

A Lightweight Hybrid CNN-BiLSTM Model with Adaptive Temporal Attention for Violence Detection in Surveillance Videos

AICHA KHALFAOUI¹, ABDELMAJID BADRI¹, ILHAM EL MOURABIT²

¹Laboratory of Materials, Energy, and Smart Systems (MESI), Hassan II University of Casablanca, Casablanca, Morocco

²Laboratory of Engineering Sciences and Biosciences (LSIB), Hassan II University of Casablanca, Casablanca, Morocco

Corresponding author: Aicha Khalfaoui (e-mail: aicha96khalfaoui@gmail.com)

ABSTRACT

Detecting violent actions in surveillance videos is a critical task for ensuring public safety in smart city environments. Although numerous deep learning approaches have been proposed, most of them favor either accuracy or computational efficiency, but rarely achieve both simultaneously. This limitation restricts their deployment on resource-constrained edge devices. In this paper, we propose a lightweight yet effective hybrid architecture that combines MobileNetV3-S for spatial feature extraction with a Bi-LSTM enhanced by an Adaptive Temporal Attention mechanism for temporal modeling. Despite its compact design (1.00 GFLOPs and 3.66 million parameters), the proposed model achieves competitive performance on four public benchmark datasets. The experimental results demonstrate that the proposed approach provides an excellent trade-off between accuracy and efficiency, making it suitable for real-time smart surveillance applications.

KEYWORDS violence detection, edge devices, adaptive temporal attention, MobileNetV3-S, Bi-LSTM.

I. INTRODUCTION

In recent years, violence detection in surveillance videos has attracted increasing attention, especially in the context of smart cities where ensuring public safety is crucial. Intelligent video surveillance systems enable authorities to respond more effectively by identifying violent behavior early. However, achieving this goal in real-world conditions is challenging. Surveillance environments are dynamic and often unpredictable, and edge devices commonly used in Internet of Things (IoT) networks face significant constraints in terms of processing power, memory, and energy. These limitations call for practical solutions that can offer both accuracy and efficiency, particularly for systems that need to operate in real time.

Effective violence detection relies on the analysis of both spatial and temporal information. Spatial features capture the appearance and context within individual frames, while temporal features reflect how motion and actions evolve

over time [1]. Traditional approaches, which often use handcrafted features, are not flexible enough to deal with the complexity of real-life scenes. Deep learning models, especially Convolutional Neural Networks (CNNs), have proven to be more effective in extracting spatial features, but their computational demands are often too high for real-time applications [2]. On the other hand, Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) models, have shown promise in modeling sequential patterns over time [3].

As each offer strong capabilities, they do not fully capture the interplay between spatial and temporal features when used separately. Moreover, deploying such models on edge devices presents practical difficulties due to limited computational resources. Even models that perform well in research settings may not be suitable for lightweight real-time systems. This makes it essential to develop streamlined architectures that deliver solid performance without compro-

missing efficiency.

Attention mechanisms have recently emerged as an effective approach to address this issue, particularly in video classification and action recognition tasks [4]. By helping models focus on the most relevant spatial areas and time steps, attention modules improve accuracy while filtering out noise and redundant information. In violence detection scenarios, these mechanisms are especially useful for identifying key motions and interactions that signal aggression.

To build on these insights, we propose a model that incorporates an Adaptive Temporal Attention (TA) mechanism. This module helps the system identify and prioritize the most meaningful moments in a video sequence, rather than treating all frames as equally important.

For spatial feature extraction, we adopt MobileNetV3s—a compact and efficient CNN designed for deployment on edge devices. To model the temporal progression of actions, we use a Bidirectional LSTM (Bi-LSTM), which allows the system to understand both past and future context in a sequence. Integrating the TA module with this architecture makes the model more attentive to critical motion cues, without adding significant computational cost.

We evaluate our method on four well-known datasets : Movie Fight, Hockey Fight, RWF-2000, and the Surveillance Camera dataset—each presenting a different set of challenges. The results show that our model achieves high accuracy across all benchmarks while maintaining a lightweight structure, making it well-suited for real-time applications in constrained environments.

In summary, our main contributions are:

- We integrate an Adaptive Temporal Attention (ATA) mechanism within a lightweight CNN–BiLSTM framework to enhance temporal feature discrimination in violence detection.
- We design a compact spatiotemporal architecture combining MobileNetV3s and Bi-LSTM, optimized for low-resource edge deployment.
- We conduct comprehensive evaluations on four diverse datasets, showing strong accuracy and generalization, even under varying conditions.

The rest of this paper is organized as follows: Section II reviews the related literature. Section III describes our proposed method. Section IV outlines the experimental setup and presents the results. Section V provides a discussion, and Section VI concludes the study with perspectives for future work.

II. RELATED WORKS

The identification of violence in videos has become an increasingly important area of research, particularly in the context of public safety and crime prevention. Early approaches to violence detection relied heavily on handcrafted features, such as Histogram of Oriented Gradients (HOG) and Histogram of Optical Flow (HOF), which were effective in setting baseline performance but struggled to address the

complexity and diversity of real-world scenarios, including occlusions and dynamic environments [5].

Deep learning has increasingly led researchers to rely on convolutional neural networks (CNNs) and their variants for automatic feature extraction. Tran et al. [6] presented 3D Convolutional Neural Networks (3D-CNNs), which showed the capacity to catch spatiotemporal patterns in video frames, thereby making them particularly suitable for detecting violent behaviors. Similarly, Simonyan and Zisserman [7] presented the Two-Stream CNN model, which aggregated RGB frame inputs with optical flow to increase temporal understanding. These architectures require large processing resources, so even if they are efficient, real-time deployment on resource-limited edge devices is challenging. Hybrid architectures that integrate CNNs with Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, have significantly improved temporal modeling. Donahue et al. [8] combined spatial feature extraction from CNNs with the temporal dependencies captured by LSTMs, achieving state-of-the-art performance on activity recognition benchmarks. Sharma et al. [9] further advanced this approach by proposing a fully integrated violence detection system using CNN and LSTM, which outperformed traditional methods. However, the recurrent nature of LSTMs and their high memory usage remain challenges for real-world deployment.

