Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.067589

ARTICLE

Leveraging Federated Learning for Efficient Privacy-Enhancing Violent Activity Recognition from Videos

Moshiur Rahman Tonmoy[1], Md. Mithun Hossain[1], Mejdl Safran[2,*], Sultan Alfarhood[2], Dunren Che[3] and M. F. Mridha[4]

[1]Department of Computer Science and Engineering, Bangladesh University of Business and Technology, Dhaka, 1216, Bangladesh
[2]Research Chair of Online Dialogue and Cultural Communication, Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh, 12372, Saudi Arabia
[3]Department of Electrical Engineering and Computer Science, Texas A&M University-Kingsville, Kingsville, TX 78363, USA
[4]Department of Computer Science, American International University-Bangladesh, Dhaka, 1229, Bangladesh

*Corresponding Author: Mejdl Safran. Email: mejdl@ksu.edu.sa

Received: 07 May 2025; Accepted: 10 September 2025; Published: 23 October 2025

ABSTRACT: Automated recognition of violent activities from videos is vital for public safety, but often raises significant privacy concerns due to the sensitive nature of the footage. Moreover, resource constraints often hinder the deployment of deep learning-based complex video classification models on edge devices. With this motivation, this study aims to investigate an effective violent activity classifier while minimizing computational complexity, attaining competitive performance, and mitigating user data privacy concerns. We present a lightweight deep learning architecture with fewer parameters for efficient violent activity recognition. We utilize a two-stream formation of 3D depthwise separable convolution coupled with a linear self-attention mechanism for effective feature extraction, incorporating federated learning to address data privacy concerns. Experimental findings demonstrate the model's effectiveness with test accuracies from 96% to above 97% on multiple datasets by incorporating the FedProx aggregation strategy. These findings underscore the potential to develop secure, efficient, and reliable solutions for violent activity recognition in real-world scenarios.

KEYWORDS: Violent activity recognition; human activity recognition; federated learning; video understanding; computer vision

1 Introduction

Human activity recognition (HAR) systems and video surveillance have been significantly enhanced by artificial intelligence (AI), enabling advancements in crowd analysis, behavior monitoring, and public safety [1,2]. These AI-driven developments facilitate the automated recognition of violent activities, which is crucial for maintaining security in public spaces. For instance, recent data indicate that the global video surveillance market is expected to reach $88.71 billion by 2030, with AI integration being the key driver of this growth [3]. However, the integration of AI in surveillance raises critical concerns regarding data security, privacy, and substantial processing demands associated with high-resolution video data [4]. Violence recognition systems often analyze sensitive and identifiable information, thereby increasing the risk of privacy infringement [5]. In response to stringent data protection regulations such as the California Consumer Privacy Act (CCPA) and the General Data Protection Regulation (GDPR), there is a growing

Copyright © 2025 The Authors. Published by Tech Science Press.
This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

imperative for AI systems that not only deliver reliable performance but also provide robust privacy safeguards [6,7].

Addressing these challenges requires innovative approaches to ensure the privacy and efficiency of AI-based automated video surveillance systems. Surveillance videos frequently contain sensitive information, which necessitates stringent privacy measures. In addition, the analysis of large-scale video data poses significant memory and processing challenges, particularly when leveraging cloud-based servers, which can lead to accessibility issues for edge devices [8]. To mitigate these obstacles, researchers have explored privacy-preserving techniques such as secure multiparty computation, federated learning (FL), and differential privacy [5]. Among these, FL has gained prominence for enabling decentralized learning without compromising data privacy. There are several paradigms and variants of FL based on different criteria. For instance, the most common is the horizontal FL, where clients share the same feature space but differ in samples (e.g., surveillance cameras at different locations training on similar video features). In vertical FL, clients share samples but differ in features (e.g., different sensors for the same event), and federated transfer learning combines both horizontal and vertical settings when clients differ in both samples and features [9,10]. Despite these advancements, many existing methods encounter limitations related to real-time scalability, high memory consumption owing to encryption protocols, and excessive computational complexity [11,12]. Furthermore, achieving a balance between high classification accuracy and robust privacy guarantees remains challenging, particularly in resource-constrained environments or scenarios with noisy data [13]. The high memory demands of deep and complex models create barriers to their deployment on resource-constrained edge devices, emphasizing the need for lightweight deep learning (DL) architectures [14].

Motivated by the aforementioned challenges, we investigate an FL-based framework for privacy-enhancing decentralized learning and efficient violence recognition from videos. We incorporate a horizontal FL approach since each surveillance camera or client would hold the same feature space, i.e., video frames with the same modalities but different samples. Our DL classifier emphasizes a lightweight design, recognizing that both FL and on-device deployment shift the computational burden of the entire network to local devices that typically lack high-end processing capabilities. To achieve competitive accuracy, our model employs a two-stream architecture that integrates multiple blocks of 3D depthwise separable convolutions with a linear self-attention mechanism, which enables efficient extraction of spatial and temporal features. In addition, we enforced information fusion between streams through residual connections, thereby enhancing information flow and convergence. The key aspects of this study can be summarized as follows:

- We propose an innovative and effective framework for violent activity recognition integrated with FL to enhance user data privacy through decentralized learning
- Our lightweight model contains only approximately 1.104 million parameters and 4.42 MB in size, significantly fewer than state-of-the-art video classifiers, which typically require approximately 100 times more parameters
- Experiments on multiple datasets demonstrated competitive recognition accuracies attained by the model
- Proposed model outperformed popular models such as ViViT and TimeSformer in a comprehensive performance analysis.

The rest of the paper is organized as follows: Section 2 presents an overview of past works. Section 3 discusses the methodology in detail. Section 4 summarizes the experiment and findings, followed by a discussion of the overall study in Section 5. Finally, Section 6 concludes the study.

