International Journal of Engineering, Management and Humanities (IJEMH)
Volume 5, Issue 5, Sep.-Oct, 2024 pp: 20-28 ISSN: 2584-2145
www.ijemh.com

An Overview of Deep Learning Methods for Violence Detection

Xiuliang Zhang[1],Tadiwa Elisha Nyamasvisva[2],Chuntao Liu[3]

[1]Student, Infrastructure University Kuala Lumpur, Faculty of Engineering,Science and Technology, Xiuliang Zhang, 232924053@s.iukl.edu.my

[2]Professor, Infrastructure University Kuala Lumpur, Faculty of Engineering, Science and Technology, Tadiwa Elisha Nyamasvisva

[3]Student, Infrastructure University Kuala Lumpur, Faculty of Engineering,Science and Technology, Chuntao Liu

Corresponding Author: Xiuliang Zhang

Date of Submission: 09-09-2024 Date of Acceptance: 25-09-2024

ABSTRACT: Violence detection in video has gained significant attention due to its critical applications in public safety, surveillance systems, and automated monitoring. Traditional methods relying on handcrafted features and manual observation have been surpassed by deep learning techniques, which offer improved accuracy and the ability to automatically extract complex patterns from video data. This paper provides an overview of state-of-the-art deep learning methods for violence detection, focusing on Convolutional Neural Networks, hybrid CNN+LSTM models, skeleton-based approaches, and Transformer architectures. Additionally, we explore the challenges faced in vision-based violence detection, including the impact of lighting conditions, model complexity, dynamic environments, and the diversity of violent behaviors. Finally, we examine popular datasets used in this domain and compare the performance of different deep learning models to offer insights for future research.

KEYWORDS: Violence Detection, Deep Learning, Convolutional Neural Network, Transformer, Dataset, Model Performance.

I. INTRODUCTION

The detection of violent behaviors in videos has become an essential research area, with numerous applications in public safety, surveillance systems, and automated monitoring. As video-based data becomes increasingly prevalent, accurately identifying and responding to instances of violence in real time is crucial. Traditional methods of violence detection, often relying on manual monitoring or handcrafted feature extraction, have proven inadequate in dealing with the complexity, diversity, and scale of modern surveillance environments[1].

In recent years, deep learning has emerged as a powerful tool for tackling the challenges of violence detection[2]. Its ability to automatically extract high-level features from raw video data has enabled significant improvements in accuracy and efficiency. Deep learning models, particularly convolutional neural network (CNN), recurrent neural network (RNN), and transformer-based architectures, have been widely adopted due to their capacity to capture both spatial and temporal information critical for understanding violent actions. Furthermore, specialized techniques using human body skeleton representations and hybrid models combining CNN with long short-term memory (LSTM) networks have also contributed to the advancement of this field.

This paper provides a comprehensive overview of the key deep learning methods employed for violence detection in videos. We discuss approaches based on CNN, CNN+LSTM architectures, body skeleton representations, and transformers, highlighting their strengths and limitations. Additionally, we explore the challenges posed by varying lighting conditions, model complexity, dynamic environments, and the diversity of violent behaviors. A detailed examination of commonly used datasets for violence detection, including Hockey Fights, Action Movies, Violent Crowd, RWF-2000, Real Life Violence Situations, UCF-Crime Selected, and others, is provided. Finally, we present a comparative performance analysis of different deep learning models, illustrating their effectiveness on benchmark datasets. This survey aims to offer valuable insights into current methodologies and guide future research efforts in the field of violence detection.

II. Deep Learning Methods for Violence Detection

2.1 Violence Detection Method Based on CNN

A seminal contribution to the field of violence detection is the Two-Stream CNN architecture introduced by Simonyan and Zisserman[3]. This model separately processes spatial information from raw RGB frames and temporal information from optical flow, allowing for the effective capture of both appearance and motion features. The dual-stream approach was a breakthrough in action recognition and has laid the foundation for subsequent research in violent behavior detection, demonstrating robust performance in capturing the complexities of violence-related actions.