By allowing models to focus on the most pertinent spatial and temporal areas, attention mechanisms have transformed video analysis. Originally presented in natural language processing by Vaswani et al. [4], self-attention methods have later been modified for video analysis. Using self-attention to capture global dependencies across frames, Wang et al. [10] developed the Non-local Neural Network, so greatly enhancing action detection performance.

When it comes to the identification of violence, attention systems have been especially useful. Whereas temporal attention modules concentrate on significant frames in a sequence, spatial attention modules stress important areas inside video frames. Using attention processes to increase accuracy in dynamic situations, Kang et al. [11] developed effective spatiotemporal modeling techniques for real-time violence detection. However, many existing approaches fail to fully exploit bidirectional motion patterns, which can provide valuable contextual information for detecting violence in complex scenarios.

The development of lightweight architectures that performs efficiently on edge devices has been driven by the increasing need for real-time violence detection in smart cities. One of the most effective CNN designs for real-time applications is MobileNetV3, which was proposed by Howard et al. [2]. It uses squeeze-and-excitation (SE) modules and depthwise separable convolutions to obtain great accuracy with less computing complexity.

Recent studies have explored ways to blend lightweight architectures with attention-based strategies to boost performance. For example, Chen et al. [12] developed a spatial-

temporal graph attention network for detecting video anomalies, demonstrating that lightweight designs can effectively work with attention mechanisms.

In a real-time violence detection system, Sharma et al. [9] also used MobileNetV3 as a feature extractor, and they achieved good performance with minimal computational overhead. However, their approach showed limited performance in complex and cluttered scenes due to the absence of advanced attention mechanisms.

In the subject of violence detection, several difficulties still exist although significant advancements.

- 1) **Inadequate Temporal Modeling:** Many of the methods used don't fully take advantage of temporal dependencies, especially when scenarios become more complex and violent behaviors evolve over time.
- 2) **Suboptimal feature selection:** Traditional CNNs, even lightweight ones like MobileNetV3, often have trouble distinguishing between important and irrelevant features in dynamic environments. Advanced attention mechanisms provide a possible answer to this problem, including adaptive temporal attention.
- 3) **Scalability on Edge Devices :** Transformers and other high-performance models sometimes surpass edge device computing capability. Particularly in Internet of Things-based monitoring systems, architectures that strike a balance between efficiency and strong performance are increasingly in demand.

To tackle these issues, we offer an innovative method that integrates MobileNetV3s with an adaptive temporal attention mechanism and a Bi-LSTM for violence detection. The MobileNetV3s backbone facilitates efficient feature extraction, whereas the TA module enhances the model's capacity to capture spatial-temporal correlations, making it appropriate for real-time deployment on edge devices.

III. METHODS

A. MOBILENETV3S

The MobileNetV3 [2] network outperforms other similar methods. Fig. 1 shows a schematic diagram of the MobileNetV3 network. This lightweight network addresses the gradient issues caused by shortcut connections and enhances the extraction of information from high-dimensional features. It incorporates a lightweight attention model, linear bottleneck inverse residual structures from MobileNetV2, depthwise separable convolutions from MobileNetV1, and replaces the previous swish function with the h-swish function. Due to its computational efficiency, the MobileNetV3 network is widely used in medical segmentation, classification, and detection.

The MobileNetV3 network is integrated with an attention mechanism module known as Squeeze and Excitation (SE). To compress the data, global average pooling is applied first. This is followed by two fully connected layers to complete the excitation process. The first fully connected layer has a vector size that is one-fourth of the number of channels and uses the ReLU activation function.

The output of the second fully connected layer matches the number of channels in the network input and employs the h-swish activation function. After the SE module determines the weight values for each channel, the initial feature vector, with dimensions of $W \times H \times C$, is multiplied by the corresponding two-dimensional weight matrix for each channel.

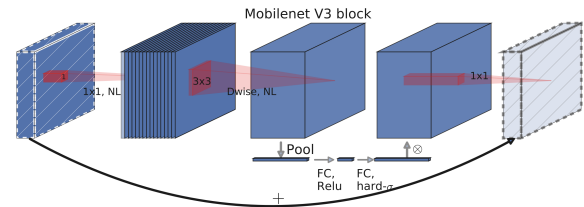


Figure 1. MobileNetV3 Architecture

B. BIDIRECTIONAL LONG-SHORT-TERM MEMORY RECURRENT NEURAL NETWORKS(BI-LSTM)

1) Long Short-Term Memory (LSTM)

In traditional Recurrent Neural Networks (RNNs), the flow of gradient information during backpropagation can either diminish or surge as the network gets deeper. This problem hampers the optimization process and lowers learning efficiency, rendering RNNs inadequate for managing long time series data [13].

LSTM networks, a specialized type of RNN, tackle these challenges with a sophisticated cell structure that includes memory mechanisms. Each LSTM cell is made up of three main components: the forget gate, the input gate, and the output gate. These elements allow the network to selectively keep or discard information, helping to alleviate the issues of vanishing or exploding gradients. By incorporating direct connections between memory states, LSTM networks effectively facilitate the learning of long-term dependencies, making it especially effective for handling sequential data. The fundamental structure of an LSTM cell is shown in Fig. 2.

Unlike traditional RNNs, which pass the cell state directly from one timestep to the next, LSTM employs a more adaptable mechanism governed by three gates. During training, the forget gate f_t , input gate i_t , and output gate o_t work together to regulate the flow of information within the cell. The operation of an LSTM unit involves several essential steps:

Each gate in the cell takes three inputs: the hidden state from the previous timestep (h_{t-1}), the current input (x_t), and the internal memory state (c_{t-1}). These inputs are transformed through nonlinear functions to manage what information to keep or discard.

The forget gate produces a value between 0 and 1, which indicates how much of the previous memory state (c_{t-1}) should be retained in the current state (c_t). The updated memory state is shaped by both the forget gate and the input

gate, which chooses relevant information from the new input to add to c_t . Lastly, the output gate decides how the hidden state (h_t) is activated, often using a nonlinear function like $\tanh$.

