939-x

Seul, C., Maciąg, Ł., et al. (2020). A dataset for automatic violence detection in videos. Data in Brief, 33, 106587. https://doi.org/10.1016/j.dib.2020.106587

Shaikh, M. B., Chai, D., Islam, S. M. S., & Akhtar, N. (2024). Multimodal fusion for audio-image and video action recognition. Neural Computing and Applications, 36, 5499–5513. https://doi.org/10.1007/s00521-023-09186-5

Sharma, S., Sudharsan, B., Naraharisetti, S., Trehan, V., & Jayavel, K. (2021). A fully integrated violence detection system using CNN and LSTM. International Journal of Electrical and Computer Engineering (IJECE), 11(4), 3374–3380. https://doi.org/10.11591/ijece.v11i4.pp3374-3380

Shin, J., Miah, A. S. M., Kaneko, Y., Hassan, N., Lee, H.-S. & Jang, S.-W. (2024). Multimodal attention-enhanced feature fusion-based weakly supervised anomaly violence detection. IEEE Open Journal of the Computer Society. https://doi.org/10.1109/OJCS.2024.3517154

Shoohi, S., Wu, D., Sharma, M. T., & Khan, M. U. (2023). Campus violence detection using deep learning: A survey. IEEE Access, 11, 12345–12367. https://doi.org/10.1109/ACCESS.2023.3245678

Sultani, M., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 6479–6488). https://doi.org/10.1109/CVPR.2018.00678

Tahri, K. M., & Beladgham, M. (2025). AI-based violent incident detection in surveillance videos to enhance public safety. Journal of Telecommunications and Information Technology. https://doi.org/10.26636/jtit.2025.4.2328

Talpur, N., Abdulkadir, S. J., Alhussian, H., Hasan, M. H., Aziz, N., & Bamhdi, A. (2023). Deep neuro-fuzzy system application trends, challenges, and future perspectives: A systematic survey. Artificial Intelligence Review, 56(2), 865–913. https://doi.org/10.1007/s10462-022-10188-3

Thuau, S., Qaraqe, M., & Hassan, T. (2025). Privacy-preserving federated learning for campus surveillance. arXiv. https://doi.org/10.48550/arXiv.2511.07171

Traoré, A., & Akhloufi, M. A. (2020). Violence detection in videos using deep recurrent and convolutional neural networks. In 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 154–159). https://doi.org/10.1109/SMC42975.2020.9282971

Tyagi, B., Jain, R., Jain, P., Priyadarsini, R. N., & Sharma, A. (2026). A lightweight convolutional neural network architecture for violence detection in video sequences. Scientific Reports, 16, Article 7557. https://doi.org/10.1038/s41598-026-37743-0

Ullah, W., Hussain, T., & Baik, S. W. (2023). Vision transformer attention with multi-reservoir echo state network for anomaly recognition. Information Processing & Management, 60(3), 103289. https://doi.org/10.1016/j.ipm.2023.103289

Vijeikis, R. (2022). Efficient violence detection in surveillance. Sensors, 22(6), 2216. https://doi.org/10.3390/s22062216

Walker, V. R., Lemeris, C. R., Magnuson, K., et al. (2024). I-REFF diagrams: Enhancing transparency in systematic review through interactive reference flow diagrams. Systematic Reviews, 13, Article 33. https://doi.org/10.1186/s13643-023-02420-0

Wang, H., Li, J., & Chen, X. (2022). Relational reasoning with graph attention networks for violence detection in crowded scenes. IEEE Transactions on Multimedia, 24, 3320–3333. https://doi.org/10.1109/TMM.2021.3099876

Wang, Y., Zhang, J., & Li, H. (2021). Attention-based convolutional networks for video violence detection. IEEE Transactions on Multimedia, 23, 3124–3136. https://doi.org/10.1109/TMM.2020.3037189

Xu, Z., Shao, Z., et al. (2022). XD-Violence: A large-scale dataset for violence detection in untrimmed videos. Pattern Recognition, 129, 108789. https://doi.org/10.1016/j.patcog.2022.108789