The integration of spatial and temporal features through Two-Stream Networks has been further refined by combining CNN for spatial processing with RNN or LSTM for temporal analysis. These networks outperform single-stream models, especially in tasks requiring the recognition of complex, dynamic actions like violent behaviors. By fusing spatial and temporal streams, Two-Stream Networks effectively capture both static and dynamic patterns, making them particularly suited for violent behavior recognition.

Graph Convolutional Network (GCN) have also emerged as a promising approach, particularly for action recognition based on skeletal data. GCN, as introduced by Jiang, M et al.[4], model the relationships between human joints as graphs, capturing both spatial and temporal dynamics. The Spatial-Temporal Graph Convolutional Network (ST-GCN) showed notable improvements in recognizing violent behaviors by modeling body part interactions, demonstrating the adaptability of GCN in handling structured data.

Another significant advancement in violence detection is the use of 3D Convolutions, which capture both spatial and temporal features within a unified framework. The C3D model, introduced by Tran et al., processes video clips directly using 3D convolutions, effectively capturing motion information across consecutive frames. This model has been highly successful in action recognition and has shown great potential in violence detection.

Building on the success of 3D convolutions, Inflated 3D ConvNet (I3D), proposed by Carreira and Zisserman, extend 2D CNN filters into 3D, allowing for enhanced spatiotemporal feature extraction. The I3D model has achieved state-of-the-art results on multiple action recognition benchmarks, underscoring its potential for applications in violent behavior recognition.

The hybrid approach of combining CNN with Recurrent Neural Network (RNN) has also shown promise. Donahue et al. introduced Long-term Recurrent Convolutional Network (LRCN), which leverage CNN for spatial feature extraction and LSTM for temporal modeling. This method improves the recognition of complex violent behaviors over extended sequences, offering a robust solution for video-based violence detection.

Recent advancements have focused on refining CNN architectures for better handling of violent behavior nuances. Sudhakaran and Lanz developed a Two-Stream 3D CNN, which integrates spatiotemporal feature extraction with a temporal segment network for feature aggregation[5]. This approach has shown superior performance on datasets like the Hockey Fight dataset, a standard in violence detection research.

Additionally, attention mechanism have been integrated into CNN framework to enhance performance. Zhou et al. introduced an Attention-Enhanced CNN that focuses selectively on the most relevant regions of video frames, improving the model’s ability to distinguish between violent and non-violent actions[1]. This approach has been evaluated on several datasets, demonstrating significant improvements in detection accuracy.

2.2 Violence Detection Method Based on CNN+LSTM

The combination of CNN and LSTM networks has emerged as a powerful approach for violence detection, effectively addressing both spatial and temporal aspects of video data[6]. CNN are well-suited for extracting spatial features from individual video frames, capturing local patterns such as edges, textures, and shapes, which are essential for identifying violent actions. However, CNN alone are insufficient for capturing the temporal dynamics inherent in video sequences, as they primarily focus on spatial relationships within single frames or short windows of frames. To overcome this limitation, LSTM networks, a type of RNN designed to handle sequential data, are integrated with CNN to model the temporal dependencies between video frames, allowing the system to understand the evolution of actions over time.

In this hybrid CNN+LSTM architecture, the CNN serves as the feature extractor, processing each frame or a set of consecutive frames to generate a rich representation of spatial features. These features are then passed to the LSTM network, which captures temporal information by maintaining a memory of previous frames, thus enabling the model to understand the progression of actions over the entire video sequence. LSTM, with their ability to mitigate vanishing gradient problems, are particularly effective in learning long-term dependencies, making them suitable for detecting violent actions that unfold over time, such as physical fights or aggressive gestures.