The LSTM operations can be mathematically expressed as follows:

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + W_{fc}C_{t-1} + b_f) \quad (1)$$

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + W_{ic}C_{t-1} + b_i) \quad (2)$$

$$C_t = f_t * C_{t-1} + i_t * \tanh(W_{cx}x_t + W_{ch}h_{t-1} + b_c) \quad (3)$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + W_{oc}C_t + b_o) \quad (4)$$

$$h_t = o_t * \tanh(C_t) \quad (5)$$

In these equations, $\sigma(\cdot)$ represents the sigmoid activation function, and $\tanh(\cdot)$ is the hyperbolic tangent activation function. The weight matrices $\mathbf{W}$ and bias vectors $\mathbf{b}$ are parameters that the network learns during training to enhance its performance. Together, these components allow LSTM networks to effectively handle complex sequential dependencies over extended time periods.

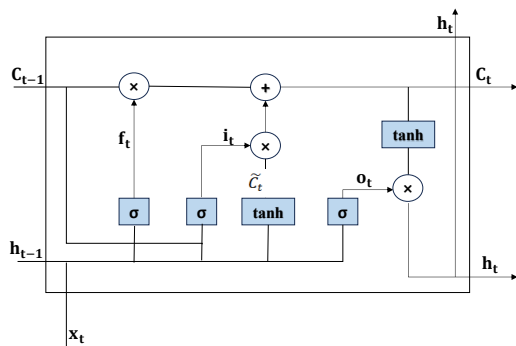


Figure 2. The LSTM Architecture

2) Bi-LSTM

LSTM networks are known for their ability to effectively transmit and process sequential information in a forward direction. However, this unidirectional approach restricts their capacity to leverage future information within a sequence, especially when it comes to reverse dependencies in a signal. To address this issue, the Bidirectional LSTM (Bi-LSTM) architecture introduces an extra LSTM layer that processes the sequence in reverse. This improvement enables the network to capture dependencies from both the past and the future, allowing it to learn sequential patterns in both directions at the same time.

As shown in Fig. 3, a Bi-LSTM is made up of two recurrent neural networks that function in opposite directions: the forward LSTM focuses on extracting long-term dependencies from past inputs, while the backward LSTM identifies future dependencies. Each neuron in the network is linked to two LSTM units, one for processing in the forward direction and one for the backward direction, leading to two hidden states, h_t and h'_t , that correspond to each direction.

At each time step t , the overall hidden state of the Bi-LSTM is represented as $H_t = [h_t, h'_t]$. In the proposed model, this hidden state is passed into a fully connected layer to carry out a nonlinear transformation of the extracted features. Ultimately, a fully connected classification layer is used to predict violent or non-violent behavior.

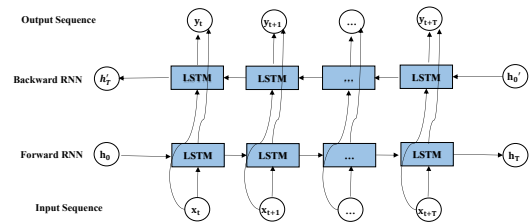


Figure 3. The Bi-LSTM Architecture

C. ADAPTIVE TEMPORAL ATTENTION (TA)

In the task of video-based violence detection, it is essential to determine if an individual has engaged in violent conduct across a sequence of consecutive acts, necessitating the model's ability to acquire adequate temporal information. Furthermore, video sequences are typically lengthy, and Long Short-Term Memory (LSTM) networks may experience a loss of temporal information in extended sequences. To resolve this issue, we employ the Temporal Adaptive (TA) block introduced by Liu et al. [14].

The convolution kernel produced by the TA module can flexibly modify its dimensions according to varying scenarios in the video frames. This improves the model's capacity to discern temporal relationships, thus enhancing the accuracy and precision in detecting violent behavior. The comprehensive structure of the TA block can be seen in Fig. 5.

Let $X \in \mathbb{R}^{C \times T \times H \times W}$ signify the feature map of a video frame, where C , T , H , and W represent the number of channels, time steps, height, and width, respectively. The TA block exclusively emphasizes temporal information, necessitating the application of global spatial average pooling to reduce the spatial dimensions. Consequently, the feature map is represented as $X \in \mathbb{R}^{C \times T}$.

A one-dimensional convolution operation is subsequently executed on X to extract temporal information from the video frames. A fully connected layer is utilized simultaneously with the global background information. In the end, the softmax activation function is employed to produce the final dynamic convolution kernel.