2 Related Works

Recent advancements in violence recognition within surveillance videos have utilized various DL architectures to enhance accuracy and real-time processing. Ullah et al. [15] conducted a comprehensive review of ML and DL methods for violence detection, highlighting existing challenges and future directions focused on surveillance scenarios. Liu et al. [16] proposed a human-centered attention mechanism that can dynamically emphasize the salient regions from videos associated with the action, leading to effective recognition. Aggarwal et al. [17] introduced a system combining MobileNetV2 with a BiLSTM layer, achieving 96% accuracy on CCTV footage. Similarly, Yadav et al. [18] presented a CNN and LSTM framework that achieved up to 98% accuracy and a processing speed of 131 frames/sec. Abbass and Kang [19] enhanced detection by integrating Convolutional Block Attention Modules (CBAM) and using data augmentation and Categorical Focal Loss to address class imbalance. Kumar et al. [20] proposed a lightweight transformer model for indoor violence detection that achieved up to 98% accuracy with occluded subjects. Mohammadi and Nazerfard [21] introduced a semi-supervised hard attention model using reinforcement learning, achieving high accuracies on the RWF and Hockey datasets, although it depends heavily on the quality of the reinforcement learning algorithms and may underperform in less controlled settings. Finally, Vijeikis et al. [22] combined a U-Net-like network with MobileNet V2 and an LSTM module, attaining around 82% accuracy and 81% precision on the RWF-2000 dataset. Many frameworks for protecting sensitive data have also been made available by recent advances in privacy-preserving video analytics. For instance, Frimpong et al. [23] developed Secrets In Motion (SIM), which employs Ciphertext Policy Attribute-Based Encryption (CP-ABE) and Multi-Key Homomorphic Encryption (MKHE) to control access to video content. Although SIM balances privacy with classification accuracy, it may face scalability issues and the computational complexity of its encryption methods, limiting its use in large-scale deployments. Feng et al. [24] introduced X-Stream, a flexible video transformer for privacy-preserving video stream analytics, featuring a declarative query interface for specifying privacy and content exposure preferences, an adaptive mechanism that selects appropriate privacy-preserving techniques at runtime, and an efficient execution engine optimized for multi-task deduplication and inter-frame inference. Gaikwad and Karmakar [25] presented the Privacy-Aware Person Search (PAPS) model for IoT surveillance, which processes data at the edge and fog layers to minimize privacy risks. Mehta et al. [26] introduced SETR-PKD, a seizure detection framework that uses optical flow features to preserve patient confidentiality. However, this method can be less effective in environments with minimal movement or where motion patterns are subtle, reducing its reliability in diverse settings. Singh et al. [27] employed ViViT for violence detection, incorporating data augmentation to enhance performance on smaller datasets. On the other hand, Pajon et al. [28] modified the Flow-Gated model and proposed an innovative approach called “Diff Gated” network, which attained improved results compared to the original model. However, they only experimented with the federated averaging (FedAvg) aggregation strategy, which assumes that data across clients follows a similar distribution, which is often unrealistic in surveillance settings. Similarly, Victor et al. [29] proposed an FL–based violence detection framework that extracts individual frames from the AIRTLab dataset and classifies them using pre-trained CNN backbones. However, their approach is limited to frame-level modeling, thus overlooking critical temporal dynamics, and is evaluated on a single dataset, raising concerns about the generalizability.

Earlier studies have primarily concentrated on enhancing human activity recognition (HAR) methods to improve the efficiency of recognition models. However, these advancements often come with significant computational demands and fail to adequately address crucial factors such as user data privacy, especially when dealing with sensitive footage of daily life or indoor activities. While some studies have explored privacy concerns such as incorporating FL, their approaches are often limited by issues such as a lack of convincing performance, poor generalizability, and high computational complexity that hinders scalability.

Additionally, many existing method poses the risk of high false positive rates and inconsistent performance across diverse real-world settings. Therefore, it is essential to develop lightweight architectures that effectively balance accuracy and model complexity. Overcoming these challenges is critical for creating scalable, secure, and efficient solutions for practical applications.

3 Methodology

In this section, we first provide an overview of Federated Learning, followed by the introduction of our proposed DL-based architecture for efficient and effective violent activity recognition, leveraging Federated Learning.

3.1 Federated Learning

Federated Learning, also known as FL, contrasts with traditional centralized learning by enabling multiple entities (known as clients) to train a model collaboratively without sharing private data [30]. Rather, clients only share model updates (e.g., weights or gradients) with the central server. The server aggregates these local updates to construct a global model and distributes the updated global parameters to the entities. The design of FL enhances data privacy and has broad applications across domains such as healthcare, surveillance, and beyond [31,32]. Aggregation strategies vary according to the target application, data distribution, and other factors. An efficient aggregation strategy is one of the popular research domains in FL, and researchers have explored various effective strategies for tackling various challenges while improving performance. In this study, we experimented with two popular strategies that are briefly discussed in the following subsections.

3.1.1 Federated Averaging (FedAvg)

Federated Averaging, widely known as FedAvg, is the most common and intuitive aggregation strategy for FL [30]. It aggregates model updates from multiple clients by simply calculating the weighted average of their locally computed updates, weighted by the size of each client's dataset. Mathematically,

$$w_t = \sum_{k=1}^{K} \frac{N_k}{N} w_t^k \tag{1}$$

where, w_t is the aggregated global model weights at round t, w_t^k is the update from the k-th client at round t, N_k is the number of data samples at the k-th client, N is total data samples across all clients, and K is the total number of participating clients.

3.1.2 Federated Proximal (FedProx)

The FedProx algorithm is an extension of FedAvg designed to address the heterogeneity of client data distributions and system constraints in FL [33]. It introduces a proximal term in the loss function that penalizes significant deviations between the local client model parameters and the global model parameters. The proximal term is scaled by the parameter μ, which controls the strength of the penalty. The optimization objective of FedProx can be expressed as:

$$\min_{w_k} \left[F_k(w_k) + \frac{\mu}{2} \| w_k - w^t \|^2 \right] \tag{2}$$

where, w^t is the global parameters in round t, w_k is the local parameters for client k after training, $F_k(w_k)$ is the local loss function for client k, μ is the proximal term coefficient ($\mu \in [0,1]$), and $\|w_k - w^t\|^2$ is the proximal term.

3.2 Proposed Architecture

The framework incorporates decentralized training based on FL principles, operating across multiple clients or data sources. As illustrated in Algorithm 1, the process begins with the initialization of the global model M_g on a central cloud server. The server then broadcasts an instance of M_g to all participating clients C (i.e., edge devices), where each client trains the local model M_c using its private data. After training, each client sends the parameters of M_c back to the server. The server aggregates these updates using a predefined strategy (e.g., FedAvg, FedProx, etc.), updates M_g with the aggregated parameters, and redistributes the updated M_g to all clients for the next training round. This iterative process continues until the global model converges or until the maximum number of rounds R is reached. The FL mechanism ensures user data privacy by keeping raw data on local devices throughout the training process. Additionally, the lightweight design of the DL classifier enables the model to run entirely on edge devices without relying on cloud communications, therefore, On-Device AI-powered deployment will ensure that user data remains secure on the device, even when deployed for real-world recognition. A visual illustration of the framework is presented in Fig. 1.