This combined architecture has shown significant improvements over traditional methods, as it allows for the simultaneous extraction of spatial and temporal features, which are both critical for accurate violence detection. The CNN+LSTM model can detect subtle cues in both appearance and motion, making it more robust to variations in camera angles, lighting conditions, and occlusions. For example, rapid movements or sudden changes in body posture, common indicators of violence, are captured by the LSTM component, while the CNN handles the spatial details that differentiate violent actions from non-violent ones.

Moreover, this method benefits from the end-to-end learning capability, where both the CNN and LSTM are trained together, optimizing the entire network to perform both spatial and temporal feature extraction seamlessly. By leveraging large, labeled video datasets for training, the CNN+LSTM model can generalize well to different environments and scenarios, improving its performance in real-world violence detection tasks, such as surveillance footage analysis or automated monitoring in public spaces.

2.3 Violence Detection Method Based on Body Skeleton

The violence detection method based on body skeleton analysis has emerged as a significant research focus in recent years. By capturing the spatial structure and motion patterns of key human joints, this method effectively identifies violent behaviors. Compared to traditional methods based on video frames or appearance features, body skeleton information offers greater robustness and generalization, especially when dealing with complex scene variations such as lighting changes, occlusions, or variations in clothing. Skeleton data can be extracted using depth cameras or pose estimation algorithms, providing an abstract representation of human movements, which allows the violence detection model to directly process high-level motion patterns without relying on low-level visual features[7].

Graph Convolutional Network (GCN) have been widely applied in skeleton-based action recognition tasks. One of the core advantages of GCN is their ability to naturally model the relationships between human joints by representing the human skeleton as a graph structure, where nodes represent joints, and edges represent the connections between them. The Spatial-Temporal Graph Convolutional Network (ST-GCN), proposed by Yan et al., represents a major advancement in this field. ST-GCN significantly improves the accuracy of violence detection by modeling both the spatial and temporal dynamics of human skeletons. It utilizes graph convolution operations to process the spatial structure of the skeleton, while temporal convolutions capture the evolution of joint movements over time. This enables ST-GCN to effectively extract hidden patterns of violent behavior, such as sudden limb movements or extreme changes in posture.

Skeleton-based violence detection models offer several advantages. First, due to their direct modeling of human posture, skeleton data is less affected by background noise or visual distractions, allowing the model to maintain stable performance even in complex environments. Additionally, this method is often computationally more efficient than pixel-based deep learning models, as skeleton data has a significantly lower dimensionality compared to raw video frames, reducing computational overhead. For resource-constrained real-time applications, such as security surveillance systems, skeleton-based violence detection provides a feasible lightweight solution.

2.4 Violence Detection Method Based on Transformer

The Transformer model, which has achieved significant breakthroughs in the field of natural language processing, has rapidly been applied to computer vision and has shown immense potential in action recognition and violence detection tasks[8]. The core advantage of the Transformer lies in its self-attention mechanism, which efficiently captures long-range dependencies in sequential data. This characteristic is particularly useful for video data, as violent behaviors often involve complex temporal dependencies, where short-term local information may be insufficient for accurately identifying the occurrence of violence.

Unlike traditional CNN, Transformer do not rely on local convolutional operations but utilize a global self-attention mechanism to dynamically weight the input sequences across both spatial and temporal dimensions. This allows Transformers to model both global and local spatial-temporal features simultaneously, enhancing their ability to capture violent behaviors. In particular, when dealing with videos that contain long intervals or complex action sequences, Transformers exhibit superior performance compared to CNN-based models, as they can focus on the most critical frames and information within the video sequence without increasing computational complexity.

For example, the Vision Transformer (ViT) is a vision model based on the Transformer architecture that processes input images by dividing them into multiple non-overlapping patches, converting them into sequential inputs for processing. Although ViT was initially designed for image classification tasks, its self-attention mechanism is equally applicable to violence detection in video sequences. ViT can extract global contextual information from different video frames, aiding in capturing the progressive evolution of violent behavior, particularly in scenarios where scenes and backgrounds are complex.