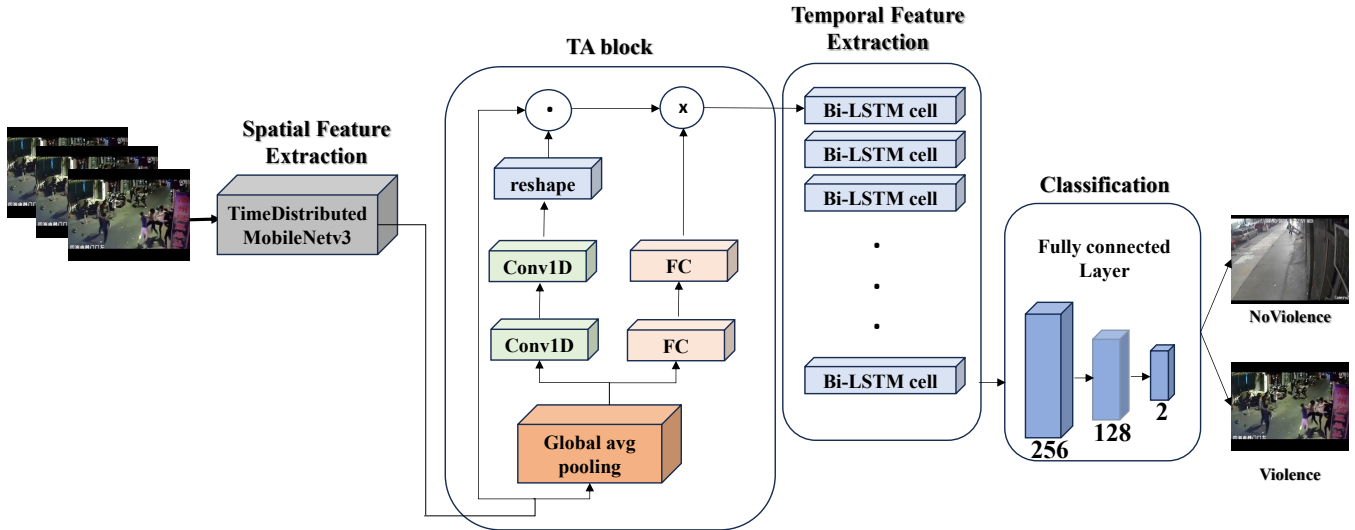


Figure 4. Model Architecture

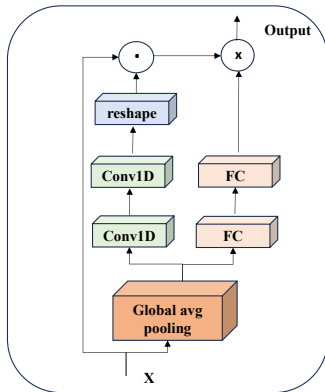


Figure 5. The TA Block

The proposed system's architecture is displayed in Fig. 4, offering a detailed overview of the essential components and their interconnections.

IV. MATERIALS AND EXPERIMENTAL RESULTS

A. DATASET AND EVALUATION METRICS

To better understand how our model performs in different scenarios, we describe the benchmark datasets used during experimentation, each offering distinct challenges and characteristics related to violence detection.

Movie Fight Dataset

The Movie Fight dataset consists of several one- to two-second video clips from various films. It has 200 clips, equally partitioned among 100 violent and 100 non-violent samples. All clips have a resolution of 360×240 pixels. We allocated 80% of the dataset for training and the remaining 20% for testing.

Hockey Fight Dataset

The Hockey Fight dataset comprises one thousand action

clips collected from National Hockey League games, designed to identify aggressive behavior. Half of the clips in the dataset show peaceful activities, whereas the other half show violent ones. The writers have carefully annotated each of the 40 frames in each clip, which has a resolution of 360×288 pixels. It should be emphasized that the Hockey Fight dataset does not include a diversity of backgrounds since it is restricted to violence in one specific environment.

Surveillance Camera Dataset

The Surveillance Camera dataset is composed of 300 two-second video clips sourced from YouTube or obtained via surveillance cameras. The dataset contains 150 violent and 150 non-violent clips which differ in size and frame count. Given the limited sample size of the dataset, we implemented a five-fold cross-validation strategy during our testing.

RWF-2000 Dataset

With its extensive coverage of violent and non-violent actions in a wide range of contexts, the RWF-2000 dataset is now the most comprehensive resource for violence detection. It comprises a collection of 2000 video clips, each containing 150 frames, recorded by surveillance cameras in actual environments. By splitting the dataset in half, the authors have made good use of both the training and testing sets. With its more realistic representation of violence and more obstacles, the RWF-2000 dataset stands out among the Movie Fight and Hockey Fight datasets.

To ensure a reliable evaluation protocol, the dataset was randomly partitioned into 70% training and 30% validation subsets. The split was performed at the video level to avoid data leakage between subsets. Class distribution was preserved in both partitions to maintain dataset balance. The training set was used to optimize model parameters, while the validation set was used for performance assessment and

early stopping.

Representative sample images from all datasets are displayed in Fig. 6, and the details of the databases used are shown in Table 1.

Table 1. Details of the databases used

Dataset	Context	Videos	Resolution
Surveillance [15]	Surveillance	150	Varying
Movie Fight [16]	Movies	200	360 × 250
Hockey Fight [16]	Hockey Games	1000	720 × 576
RWF-2000 [17]	Surveillance	2000	Varying



(a)



(b)



(c)



(d)

Figure 6. Sample images from the Hockey Fight dataset (a), Movie Fight dataset (b), RWF-2000 dataset (c), and the Surveillance dataset (d).

In our study on violence detection using the Bi-directional Motion Attention and MobileNetV3-LSTM framework, evaluating the model's performance is essential, and metrics like precision, recall, accuracy, and F1-score play a pivotal role in this process. Precision reflects how well the model identifies violent instances without misclassifying non-violent ones, focusing on reducing false positives (FP). Recall, conversely, measures the model's capability to capture all actual violent instances, minimizing missed detections or false negatives (FN). Accuracy gives a broad view of the model's overall correctness by considering both correctly predicted violent and non-violent instances. The F1 score, which harmonizes precision and recall, is especially valuable in scenarios with imbalanced data, as it balances the trade-off between identifying true positives and avoiding errors.

These metrics are mathematically defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (6)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (7)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

where TP (True Positives) represent correctly identified violent instances, FP (False Positives) are non-violent instances wrongly classified as violent, TN (True Negatives) are correctly identified non-violent instances, and FN (False Negatives) are violent instances incorrectly labeled as non-violent. Together, these metrics provide a robust framework for evaluating the effectiveness of our violence detection model, ensuring a thorough understanding of its strengths and limitations.

B. IMPLEMENTATION DETAILS

In developing this model, we carefully chose hyperparameters to achieve a balance between performance and computational efficiency. To address memory constraints, we configured the batch size to 8 and trained the model for 100 epochs to guarantee comprehensive learning. The learning rate was set to 0.001, utilizing the AdamW optimizer to adjust the weights, with a weight decay of 1×10^{-5} implemented to reduce overfitting. To avert redundant training, we adopted an early stopping mechanism with a patience of 10 epochs, which stops training if the validation accuracy does not enhance. To optimize processing, we adjusted the input images to a resolution of 112×112 to reduce computational complexity and memory consumption, which are critical constraints in edge-based surveillance systems, and limited the number of frames per video to 10 to maintain consistency across inputs.

We used random horizontal flipping and color (adjusting brightness, contrast, and saturation) to improve the generalizing capacity of the model. Designed to run bidirectionally and so capture temporal dependencies in the data, the LSTM layer has a hidden size of 64. We also incorporated a ReduceLROnPlateau scheduler, which reduces the learning rate by a factor of 0.5 if the validation accuracy plateaus for 5 consecutive epochs.

To further optimize training, we enabled mixed precision training using autocast and GradScaler, which not only speeds up the process but also reduces memory usage. All experiments were conducted on Google Colab, leveraging the powerful Tesla A100 GPU to handle the computational demands of the training process efficiently. Together, these hyperparameter values taken guarantee that the model can provide strong performance for the purpose of violence detection and be computationally efficient.