Algorithm 1: Workflow of the federated learning empowered recognition framework

Require: Number of rounds R, set of clients C, global model M_g, aggregation function A

Initialize global model M_g
for each round $r = 1$ to R **do**
    Broadcast M_g to all clients in C
    for each client $c \in C$ in parallel **do**
        Receive global model M_g on client c
        Train local model M_c on client data
        Send updated model M_c to the server
    end for
    Aggregate local models: $M_g \leftarrow A(\{M_c | c \in C\})$
end for
Return final global model M_g

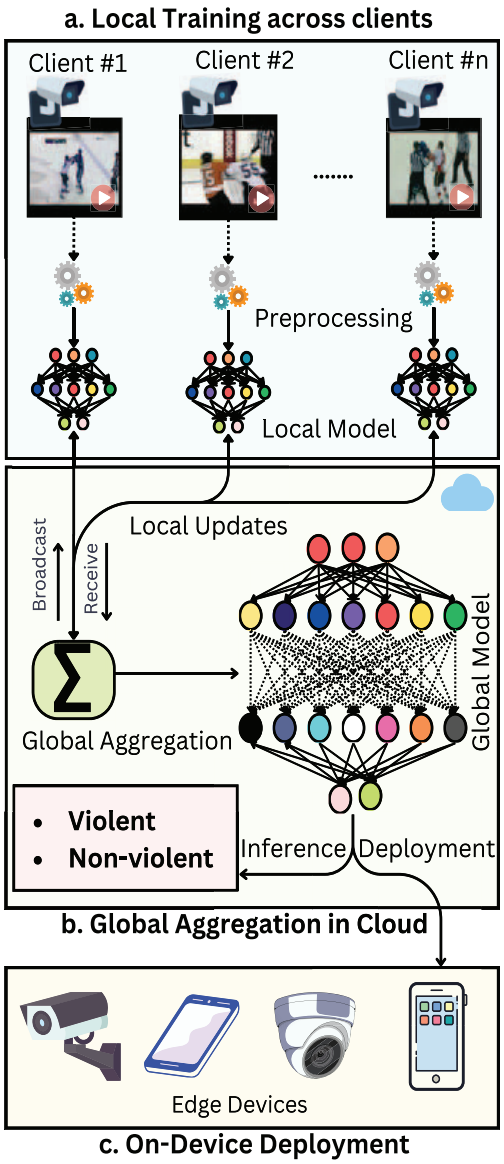

Figure 1: Overview of the FL empowered violent activity recognition

Fig. 2 illustrates the proposed classifier architecture and its key components. Fig. 2a provides an overview of the model, which comprises two-stream feature extraction, where each stream uses different filter sizes during the convolution operation to capture the diverse characteristics of the input frames. Fig. 2b shows the building block of the two-stream formation, the DepthConv block, which employs 3D depthwise separable convolution. In this process, depthwise convolution treats each input channel independently, followed by pointwise convolution to combine the channels. This mechanism efficiently reduces the computational burden while extracting essential features [34]. Mathematically, given an input tensor $\mathbf{X} \in \mathbb{R}^{C_{\text{in}} \times D \times H \times W}$, first a 3D convolution is applied independently to each input channel:

$$\mathbf{X}_d = \mathbf{W}_d * \mathbf{X}, \quad \mathbf{W}_d \in \mathbb{R}^{C_{\text{in}} \times K_d \times K_h \times K_w}$$

where $*$ represents the convolution operation and $\mathbf{W}_d$ is the depthwise kernel. Next, a $1 \times 1 \times 1$ convolution is applied to project the feature maps to C_{out} channels:

$$\mathbf{X}_p = \mathbf{W}_p * \mathbf{X}_d + \mathbf{b}_p, \quad \mathbf{W}_p \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}} \times 1 \times 1 \times 1}$$

where $\mathbf{W}_p$ is the pointwise kernel, and $\mathbf{b}_p$ is the bias. We also employed batch normalization within the DepthConv block to enhance training efficiency further. Additionally, the GELU activation function is used throughout the architecture to introduce non-linearity, as GELU offers smoother and more stable gradient propagation compared to ReLU, making it a preferred choice for many DL tasks [35].

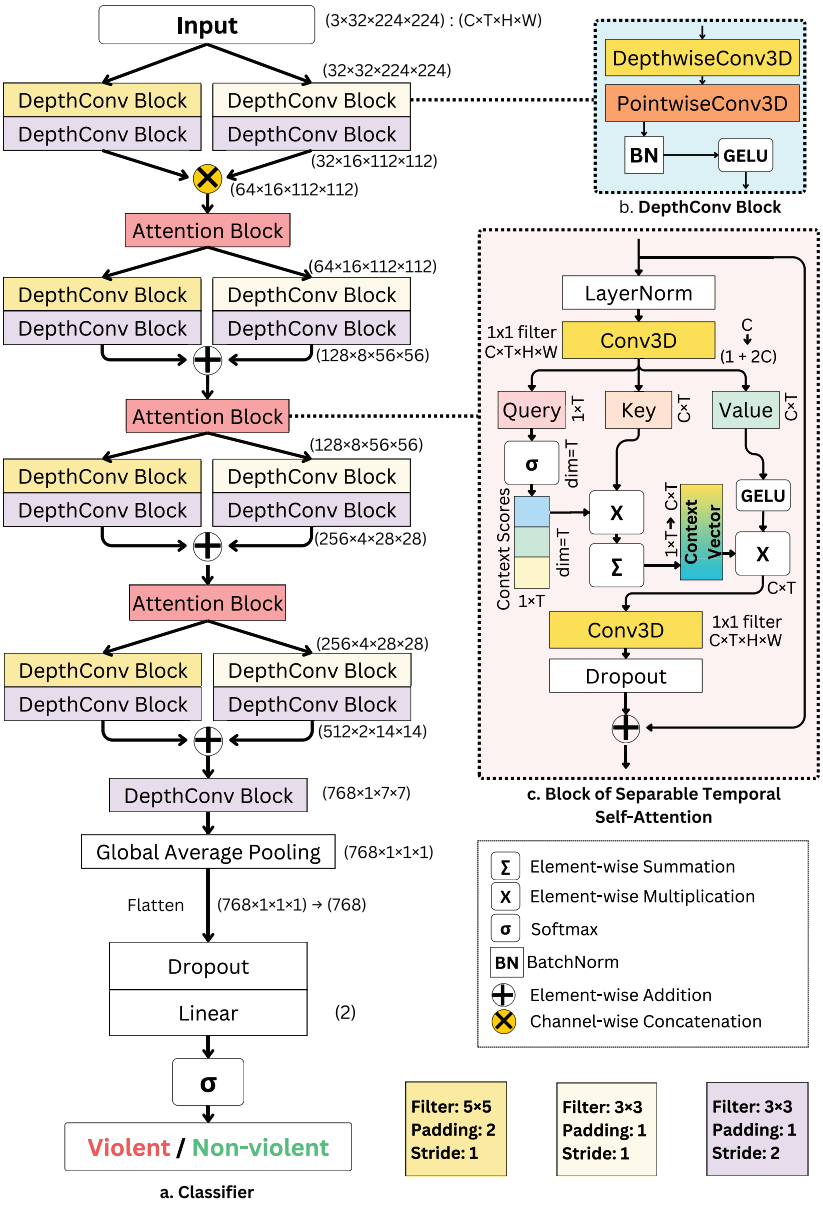

Figure 2: Outline of employed classifier and its components: (**a**) the classifier with two-stream feature extraction, (**b**) DepthConv Block comprising Depthwise Separable 3D Convolution, (**c**) Attention Block employing Separable Self-Attention over the temporal dimension