In the field of violence detection, Transformer-based approaches often employ hybrid architectures that combine Transformers with Convolutional Neural Networks to leverage the advantages of both methods. For instance, some studies have utilized CNN to extract spatial features at the frame level before processing the temporal sequences with Transformers. This combination ensures high-quality extraction of spatial features while enhancing the model's ability to handle temporal dependencies. Such models have demonstrated excellent performance across various violence detection datasets, proving their generalization capabilities in complex environments.

III. Challenges in Vision-Based Violence Detection

3.1 Impact of Lighting Conditions

Lighting conditions play a critical role in the effectiveness of vision-based violence detection models. Variations in illumination, such as low light, overexposure, or rapidly changing light, can significantly affect the clarity of video frames and degrade the performance of models that rely on spatial features extracted from RGB images. Poor lighting can obscure important visual cues, such as facial expressions or body movements, which are essential for identifying violent behaviors. Additionally, artificial lighting in indoor environments or nighttime surveillance footage introduces noise and shadows, further complicating the task. Although techniques like histogram equalization and contrast enhancement can mitigate some of these effects, handling extreme variations in lighting remains a persistent challenge. Future advancements may require more robust preprocessing techniques or adaptive models that can dynamically adjust to fluctuating lighting conditions.

3.2 Model Complexity

The complexity of deep learning models used for violence detection, particularly architectures like CNN and Transformer, poses both computational and practical challenges. These models often require significant computational resources, including high-end GPUs, large memory capacity, and extensive training time. This makes it difficult to deploy these models in real-time applications, such as surveillance systems with limited hardware capabilities. Moreover, large and complex models are prone to overfitting, especially when training on limited datasets, which may result in poor generalization to new environments or unseen violent behaviors. While model pruning, quantization, and knowledge distillation have been explored as strategies to reduce model complexity, the balance between model accuracy and computational efficiency remains an ongoing challenge. Developing lightweight models capable of operating effectively in resource-constrained environments is crucial for the broader adoption of vision-based violence detection systems.

3.3 Dynamic Environments

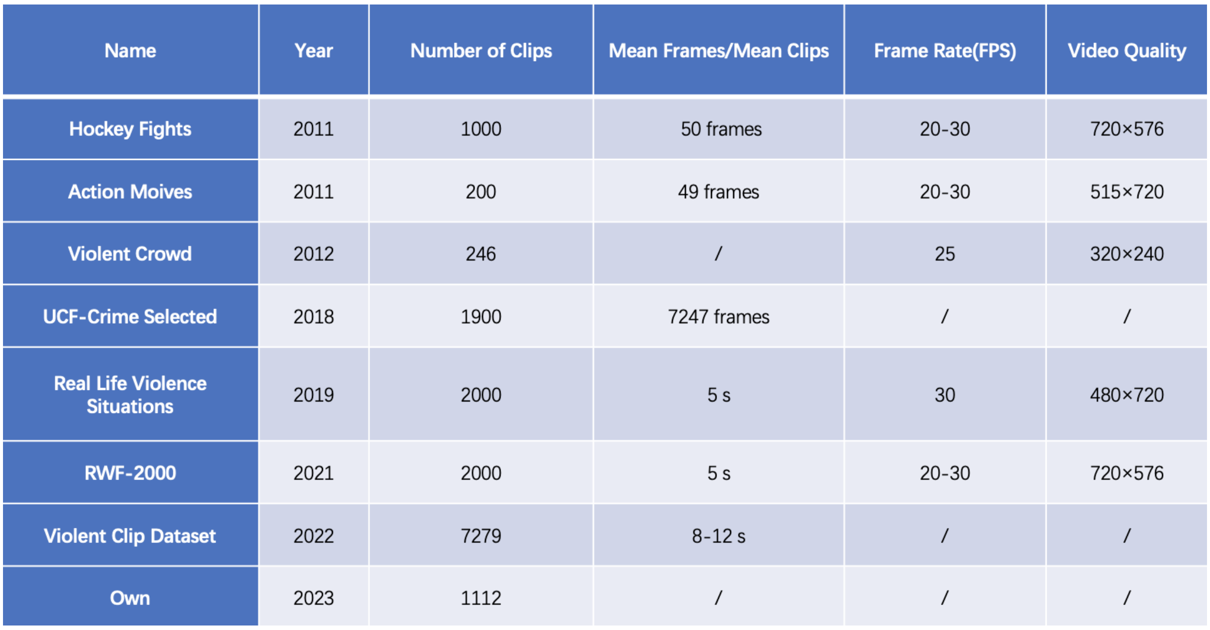
Dynamic environments, where elements such as background, camera angles, and crowd density constantly change, add significant complexity to violence detection tasks. In real-world surveillance scenarios, environments are rarely static; people move unpredictably, and camera viewpoints vary, causing frequent occlusions and making it difficult to consistently capture relevant information. These environmental changes can introduce substantial noise into the system, leading to false positives or missed detections. For example, in crowded spaces like stadiums or streets, a high density of people can obscure violent actions, while in open areas, distant violent actions may be challenging to detect due to limited spatial resolution. Effective violence detection models must be resilient to such environmental variations, requiring robust feature extraction techniques and more adaptive learning strategies that can dynamically adjust to different contexts and conditions.