Table 2. Comparison of RNN-based CNN models for violence detection.

Model	Accuracy (%)	Parameters (M)
VGG16 + Bi-LSTM	94.50	20.39
ResNet18 + Bi-LSTM	80.50	11.55
Xception + Bi-LSTM	94.00	22.01
InceptionV3 + Bi-LSTM	93.99	22.95
DenseNet121 + Bi-LSTM	96.49	7.66
EfficientNet-B0 + Bi-LSTM	88.00	4.80
MobileNetV3-Small + Bi-LSTM	95.50	1.33

Table 3. Comparison of different CNN-based hybrid models for violence detection.

Model	Accuracy (%)	Parameters (M)
MobileNetV3s + LSTM	95.00	1,153,138
MobileNetV3s + Bi-LSTM	95.50	1,333,746
MobileNetV3s + ConvLSTM	93.50	2,709,618
MobileNetV3s + Bi-GRU	95.00	1,252,082

C. MAIN RESULTS

From the comparative observation in Table 2, it can be noted that the MobileNetV3s + Bi-LSTM architecture has the optimal tradeoff between parameters and performance. Even though DenseNet121 + Bi-LSTM achieved the highest accuracy of 96.49%, with MobileNetV3s + Bi-LSTM achieving 95.50% accuracy, the former model had a parameter count of 1,333,746. This is over five times more than the DenseNet121 was. The drastic reduction in model size with minimal compromise in accuracy makes MobileNetV3s + Bi-LSTM the most effective for resource-limited settings or real-time embedded systems. Its lightweight and strong predictive performance justifies it as the best architecture in our case.

The comparison of various hybrid architectures shown in Table 3 indicates that MobileNetV3s + Bi-LSTM is the best performing hybrid having an accuracy of 95.50 and a parameter count of 1,333,746 which is relatively low when compared to other variants. While MobileNetV3s + ConLSTM was also observed to perform worse with a higher parameter count of 2,709,618, accuracy stood at 93.50 suggesting that it is inefficient in terms of a parameter to accuracy ratio. However, MobileNetV3s + Bi-GRU and MobileNetV3s + LSTM performed equally in accuracy 95.00%, which is lower than the result of Bi-LSTM. In this regard, it can be concluded that MobileNetV3s + Bi-LSTM along with its classification performance is more complex than other models making it the best fit for our case.

The given Table 4 outlines the performance measures (precision, recall, F1-score, and accuracy) of the violence detection model on four datasets: Movie Fight, Hockey-Fight, Surveillance, and RWF-2000. The model achieved near-perfect classification on the Movie Fight dataset, reaching 100% on all metrics, which implies that distinguishing violence features in this context is quite easy. Performance remains strong on Hockey-Fight as well, with all metrics exceeding 98%, indicating that the model performs well in

sports scenes too. A slight drop is noted on Surveillance and RWF-2000, more specifically in NonViolence precision and recall on the Surveillance dataset, most probably due to changing illumination, viewing angles, and occlusions present in real-world footage. Regardless, the model maintains strong (>0.89) F1-scores and ($>90\%$) accuracy on all datasets, proving that the model is robust to varying conditions. Including the standard deviations helps show the consistency of results, as portrayed within the Movie Fight, Hockey-Fight, and Surveillance datasets, where there is less variability in controlled environments compared to more challenging ones. These results demonstrate that the model is adaptable while highlighting the issues that arise from using real-world surveillance and crowd-sourced video data.

The normalized confusion matrices are shown in Fig. 7 for the four datasets, *Hockey Fight*, *Movies*, *RWF-2000*, and *Surveillance*, and provide key insights into the performance of the violence detection model across diverse video domains. Starting with the *Movies* dataset, the model demonstrates perfect classification accuracy, with both “NonViolence” classes achieving 100% recognition. This exceptional performance likely stems from the high-quality, well-structured, and clearly segmented visual content typical of cinematic productions, where violent actions are often exaggerated and well-defined, making them easier to detect by learning-based models.

In the case of the *Hockey Fight* dataset, the model also performs remarkably well, with 99% accuracy for “NonViolence” and 98% for “Violence”. These results suggest that although the dataset features fast-paced, action-intensive scenes, the physical interactions characteristic of hockey fights are distinct enough to allow effective discrimination between violent and non-violent instances. The slight drop compared to the *Movies* dataset may reflect the chaotic and overlapping nature of real-time sports footage.

For the *RWF-2000* dataset, which consists of real-world surveillance videos, the model shows decent performance with a 90% accuracy for “NonViolence” and 92% for “Violence”. However, misclassification rates of 10% and 8% respectively indicate the increasing difficulty in distinguishing violent acts in more complex, real-world environments.

Surveillance videos often involve occlusions, varying lighting conditions, camera motion, and subtle actions that can blur the line between aggressive and non-aggressive behavior, thus challenging even robust classifiers. Finally, the *Surveillance* dataset poses the greatest challenge, with 13% of “NonViolence” samples incorrectly predicted as “Violence” and 6% vice versa. While the model still performs reasonably well (87% and 94% accuracy respectively), this confusion suggests the presence of ambiguous or borderline scenes that complicate decision boundaries. The results underscore the difficulty of real-world deployment, where environmental noise, diverse human behavior, and camera perspectives significantly impact detection accuracy.

For the four datasets : Hockey, Movie Fight, RWF-

Table 4. Performance metrics for violence detection across different datasets. Values are presented as mean $\pm$ standard deviation.