Fig. 2c illustrates the attention block, which is a key component of our model that enhances performance while minimizing computational complexity. This block employs a separable temporal self-attention mechanism that is integrated with a residual connection. The MobileViTV2 image classification model inspires the self-attention mechanism [36], which operates with linear complexity, making it well-suited for mobile devices. In our implementation, we adapted the attention mechanism to use 3D convolution and GELU activation instead of the linear operations with ReLU non-linearity for improved performance in video-based tasks. Given an input tensor $\mathbf{X} \in \mathbb{R}^{C\times D\times H\times W}$, where C is the embedding dimension, the input is first normalized using Layer Normalization. Next, a $1 \times 1 \times 1$ convolution projects the normalized input into query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) tensors:

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = \text{split}(\mathbf{W}_{qkv} * \mathbf{X}_n)$$

where $\mathbf{W}_{qkv} \in \mathbb{R}^{(1+2C)\times C\times 1\times 1\times 1}$ is the convolutional kernel. The context scores (A) are then computed using softmax over the temporal dimension of the Q, and the K is weighted by the context scores and summed along the attention dimension for context vector (C):

$$\mathbf{A} = \text{softmax}(\mathbf{Q})$$
$$\mathbf{C} = \sum_d \mathbf{A} \odot \mathbf{K}$$

where $\odot$ denotes element-wise multiplication. The output is refined by applying GELU activation to the value tensor (V), then modulating it with the expanded context vector (C):

$$\mathbf{Y}_a = \text{GELU}(\mathbf{V}) \odot \mathbf{C}$$

A final $1 \times 1 \times 1$ convolution projects back to the embedding dimension:

$$\mathbf{Y} = \mathbf{W}_o * \mathbf{Y}_a$$

where $\mathbf{W}_o \in \mathbb{R}^{C\times C\times 1\times 1\times 1}$. Lastly, dropout is applied, and a residual connection is added:

$$\mathbf{X}_{\text{out}} = \mathbf{X} + \text{Dropout}(\mathbf{Y})$$

In the end, we employed global average pooling to summarize the overall influence of the extracted features for classification. The resulting tensor is flattened and passed through a linear layer with a dropout operation to prevent overfitting. The final classification was achieved using softmax scores.

4 Experiment and Result

This section outlines the experimental details and findings of our study, starting with the data preparation and training setup, followed by performance analysis and ablation study, and concluding with the performance comparison with baseline models.

4.1 Experimental Data, and Setup

We conducted experiments using three separate datasets to evaluate the performance of our proposed model, simulating decentralized learning. The first dataset is the Hockey Fight (HF) detection dataset [37], which consists of 1000 clips (500 fight and 500 non-fight) from the National Hockey League (NHL). Each clip contains 50 frames with a resolution of 720 × 576 pixels. The second dataset is the Weapon Violence Dataset 2.0 (WVD) [38], a synthetic dataset containing 334 videos, which are divided into 60 instances of

hot violence (representing violence with firearms), 54 instances of cold violence (representing violence with traditional weapons), and 54 instances labeled as no violence. Lastly, we used the Firearm Action Recognition dataset [39], consisting of 398 videos representing actions with no gun (118 videos), handguns (141 videos), and machine guns (139 videos).

To prepare the datasets for the experiment, we downsampled the resolution of the original clips to 224×224 pixels and uniformly sampled 32 frames from each clip to ensure consistency. In the HF dataset, we replaced the original class names "fight" and "non-fight" with "violent" and "non-violent." Additionally, we merged the "cold violence" and "hot violence" clips from the WVD dataset into a single "violent" class. The datasets were then split into three parts, maintaining a 60-20-20 ratio for training, validation, and test evaluation. Extensive on-the-fly augmentation was applied during training using the Albumentations library [40], as summarized in Table 1. All three data splits were normalized with a mean of (0.485, 0.456, 0.406) and a standard deviation of (0.229, 0.224, 0.225).

Table 1: Overview of employed on-the-fly data augmentation strategies

Attribute	Probability
Horizontal flip	50%
Random Brightness/Contrast	50%
Shift, Scale, Rotate (limits: 0.05, 0.05, 30°)	50%
RGB Shift (shift limits: 15)	50%
Hue, Saturation, Value shift (limits: 20, 30, 20)	50%

The experimental implementations were carried out using Python and PyTorch to simulate the FL training system. The training data were partitioned into two subsets to simulate two participating local clients, each using their respective local data. Each client was trained on their local data for two consecutive epochs per round. To enhance the training process, we implemented callback strategies, including learning rate reduction and early stopping, based on the validation set performance of the global model after aggregation. If the validation accuracy showed no improvement for five consecutive rounds, the learning rate was reduced by 50% globally. Early stopping was triggered if no improvement in validation accuracy occurred over 30 consecutive rounds. All training and testing experiments were conducted in the Kaggle Notebook environment, utilizing two NVIDIA Tesla T4 GPUs with 29 GB of RAM. A summary of the experimental settings for the proposed model is provided in Table 2.

Table 2: Summary of the experimental settings

Attribute	Value
Frame shape	$3 \times 224 \times 224$
No. of frames	32
Batch size	8
Initial learning rate	$1e-3$
Minimum learning rate	$1e-8$
Max communication round	500
Local epoch	2
LR reduction patience	5
Early stopping patience	30

(Continued)

Table 2 (continued)

Attribute	Value
Weight decay	$1e-3$
Dropout rate	40%
Activation	GELU
Optimizer	Adam
Loss	Categorical crossentropy

4.2 Performance Analysis

The model consistently outperformed across all datasets when trained with the FedProx strategy compared to FedAvg. As shown in Table 3, FedProx achieved a test accuracy of 97.50% and an AUC of 0.9958 on the HF dataset, whereas FedAvg reached only 94.99%, marking a 2.58% drop in accuracy. A similar trend was observed on the WVD dataset, where FedProx achieved 97.06% accuracy and an AUC of 0.9616—an improvement of nearly 10% over the 88.24% accuracy obtained using FedAvg. On the Firearm dataset, FedProx also led to an 8.46% increase in accuracy compared to FedAvg. It is important to note that the small number of test samples per class had a significant impact on the final accuracy. For example, as illustrated in Fig. 3, a single misclassification out of 34 samples in WVD yielded an accuracy of 97.06%. Similarly, five misclassifications out of 200 samples in HF resulted in 97.50% accuracy, while three misclassified samples out of 80 in the Firearm dataset led to a 96.25% accuracy.