3.4 Diversity of Violent Behaviors

Violent behaviors are highly diverse, ranging from physical altercations to more subtle forms of aggression, such as threats or intimidation. This diversity complicates the development of generalized detection models. Actions like pushing, punching, or kicking may follow predictable motion patterns, but more complex or less overt forms of violence, such as verbal aggression or psychological intimidation, are much harder to identify based solely on visual cues. Moreover, cultural differences in behavior can further complicate detection, as the same actions may be perceived as aggressive in some contexts but not in others. Models trained on specific datasets may struggle to generalize to different types of violence or diverse populations. To address this challenge, future research may focus on multi-modal approaches that incorporate audio, physiological data, or contextual information alongside video feeds, enabling a more comprehensive understanding of violent behaviors. Additionally, fine-tuning models on a wide range of datasets and incorporating domain adaptation techniques may help improve the detection of diverse forms of violence in different settings.

IV. Datasets for Violence Detection

Datasets play a crucial role in developing and evaluating models for violence detection. They provide structured video data that allow researchers to train and test algorithms across various scenarios. In this section, we examine several prominent datasets frequently used in violence detection research, including Hockey Fights, Action Movies, Violent Crowd, RWF-2000, Real Life Violence Situations, UCF-Crime Selected, Own, and Violent Clip Dataset. These datasets vary in terms of the environments they capture, the diversity of violent behaviors, and the complexity of the scenes. Summary of dataset characteristics as show in figure 1.



Name	Year	Number of Clips	Mean Frames/Mean Clips	Frame Rate(FPS)	Video Quality
Hockey Fights	2011	1000	50 frames	20-30	720×576
Action Moives	2011	200	49 frames	20-30	515×720
Violent Crowd	2012	246	/	25	320×240
UCF-Crime Selected	2018	1900	7247 frames	/	/
Real Life Violence Situations	2019	2000	5 s	30	480×720
RWF-2000	2021	2000	5 s	20-30	720×576
Violent Clip Dataset	2022	7279	8-12 s	/	/
Own	2023	1112	/	/	/

Figure 1.Summary of dataset characteristics

4.1 Hockey Fights Dataset

The Hockey Fights dataset is one of the earliest and most widely used datasets for violence detection. It contains 1,000 video clips of hockey games, where 500 clips involve fights and 500 represent normal gameplay. Each clip lasts around 2 seconds, recorded at 30 frames per second (FPS). The dataset focuses on detecting physical altercations between players on the ice. Although the controlled environment of an ice rink simplifies scene understanding, it also limits the variety of violence types to primarily brawling and shoving. Despite these limitations, the dataset has been instrumental in establishing baseline methods for violence detection, particularly in the early development of binary classifiers for violent and non-violent scenes.