Dataset	Precision (%)		Recall (%)		F1-score (%)		Accuracy (%)
	Violence	Non-Violence	Violence	Non-Violence	Violence	Non-Violence	
Movie Fight	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0
Hockey Fight	98.0 $\pm$ 2.0	98.0 $\pm$ 2.0	98.0 $\pm$ 2.0	98.0 $\pm$ 2.0	98.0 $\pm$ 2.0	98.0 $\pm$ 2.0	98.0 $\pm$ 2.0
Surveillance	88.0 $\pm$ 11.0	92.0 $\pm$ 10.0	93.0 $\pm$ 7.0	86.0 $\pm$ 14.0	91.0 $\pm$ 9.0	89.0 $\pm$ 12.0	90.0 $\pm$ 10.0
RWF-2000	90.0 $\pm$ 8.0	92.0 $\pm$ 5.0	92.0 $\pm$ 4.0	90.0 $\pm$ 8.0	91.0 $\pm$ 6.0	91.0 $\pm$ 6.0	91.0 $\pm$ 6.0

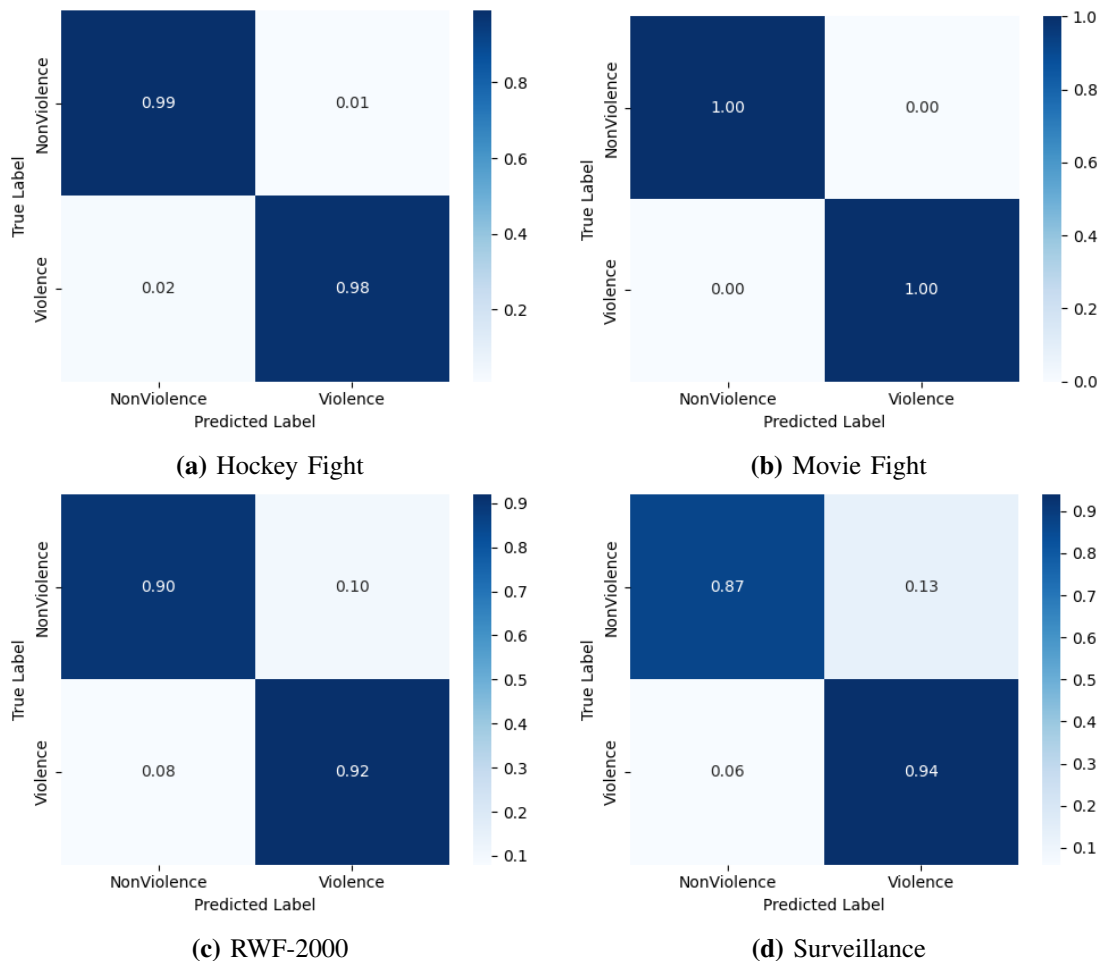


Figure 7. Normalized confusion matrices on the four datasets used.

2000 and Surveillance, the ROC curves (Receiver Operating Characteristic) shown in Fig. 8 represent an overall view of the classification performance across several domains in violence detection tasks. With both “Violence” and “Non-violence” classes obtaining an Area Under the Curve (AUC) of 0.99, suggesting near-perfect discrimination with minimal false positives and false negatives, the Hockey dataset shows an extraordinary classification ability. Likewise, the Movie Fight dataset shows optimal performance with an AUC of 1.00 for both classes, implying that in this controlled, possibly less noisy dataset, the classifier precisely separates the two categories. With an AUC of 0.94 for both “fight” and “no_fight,” the RWF-2000 dataset known for its real-world

complexity presents somewhat lower but still respectable performance. However, this shows a consistent classification with considerable degradation due to scene complexity, lighting changes, and video quality. Lastly, despite different settings and camera angles, the Surveillance dataset, which is the most difficult, achieves a decent AUC of 0.93 for both “Violence” and “Non-violence.” This implies that although the model has good performance, scene classification in complicated, real-world surveillance circumstances is somewhat more challenging and uncertain. From carefully selected to real-world environments, the ROC analyses collectively confirm the resilience of the proposed violence detection methodology throughout a range of datasets.