Table 3: Classwise performance on experimental test sets across global aggregation strategies (Acc. = Accuracy, P = Precision, R = Recall)

Dataset	Strategy	Classes	Acc (%)	P (%)	R (%)	F1-score	AUC	Support
HF	FedAvg	Non-violent	–	94.79	94.79	0.9479	–	96
		Violent	–	95.19	95.19	0.9519	–	104
		Overall	94.99	94.99	94.99	0.9499	0.9874	200
	FedProx	Non-violent	–	97.90	96.88	0.9738	–	96
		Violent	–	97.14	98.08	0.9761	–	104
		Overall	97.50	97.50	97.50	0.9750	0.9958	200
WVD	FedAvg	Non-violent	–	75.00	75.00	0.7500	–	8
		Violent	–	92.31	92.31	0.9231	–	26
		Overall	88.24	88.24	88.24	0.8824	0.8654	34
	FedProx	Non-violent	–	100.00	87.50	0.9333	–	8
		Violent	–	96.30	100.00	0.9811	–	26
		Overall	97.06	97.17	97.06	0.9699	0.9616	34
Firearm	FedAvg	machine_gun	–	86.96	76.92	0.8163	–	26
		Handgun	–	80.65	100.00	0.8929	–	25
		no_gun	–	100.00	89.66	0.9456	–	29
		Overall	88.75	89.71	88.75	0.8871	0.9802	80
	FedProx	machine_gun	–	96.00	92.31	0.9412	–	26
		Handgun	–	92.31	96.00	0.9412	–	25
		no_gun	–	100.00	100.00	1.0000	–	29
		Overall	96.25	96.30	96.25	0.9625	0.9952	80

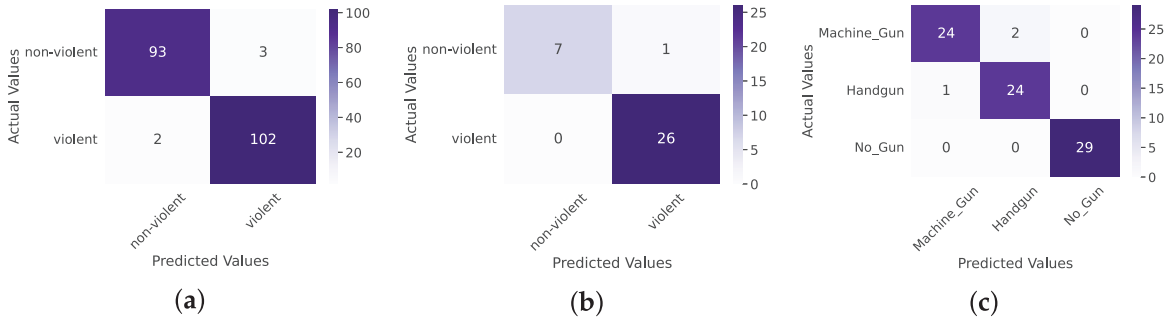

Figure 3: Confusion matrix of the test set evaluation using the FedProx strategy. (**a**) HF dataset; (**b**) WVD dataset; (**c**) Firearm dataset

To analyze the impact of FedAvg and FedProx on performance, Fig. 4 illustrates the training accuracy and loss across the communication rounds for each client using the HF dataset. Fig. 4a,c shows that FedAvg completed only 43 rounds before triggering early stopping, while Fig. 4b,d demonstrates that FedProx sustained training for 78 rounds, progressively enhancing robustness, delaying the early stopping threshold, and improving global validation accuracy. Fig. 4 also illustrates the average training accuracy and loss with the global validation performance. Although FedAvg initially achieved a peak validation accuracy of 95.50%, its performance deteriorated in subsequent rounds, as shown in Fig. 4e. In contrast, FedProx exhibited a steady improvement in validation accuracy, aligned with the average training accuracy, ultimately reaching 98.00% validation accuracy, as seen in Fig. 4f. A common trend observed in Fig. 4 is that validation accuracy consistently outperformed training accuracy. This can be attributed to regularization techniques, such as dropout and L2 regularization, applied during training. Furthermore, the validation accuracy reflects the performance of the aggregated model, which is inherently more robust than individual client models, contributing to its superior validation performance.

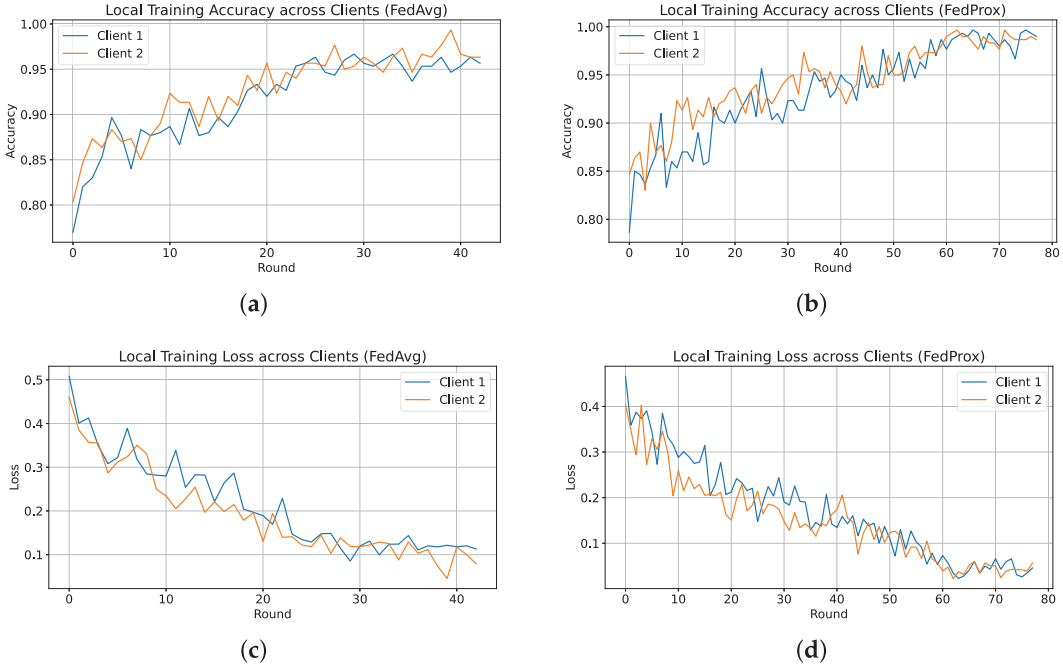

Figure 4: (Continued)