4.2 Action Movies Dataset

The Action Movies dataset consists of a collection of violent and non-violent clips extracted from Hollywood action films. The dataset includes a broad range of violent acts, such as shootings, explosions, hand-to-hand combat, and chase scenes. The diversity of violent actions and cinematic techniques, including different camera angles, lighting conditions, and effects, makes this dataset a challenging benchmark for violence detection. However, it can introduce biases, as the violent actions depicted in movies are often exaggerated or dramatized, which might not reflect real-world scenarios. This dataset is valuable for testing models' ability to detect violence in more cinematic and exaggerated environments, offering a more varied and visually rich data source compared to sports or surveillance footage.

4.3 Violent Crowd Dataset

The Violent Crowd dataset is designed to capture violent incidents in crowded public spaces, such as protests, riots, or large gatherings. It contains over 246 video sequences of both violent and non-violent crowd behaviors, often recorded under uncontrolled conditions. The clips are sourced from real-world events, with varying camera angles, lighting, and levels of occlusion due to the density of people. The dataset is particularly useful for evaluating models in complex environments where violence is harder to discern due to the chaotic and dense nature of crowds. The ability to detect subtle signs of violence in a sea of human activity is a key challenge for models trained on this dataset, making it crucial for advancing public safety and surveillance applications.

4.4 RWF-2000 Dataset

The RWF-2000 (Real World Fights) dataset is a large-scale collection of 2,000 videos, half of which depict real-life violent altercations recorded in public spaces using handheld cameras or surveillance systems. The videos cover a wide range of real-world fight scenes, including street brawls, scuffles, and group fights. This dataset is unique due to the diversity in video quality, lighting conditions, and camera angles, making it highly representative of the challenges faced in real-world applications. RWF-2000 has become a widely used benchmark in violence detection, allowing researchers to develop models that generalize well to unstructured, non-cinematic settings. This dataset tests a model's ability to handle noisy data, unpredictable lighting, and sudden motion.

4.5 Real Life Violence Situations Dataset

The Real Life Violence Situations dataset consists of videos captured in everyday environments such as streets, malls, schools, and parks. The dataset includes both minor altercations and severe violence, covering a wide range of real-life violent incidents. This dataset is particularly challenging because it involves various unpredictable factors such as occlusions, rapid movements, and lighting variations, all of which make it difficult to accurately detect violence. The data is sourced from publicly available video platforms, surveillance footage, and personal recording devices, making it one of the most diverse datasets in terms of violent behavior and environmental complexity. This dataset is crucial for evaluating models that aim to detect subtle forms of violence in natural, unedited footage.

4.6 UCF-Crime Selected Dataset

The UCF-Crime Selected dataset is a subset of the larger UCF-Crime dataset, focusing specifically on violent crimes such as assaults, robberies, and armed violence. It contains over 1,900 long-duration videos from public surveillance cameras, including both violent and non-violent criminal activities. Unlike many datasets that provide short clips, UCF-Crime includes entire video sequences where violent acts may occur sporadically, requiring models to detect violence over extended periods of time. The dataset is particularly useful for evaluating the robustness of violence detection models in real-time monitoring systems, where the ability to locate violence within continuous video streams is essential.

4.7 Own Dataset

The Own dataset, developed specifically for research in violence detection, comprises a custom collection of violent and non-violent clips recorded in controlled and real-world environments. It typically includes a variety of violent acts such as physical altercations, aggressive gestures, and acts of vandalism. The dataset can be tailored for specific research objectives, such as testing new detection algorithms or creating benchmarks for a particular form of violence. It serves as a supplemental dataset that can complement existing publicly available datasets, offering additional variety or focus on niche scenarios that may not be present in other collections.