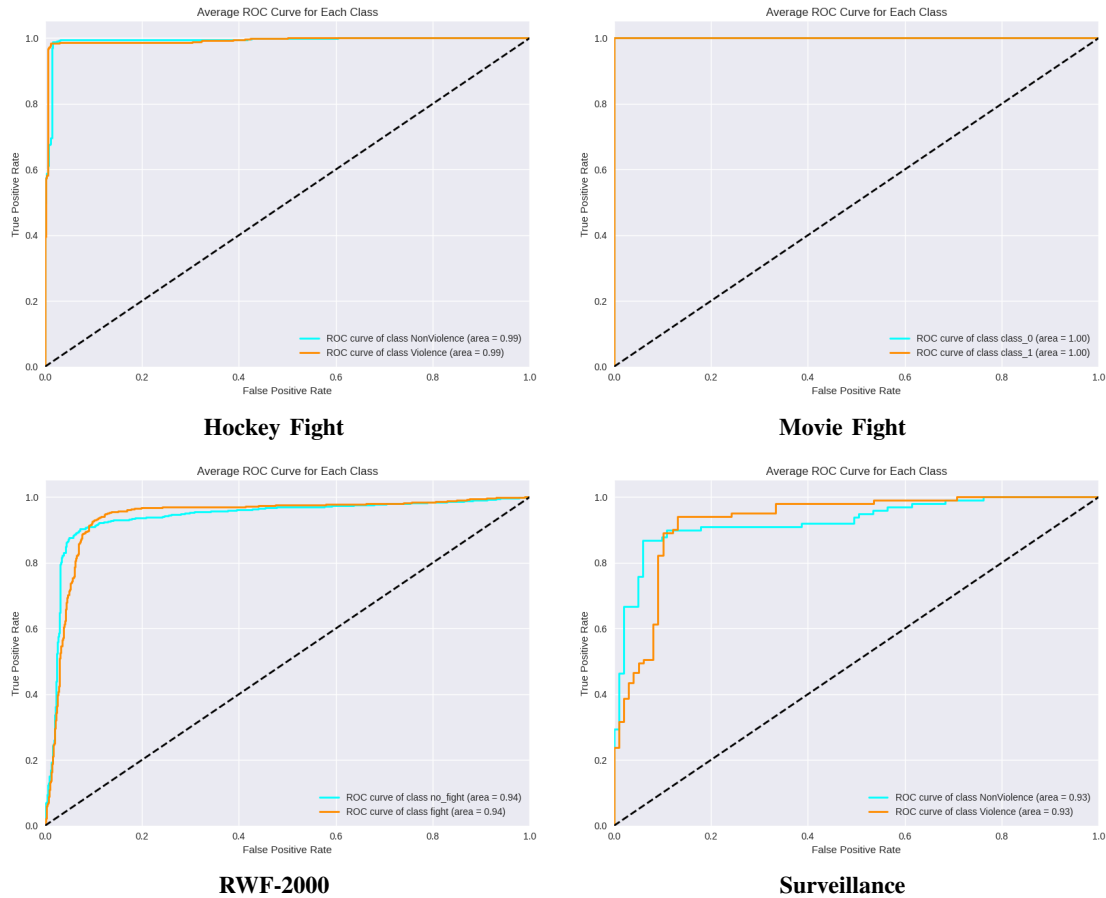


Figure 8. ROC curves and the AUC values for the four datasets used.

The impact of the TA mechanism on the accuracy of the MobileNetV3 + Bi-LSTM model is evident from the ablation study presented in Table 5. The base model without TA achieves a high accuracy of 95.50%, while maintaining efficiency with 1.33M parameters, 0.90 GFLOPs, and a model size of 5.09 MB. When the TA mechanism is introduced, there is a remarkable increase in accuracy to 98.60%, albeit with a significant rise in model complexity. Specifically, TA increases the parameter count to 3.66M, raises the computational demand slightly to 1.00 GFLOPs, and expands the memory usage to 13.95 MB. These findings suggest that while TA enhances the model's ability to learn precise temporal features, it also challenges its applicability in resource-constrained environments. Therefore, model design must carefully balance accuracy and computational efficiency, especially for deployment in embedded or real-time systems.

Table 6 presents a comparison of violence detection model accuracies on multiple datasets (Movie Fight, Hockey Fight, Surveillance, and RWF-2000), along with their computational costs in terms of GFLOPs and number of parameters. As previously mentioned, our proposed model achieves the highest or near-highest accuracy across all datasets, 100%, 98.60%, 90.50%, and 91.30%, respectively

while maintaining a low computational cost (1.00 GFLOPs) and a moderate number of parameters (3.66M). This demonstrates a better trade-off between performance and efficiency compared to other models such as I3D, C3D, and SlowFast, which require significantly more computational resources. The inclusion of GFLOPs and parameter counts further highlights the practicality of our model for deployment in resource-constrained environments.

V. DISCUSSION

The proposed lightweight architecture, combining MobileNetV3s, Bi-LSTM, and an Adaptive Temporal Attention (ATA) mechanism, demonstrates that it is possible to achieve high accuracy in violence detection while maintaining a low computational cost. This characteristic makes the model particularly well-suited for real-time surveillance applications on edge devices. The ATA module significantly enhances the model's ability to focus on the most relevant temporal segments, enabling it to better capture subtle yet critical temporal patterns associated with violent actions.

Our model achieved notable performance across four different datasets: Movie Fight, Hockey Fight, Surveillance Camera, and RWF-2000, with accuracies of 100%, 98.60%, 90.50%, and 91.30%, respectively. These results demon-

Table 5. Performance and complexity comparison of MobileNetV3-based architectures for violence detection.

Model	Accuracy (%)	Parameters (M)	GFLOPs	Size (MB)
MobileNetV3 + Bi-LSTM	95.50	1.33	0.90	5.09
MobileNetV3 + TA + Bi-LSTM	98.60	3.66	1.00	13.95

Table 6. Comparison of existing violence detection models on multiple datasets with computational complexity.

Model	Movie Fight	Hockey Fight	Surveillance	RWF-2000	GFLOPs	Parameters (M)
ConvLSTM [18]	100%	97.10%	–	77.00%	15.69	9.76
ConvBiLSTM [19]	99.66%	98.78%	–	79.10%	24.52	11.20
FightCNN-based Attention [11]	99.50%	96.10%	71.00%	82.00%	–	4.07
ViT Large-16 [20]	99.50%	98.00%	84.60%	–	–	–
Two-cascade TSM [21]	–	97.50%	–	87.75%	–	8.49
R(2+1)D-18 [22]	100%	95.00%	–	81.25%	4.42	33.20
R3D-50 [23]	100%	96.00%	85.21%	80.90%	39.98	46.20
VD-Net [24]	99.00%	98.50%	92.50%	–	15.30	49.28
Flow Gated Network [17]	100%	98.00%	–	84.50%	17.20	0.25
SlowFast [25]	100%	97.50%	–	85.00%	23.89	33.79
I3D (RGB) [26]	100%	98.50%	84.67%	85.75%	111.33	12.29
C3D [6]	100%	96.00%	–	82.75%	154.19	78.00
Our Model	100%	98.60%	90.50%	91.30%	1.00	3.66

strate the generalizability and robustness of the proposed architecture under various lighting conditions, camera angles, and scene complexities. With just 1.00 GFLOPs and 3.66 million parameters, the proposed hybrid model not only ensures computational efficiency but also improves classification performance compared to baseline methods, making it well-suited for deployment in resource-constrained environments.