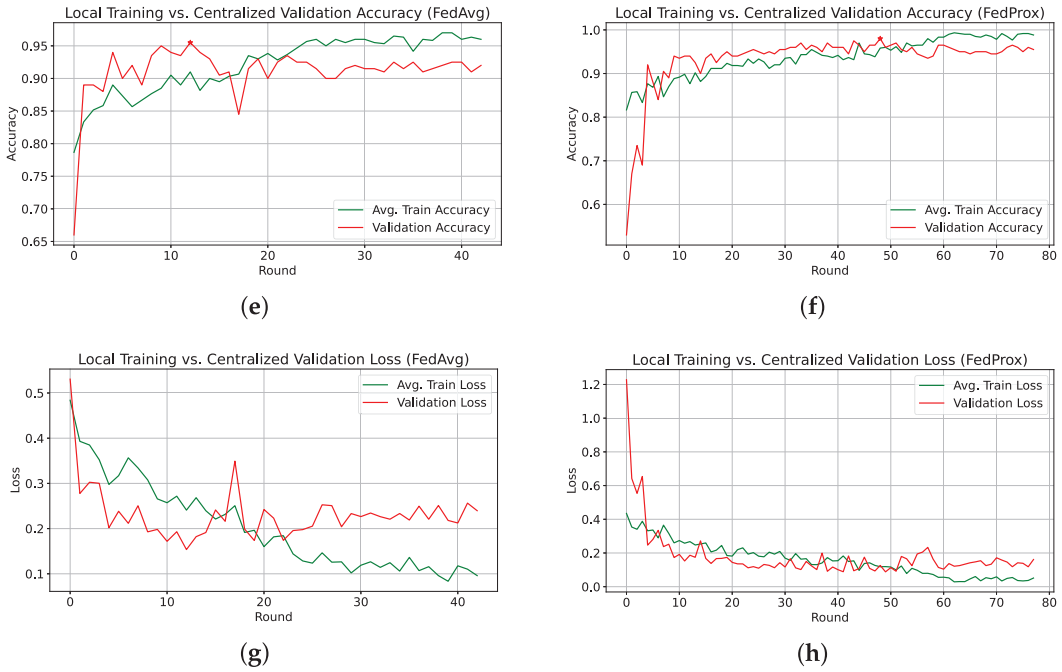

Figure 4: Overview of local training and centralized validation performance using FedProx and FedAvg aggregation strategies on the HF dataset. (**a**) Local Train Accuracy with FedAvg; (**b**) Local Train Accuracy with FedProx; (**c**) Local Train Loss with FedAvg; (**d**) Local Train Loss with FedProx; (**e**) Avg. Train vs. Global Validation Accuracy with FedAvg; (**f**) Avg. Train vs. Global Validation Accuracy with FedProx; (**g**) Avg. Train vs. Global Validation Loss with FedAvg; (**h**) Avg. Train vs. Global Validation Loss with FedProx

4.3 Ablation Study

To validate the effectiveness of the model's architecture, we conducted an ablation study focusing on the dual-stream architecture and the attention block, which are the core components of our proposed model. Table 4 summarizes the results of various configurations evaluated on the test set. The findings show a significant decline in performance across all metrics when attention blocks are excluded or partially included. Without any attention blocks, the model achieved the lowest accuracy of 94.00% with mono-stream and 93.50% with dual-stream. However, the incremental inclusion of attention blocks led to improved performance in both mono- and dual-stream formations. The proposed architecture, incorporating all three attention blocks in the dual-stream formation for effective feature extraction, achieved a test accuracy of 97.50% and an AUC score of 0.9958, underscoring the crucial role of the attention mechanism and the effectiveness of multi-stream feature extraction in enhancing the model's performance.

Table 4: Summary of the test set performance during ablation study of the proposed model (✗ → none, Acc. = Accuracy, P = Pecision, R = Recall)

Stream	Attention block	Params (M)	Acc. (%)	P (%)	R (%)	F1-score	AUC
Mono	✗	0.701	94.00	94.08	94.00	0.9400	0.9823
	Only first block	0.713	95.00	95.00	95.07	0.9499	0.9864

(Continued)

Table 4 (continued)

Stream	Attention block	Params (M)	Acc. (%)	P (%)	R (%)	F1-score	AUC
	First two blocks	0.763	94.50	94.71	94.67	0.9450	0.9921
	All three blocks	0.961	96.00	95.99	95.99	0.9599	0.9923
Dual	✗	0.843	93.50	93.50	93.50	0.9350	0.9763
	Only first block	0.856	94.50	94.50	94.50	0.9450	0.9878
	First two blocks	0.906	96.00	96.08	96.00	0.9600	0.9936
	All three blocks	**1.104**	**97.50**	**97.50**	**97.50**	**0.9750**	**0.9958**

4.4 Performance Comparision

To benchmark the performance of the proposed model against established video classification approaches, we conducted experiments using transfer learning with two widely recognized transformer-based video classifiers: ViViT [41] and TimeSformer [42]. For a fair comparison, both baselines were trained under the same experimental setup (input resolution, frame sampling, preprocessing, splits, and federated configuration) and optimized with the FedProx aggregation strategy, since our proposed model achieved its best performance with FedProx. We employed an additive fine-tuning approach with their pre-trained base architectures. Specifically, we froze the pre-trained layers of each model and appended two trainable dense layers at the end, incorporating a GELU non-linearity between them. This configuration added over one million trainable parameters to each model, enabling task-specific learning while leveraging the representational power of the pre-trained transformers. Table 5 presents the performance metrics obtained from the experiments. Among the baseline models, TimeSformer achieved the highest test accuracy of 96.00% for the HF dataset and 97.05% for the WVD dataset, highlighting its strength in video classification tasks. However, these results slightly lag behind the performance of our proposed model, emphasizing the effectiveness of our approach. Similarly, ViViT recorded the lowest F1-score of 0.945 for HF and 0.9101 for WVD, compared to 0.975 and 0.9699 achieved by our model. Both ViViT and TimeSformer showed significantly poorer results for the Firearm dataset, while our proposed model attained over 96% accuracy. Fig. 5 provides a visual comparison of the F1-scores produced by these models, demonstrating the superior capability of the proposed model in recognizing violent videos.

Table 5: Overview of the test set performance of the fine-tuned baseline models

Model	Params (M)	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score	AUC
ViViT	554.2	HF	94.50	94.50	94.50	0.9450	0.9905
		WVD	94.12	96.43	87.50	0.9101	0.9856
		Firearm	80.00	85.48	77.59	0.7919	0.9432
TimeSformer	777.2	HF	96.00	96.02	96.00	0.9600	0.9915
		WVD	97.05	98.14	93.75	0.9572	**0.9903**
		Firearm	78.75	78.79	78.75	0.7872	0.8962
Proposed	1.104	HF	**97.50**	**97.50**	**97.50**	**0.9750**	**0.9958**
		WVD	**97.06**	**97.17**	**97.06**	**0.9699**	0.9616
		Firearm	**96.25**	**96.30**	**96.25**	**0.9625**	**0.9952**

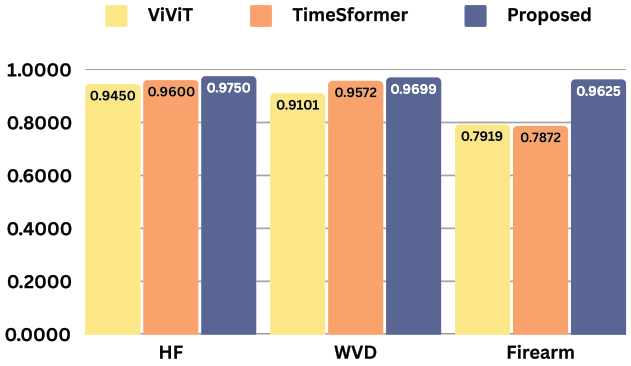

Figure 5: F1-score visual comparison among baseline models and the proposed one

5 Discussion

This study introduces a DL-based method for violent activity recognition that prioritizes user data privacy by employing FL and computational efficiency by incorporating a lightweight design tailored for edge deployment. With only approximately 1.104 million parameters and a model size of 4.42 MB, our model achieves competitive performance, with test set accuracies ranging from 96.25% to 97.50%. This strong performance is a result of our efficient architectural design, which integrates lightweight operations such as depthwise separable 3D convolutions and linear attention mechanisms, and notably the use of the FedProx aggregation method. However,