4.8 Violent Clip Dataset

The Violent Clip Dataset is a specialized collection of short clips (each lasting around 5–10 seconds) specifically focusing on violent interactions, such as fights, assaults, and aggressive behavior. The dataset includes 7279 labeled video clips from various sources, including YouTube, surveillance cameras, and personal video recordings. This dataset emphasizes fast and sudden bursts of violence, making it an excellent resource for testing models that need to respond quickly to rapid events. The videos are pre-labeled for violence, allowing researchers to focus on refining detection models rather than spending time on data annotation.

V. Performance Comparison of Deep Learning Models

In this section, we present a performance comparison of various deep learning models on the Violent Flow dataset, specifically focusing on the accuracy rates achieved by different algorithms. The models are categorized based on their architectures: CNN, CNN+LSTM, Skeleton-based CNN, and Transformer, The accuracy comparison is shown in Figure 2.

The Violent Flow dataset serves as a robust benchmark for evaluating violence detection models due to its diverse and challenging real-world scenarios. Variations in light conditions, camera angles, and types of violent behaviors provide a comprehensive testing ground for assessing the effectiveness of different approaches.

According to the study by Appavu et al., the CNN-based model achieved the highest accuracy of 99.68% on this dataset[9]. This demonstrates the model's strong capability in capturing spatial features and recognizing violent patterns from individual frames in the video sequences. In comparison, the CNN+LSTM model evaluated by Ullah et al. yielded an accuracy of 98.21%[10]. This method benefits from the combination of CNN for spatial feature extraction and LSTM for temporal sequence analysis, which is particularly useful in video-based violence detection.

The Skeleton-based CNN model discussed by Su et al. achieved a lower accuracy of 94.5%, reflecting the inherent challenges of using skeleton data to detect violent actions, especially in complex scenes where pose estimation may not be accurate or complete[11]. Nevertheless, this method remains effective in scenarios where human body movement is a key indicator of violent behavior.

Lastly, the Transformer-based approach, as presented by Akti et al., achieved an accuracy of 98%[12]. This method leverages the Transformer’s capability to model long-range dependencies and capture both spatial and temporal patterns in video data, providing a competitive performance compared to the CNN+LSTM models.

Overall, the comparison reveals that while CNN-based models lead in terms of accuracy on the Violent Flow dataset, both CNN+LSTM and Transformer models offer strong alternatives with high performance. Skeleton-based models, although slightly less accurate, can be valuable in specialized contexts where body movement plays a critical role in identifying violence. The choice of model thus depends on the specific requirements of the application, with CNN offering the highest accuracy and CNN+LSTM and Transformer models excelling in temporal analysis.