While our study demonstrates promising results, this study focuses on simulating the feasibility of lightweight models within FL frameworks for violent activity recognition and the key next step is to validate the proposed model on representative edge hardware and quantify practical deployment trade-offs. Future studies with large-scale and more diverse datasets will enhance the statistical reliability and generalizability of our findings. Moreover, although FedProx can effectively handle challenges posed by heterogeneous data distribution via its proximal penalty term, extended experiments with more clients and varying data distributions will better reflect the robustness of our model and FL strategy against the heterogeneity of real-world surveillance systems. In addition, the exploration of the impact of tailored augmentations specific to violent activity characteristics and the role of varying experimental setups (e.g., varying batch sizes) is left for future research. On the other hand, models can sometimes learn spurious or irrelevant patterns while still producing correct predictions; integrating explainable AI (XAI) techniques (e.g., attention-map

visualizations, and gradient-based methods) is an important future direction to provide impactful qualitative analyses and to clarify model decision-making, thereby increasing the trustworthiness of HAR applications.

By pursuing these directions, we anticipate establishing a versatile, privacy-preserving and trustworthy activity recognition framework that can adapt to a wide range of applications—from public safety monitoring to assistive care, while maintaining the stringent data-protection guarantees required in sensitive environments.

6 Conclusion

We propose an effective and efficient DL model for automated violent activity recognition from videos, designed with a focus on real-world, resource-constrained applications. Our approach emphasizes decentralized training, eliminating reliance on cloud servers, and ensuring enhanced user data privacy by keeping data confined to the native device. The proposed lightweight architecture achieved competitive test set accuracies across multiple experimental datasets, outperforming larger pre-trained models that have more than 100 times the number of parameters. These findings not only highlight the model's effectiveness but also demonstrate its generalization potential across diverse scenarios. Additionally, we found that the FedProx aggregation strategy improves the performance of the employed model over the FedAvg strategy. This study puts a significant step toward developing privacy-enhancing and efficient solutions for violent activity recognition, deployment on edge devices, and beyond, facilitating the broader adoption of secure and effective AI systems in sensitive applications.

Acknowledgement: The authors extend their appreciation to the Research Chair of Online Dialogue and Cultural Communication, King Saud University, Saudi Arabia, for funding this research.

Funding Statement: This work was supported by the Research Chair of Online Dialogue and Cultural Communication, King Saud University, Saudi Arabia.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Moshiur Rahman Tonmoy; methodology, Moshiur Rahman Tonmoy and Md. Mithun Hossain; software, Moshiur Rahman Tonmoy and Md. Mithun Hossain; validation, Mejdl Safran, Sultan Alfarhood and M. F. Mridha; formal analysis, Md. Mithun Hossain; investigation, Moshiur Rahman Tonmoy and Md. Mithun Hossain; resources, Mejdl Safran and M. F. Mridha; data curation, Md. Mithun Hossain; writing—original draft preparation, Moshiur Rahman Tonmoy and Md. Mithun Hossain; writing—review and editing, Mejdl Safran, Dunren Che and M. F. Mridha; visualization, Dunren Che and Sultan Alfarhood; supervision, M. F. Mridha; project administration, M. F. Mridha; funding acquisition, Mejdl Safran and Sultan Alfarhood. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available in the respective repository as referenced in this study.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Chaudhary D, Kumar S, Dhaka VS. Video based human crowd analysis using machine learning: a survey. Comput Methods Biomech Biomedical Eng Imag Visual. 2022;10(2):113–31.
2. Tripathi G, Singh K, Vishwakarma DK. Convolutional neural networks for crowd behaviour analysis: a survey. Vis Comput. 2019;35(5):753–76. doi:10.1007/s00371-018-1499-5.
3. MarketsandMarkets. Video surveillance industry worth $88.71 billion by 2030; 2025 [Internet]. [cited 2025 Jul 22]. Available from: https://www.marketsandmarkets.com/PressReleases/global-video-surveillance-market.asp.

4. Badidi E, Moumane K, El Ghazi F. Opportunities, applications, and challenges of edge-ai enabled video analytics in smart cities: a systematic review. IEEE Access. 2023;11:80543–72. doi:10.1109/access.2023.3300658.
5. Liu Y, Liu S, Zhu X, Li J, Yang H, Teng L, et al. Privacy-preserving video anomaly detection: a survey. arXiv:2411.14565. 2025.
6. ElBaih M. The role of privacy regulations in AI development (A discussion of the ways in which privacy regulations can shape the development of AI). 2023. [cited 2025 Jul 22]. Available from: http://dx.doi.org/10.2139/ssrn.4589207.
7. Gupta R. Safeguarding digital privacy with ai-driven solutions. ESP Int J Adv Comput Technol (ESP-IJACT). 2024;2(1):126–42.
8. Do TTT, Huynh QT, Kim K, Nguyen VQ. A survey on video big data analytics: architecture, technologies, and open research challenges. Appl Sci. 2025;15(14):8089. doi:10.3390/app15148089.
9. Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: concept and applications. ACM Trans Intell Syst Technol. 2019;10(2):12. doi:10.1145/3298981.
10. Shenaj D, Rizzoli G, Zanuttigh P. Federated learning in computer vision. IEEE Access. 2023;11:94863–84. doi:10.1109/access.2023.3310400.
11. Alsmirat MA, Obaidat I, Jararweh Y, Al-Saleh M. A security framework for cloud-based video surveillance system. Multimedia Tools Appl. 2017;76(21):22787–802. doi:10.1007/s11042-017-4488-1.
12. Shifa A, Asghar MN, Noor S, Gohar N, Fleury M. Lightweight cipher for H.264 videos in the Internet of multimedia things with encryption space ratio diagnostics. Sensors. 2019;19(5):1228. doi:10.3390/s19051228.
13. Bagdasaryan E, Poursaeed O, Shmatikov V. Differential privacy has disparate impact on model accuracy. Adv Neural Inf Process Syst. 2019;32. [cited 2025 Jul 12]. Available from: https://proceedings.neurips.cc/paper/2019/hash/fc0de4e0396fff257ea362983c2dda5a-Abstract.html.
14. Tonmoy MR, Rakib AF, Rahman R, Adnan MA, Mridha MF, Huang J, et al. A lightweight visual font style recognition with quantized convolutional autoencoder. IEEE Open J Comput Soc. 2024;5(2):120–30. doi:10.1109/ojcs.2024.3378709.
15. Ullah FUM, Obaidat MS, Ullah A, Muhammad K, Hijji M, Baik SW. A comprehensive review on vision-based violence detection in surveillance videos. ACM Comput Surv. 2023;55(10):1–44. doi:10.1145/3561971.
16. Liu S, Li Y, Fu W. Human-centered attention-aware networks for action recognition. Int J Intell Syst. 2022;37(12):10968–87. doi:10.1002/int.23029.
17. Aggarwal S, Ranjan R, Sinha M, Pal V, Kushwaha R. CNN and BiLSTM based framework for real life violence detection from CCTV videos. In: 2024 IEEE Region 10 Symposium (TENSYMP); 2024 Sep 27–29; New Delhi, India. p. 1–6.
18. Yadav V, Kumar S, Goyal A, Bhatla S, Sikka G, Kaur A. Integrated violence and weapon detection using deep learning. In: 2024 First International Conference on Pioneering Developments in Computer Science & Digital Technologies (IC2SDT); 2024 Aug 2–4; Delhi, India.
19. Abbass MAB, Kang HS. Violence detection enhancement by involving convolutional block attention modules into various deep learning architectures: comprehensive case study for UBI-fights dataset. IEEE Access. 2023;11:37096–107. doi:10.1109/access.2023.3267409.
20. Kumar A, Shetty A, Sagar A, Charushree A, Kanwal P. Indoor violence detection using lightweight transformer model. In: 2023 4th International Conference for Emerging Technology (INCET); 2023 May 26–28; Belgaum, India. p. 1–6.
21. Mohammadi H, Nazerfard E. Video violence recognition and localization using a semi-supervised hard attention model. Expert Syst Appl. 2023;212:118791. doi:10.1016/j.eswa.2022.118791.
22. Vijeikis R, Raudonis V, Dervinis G. Efficient violence detection in surveillance. Sensors. 2022;22(6):2216. doi:10.3390/s22062216.
23. Frimpong E, Khan T, Michalas A. Secrets in motion: privacy-preserving video classification with built-in access control. In: 2024 9th International Conference on Smart and Sustainable Technologies (SpliTech); 2024 Jun 25–28; Bol and Split, Croatia. p. 1–6.

24. Feng D, Wang L, Chen S, Tung L, Liu F. X-stream: a flexible, adaptive video transformer for privacy-preserving video stream analytics. In: IEEE INFOCOM 2024-IEEE Conference on Computer Communications; 2024 May 20–23; Vancouver, BC, Canada. p. 1–10.
25. Gaikwad B, Karmakar A. Real-time distributed video analytics for privacy-aware person search. Comput Vis Image Underst. 2023;234(3):103749. doi:10.1016/j.cviu.2023.103749.
26. Mehta D, Sivathamboo S, Simpson H, Kwan P, O'Brien T, Ge Z. Privacy-preserving early detection of epileptic seizures in videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham, Switzerland: Springer Nature; 2023. p. 210–9.
27. Singh S, Dewangan S, Krishna GS, Tyagi V, Reddy S, Medi PR. Video vision transformers for violence detection. arXiv:2209.03561. 2022.
28. Pajon Q, Serre S, Wissocq H, Rabaud L, Haidar S, Yaacoub A. Balancing accuracy and training time in federated learning for violence detection in surveillance videos: a study of neural network architectures. J Comput Sci Technol. 2024;39(5):1029–39. doi:10.1007/s11390-024-3702-7.
29. Victor EDS, Lacerda TB, Miranda PB, Nascimento AC, Furtado APC. Federated learning for physical violence detection in videos. In: 2022 International Joint Conference on Neural Networks (IJCNN); 2022 Jul 18–23; Padua, Italy. p. 1–8.
30. McMahan B, Moore E, Ramage D, Hampson S, BAy Arcas. Communication-efficient learning of deep networks from decentralized data. In: Singh A, Zhu J, editors. Proceedings of the 20th international conference on artificial intelligence and statistics. Vol. 54. Westminster, UK: PMLR; 2017. p. 1273–82.
31. Guan H, Yap PT, Bozoki A, Liu M. Federated learning for medical image analysis: a survey. Pattern Recognit. 2024;151(3):110424. doi:10.1016/j.patcog.2024.110424.
32. Mistry D, Tonmoy MR, Anower MS, Hasan ASMT. Federated transfer learning for vision-based fall detection. In: Arefin MS, Kaiser MS, Bhuiyan T, Dey N, Mahmud M, editors. Proceedings of the 2nd International Conference on Big Data, IoT and Machine Learning. Singapore: Springer Nature Singapore; 2024. p. 961–75. doi:10.1007/978-981-99-8937-9_64.
33. Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V. Federated optimization in heterogeneous networks. In: Proceedings of Machine Learning and Systems; 2020 Mar 2–4; Austin, TX, USA. p. 429–50.
34. Tonmoy MR, Shams MA, Adnan MA, Mridha MF, Safran M, Alfarhood S, et al. X-Brain: explainable recognition of brain tumors using robust deep attention CNN. Biomed Signal Process Control. 2025;100(18):106988. doi:10.1016/j.bspc.2024.106988.
35. Hendrycks D, Gimpel K. Gaussian error linear units (GELUs). arXiv:1606.08415. 2023.
36. Mehta S, Rastegari M. Separable self-attention for mobile vision transformers. Trans Mach Learn Res. 2023. [cited 2025 Jul 12]. https://openreview.net/forum?id=tBl4yBEjKi.
37. Bermejo Nievas E, Deniz Suarez O, Bueno García G, Sukthankar R. Violence detection in video using computer vision techniques. In: Real P, Diaz-Pernil D, Molina-Abril H, Berciano A, Kropatsch W, editors. Computer analysis of images and patterns. Berlin/Heidelberg, Germany: Springer; 2011. p. 332–9. doi: 10.1007/978-3-642-23678-5_39.
38. Nadeem MS, Kurugollu F, Atlam HF, Franqueira VNL. Weapon violence dataset 2.0: a synthetic dataset for violence detection. Data Brief. 2024;54(8):110448. doi:10.1016/j.dib.2024.110448.
39. Ruiz-Santaquiteria J, Muñoz JD, Maigler FJ, Deniz O, Bueno G. Firearm-related action recognition and object detection dataset for video surveillance systems. Data Brief. 2024;52(24):110030. doi:10.1016/j.dib.2024.110030.
40. Buslaev A, Iglovikov VI, Khvedchenya E, Parinov A, Druzhinin M, Kalinin AA. Albumentations: fast and flexible image augmentations. Information. 2020;11(2):125. doi:10.3390/info11020125.
41. Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Schmid C. ViViT: a video vision transformer. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 6836–46.
42. Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding?. In: Meila M, Zhang T, editors. Proceedings of the 38th international conference on machine learning. Vol. 139. Westminster, UK: PMLR; 2021. p. 813–24.