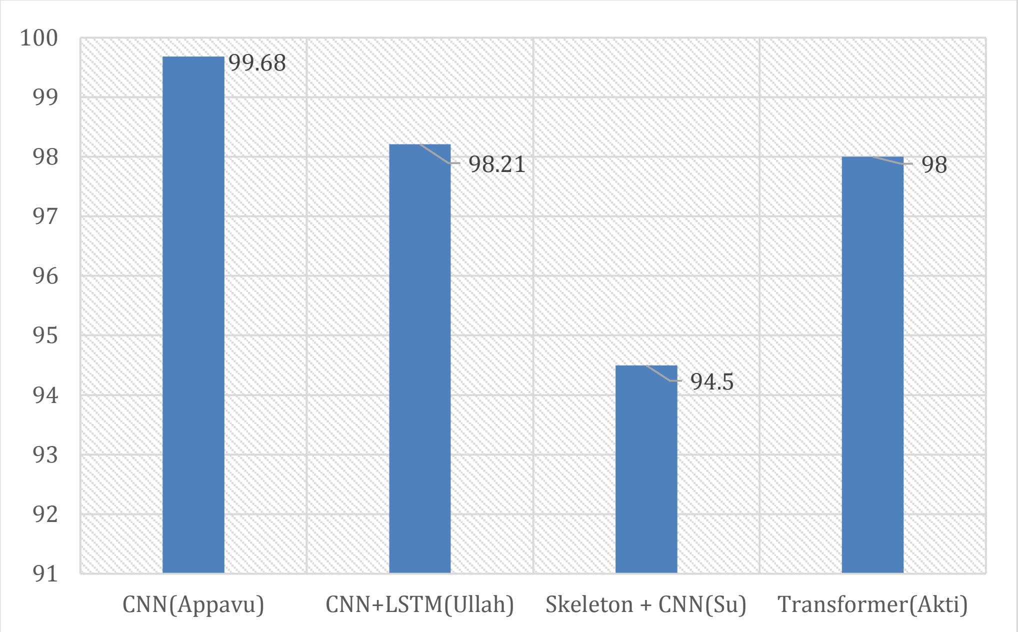


Figure 2.Accuracy obtained by selected items in the Violent Flow dataset.

VI. CONCLUSION

In this paper, we have reviewed the latest deep learning methods applied to violence detection in video, highlighting the strengths and limitations of various models, including CNN, CNN+LSTM architectures, skeleton-based approaches, and Transformer networks. Each method demonstrates unique capabilities in capturing spatial and temporal features essential for identifying violent behaviors, offering different solutions to the challenges posed by lighting conditions, model complexity, dynamic environments, and the diversity of violent actions. By examining widely used datasets and comparing model performance, we provide a comprehensive understanding of current methodologies. Future research should focus on addressing the remaining challenges while improving the generalization and efficiency of violence detection models in real-world scenarios.

REFERENCES

[1]. Zhou, P., et al. Violent interaction detection in video based on deep learning. in Journal of physics: conference series. 2017. IOP Publishing.

[2]. Bermejo Nievas, E., et al. Violence detection in video using computer vision techniques. in Computer Analysis of Images and Patterns: 14th International Conference, CAIP 2011, Seville, Spain, August 29-31, 2011, Proceedings, Part II 14. 2011. Springer.

[3]. Simonyan, K. and A. Zisserman, Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems, 2014. **27**.

[4]. Jiang, M., et al. Inception spatial temporal graph convolutional networks for skeleton-based action recognition. in 2022 International Symposium on Control Engineering and Robotics (ISCER). 2022. IEEE.

[5]. Sudhakaran, S. and O. Lanz. Learning to detect violent videos using convolutional long short-term memory. in 2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS). 2017. IEEE.

[6]. Mutegeki, R. and D.S. Han. A CNN-LSTM approach to human activity recognition. in 2020 international conference on artificial intelligence in information and communication (ICAIIC). 2020. IEEE.

[7]. Huang, L., et al. Part-level graph convolutional network for skeleton-based action recognition. in Proceedings of the AAAI conference on artificial intelligence. 2020.

[8]. Ulhaq, A., et al., Vision transformers for action recognition: A survey. arXiv preprint arXiv:2209.05700, 2022.

[9]. Appavu, N. Violence Detection Based on Multisource Deep CNN with Handcraft Features. in 2023 IEEE International Conference on Advanced Systems and Emergent Technologies (IC_ASET). 2023. IEEE.

[10]. Ullah, F.U.M., et al., AI-assisted edge vision for violence detection in IoT-based industrial surveillance networks. IEEE Transactions on Industrial Informatics, 2021. **18**(8): p. 5359-5370.

[11]. Su, Y., et al. Human interaction learning on 3d skeleton point clouds for video violence recognition. in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. 2020. Springer.

[12]. Aktı, Ş., et al. Fight detection from still images in the wild. in Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2022.