Furthermore, while Bi-LSTM captures long-term dependencies in the temporal domain which is essential for analyzing motion patterns in video sequences, the use of MobileNetV3s enables efficient spatial feature extraction. The ATA mechanism addresses a common limitation in prior works that treat all temporal frames equally, potentially diluting the impact of key moments.

These results suggest that combining lightweight CNNs with attention-enhanced temporal modeling presents a promising approach for scalable and deployable violence detection systems. However, some challenges remain, such as false positives in crowded or fast-moving scenes, which require further investigation.

VI. CONCLUSION

This work presented a lightweight and efficient framework for real-time violence detection in surveillance videos, explicitly designed for edge computing scenarios. By integrating MobileNetV3s with a Bi-LSTM enhanced by an Adaptive Temporal Attention (ATA) mechanism, the proposed architecture achieves a compelling balance between detection accuracy and computational efficiency.

Extensive experiments on four benchmark datasets validate the robustness of the approach, with a reduced complexity of 1.00 GFLOPs and 3.66 million parameters. When evaluated on an NVIDIA Tesla A100 GPU, the model operates between 104.93 and 127.88 FPS, significantly

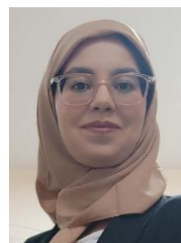
surpassing the 25–30 FPS threshold required for real-time video surveillance. The ablation study further confirms the critical role of the ATA module in strengthening temporal feature modeling.

While occasional false positives may arise in highly dynamic or crowded environments and performance may be affected under challenging illumination conditions, the compact design and high inference speed make the framework well-suited for deployment in practical smart surveillance systems. Future research will investigate advanced temporal architectures and multi-modal fusion strategies to enhance robustness in complex real-world settings.

References

- [1] H. M. B. Jahlan and L. A. Elrefaei, “Detecting violence in video based on deep features fusion technique,” arXiv preprint arXiv:2204.07443, 2022.
- [2] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan et al., “Searching for mobilenetv3,” in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1314–1324.
- [3] A. Traoré and M. A. Akhloufi, “Violence detection in videos using deep recurrent and convolutional neural networks,” in 2020 IEEE international conference on systems, man, and cybernetics (SMC). IEEE, 2020, pp. 154–159.
- [4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems, vol. 30, 2017.
- [5] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05), vol. 1. Ieee, 2005, pp. 886–893.
- [6] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497.
- [7] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” Advances in neural information processing systems, vol. 27, 2014.
- [8] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2625–2634.

- [9] S. Sharma, B. Sudharsan, S. Narahariseti, V. Trehan, and K. Jayavel, "A fully integrated violence detection system using cnn and lstm," *International Journal of Electrical & Computer Engineering* (2088-8708), vol. 11, no. 4, 2021.
- [10] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.
- [11] M.-S. Kang, R.-H. Park, and H.-M. Park, "Efficient spatio-temporal modeling methods for real-time violence recognition," *IEEE Access*, vol. 9, pp. 76 270–76 285, 2021.
- [12] H. Chen, X. Mei, Z. Ma, X. Wu, and Y. Wei, "Spatial-temporal graph attention network for video anomaly detection," *Image and Vision Computing*, vol. 131, p. 104629, 2023.
- [13] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.
- [14] Z. Liu, L. Wang, W. Wu, C. Qian, and T. L. Tam, "Temporal adaptive module for video recognition. in 2021 ieee," in *CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 13 688–13 698.
- [15] Ş. Aktı, G. A. Tataroğlu, and H. K. Ekenel, "Vision-based fight detection from surveillance cameras," in *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, 2019, pp. 1–6.
- [16] E. Bermejo Nieves, O. Deniz Suarez, G. Bueno García, and R. Sukthankar, "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*. Springer, 2011, pp. 332–339.
- [17] M. Cheng, K. Cai, and M. Li, "Rwf-2000: An open large scale video database for violence detection," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 4183–4190.
- [18] S. Sudhakaran 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)*. IEEE, 2017, pp. 1–6.
- [19] R. Halder and R. Chatterjee, "Cnn-bilstm model for violence detection in smart surveillance," *SN Computer science*, vol. 1, no. 4, p. 201, 2020.
- [20] L. Ciampi, C. Santiago, J. P. Costeira, F. Falchi, C. Gennaro, and G. Amato, "Unsupervised domain adaptation for video violence detection in the wild," in *IMPROVE*, 2023, pp. 37–46.
- [21] Q. Liang, Y. Li, B. Chen, and K. Yang, "Violence behavior recognition of two-cascade temporal shift module with attention mechanism," *Journal of Electronic Imaging*, vol. 30, no. 4, pp. 043 009–043 009, 2021.
- [22] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.
- [23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [24] M. Khan, A. El Saddik, W. Gueaieb, G. De Masi, and F. Karray, "Vd-net: An edge vision-based surveillance system for violence detection," *IEEE Access*, vol. 12, pp. 43 796–43 808, 2024.
- [25] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.
- [26] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.



AICHA KHALFAOUI She was born in Errachidia, Morocco, on February 28, 1996. She received the M.Sc. degree in Industrial Computing and Instrumentation Engineering from the Faculty of Sciences and Techniques of Errachidia, Morocco, and the Ph.D. degree in Artificial Intelligence and Embedded Systems from the Faculty of Sciences and Technology of Mohammedia, Hassan II University of Casablanca, Morocco.

She is currently a researcher at the Laboratory of Materials, Energy, and Smart Systems (LMESI), Hassan II University of Casablanca, Mohammedia, Morocco.

Her research interests include artificial intelligence, deep learning, computer vision, image processing, embedded systems, model compression, and intelligent real-time surveillance systems.



ABDELMAJID BADRI is holder of a doctorate in Electronics and Image Processing in 1992 at the University of Poitiers-France. In 1996, he obtained the diploma of the authorization to Manage Researches (HDR) at the University of Poitiers-France, image processing. He was a University Professor (PES-C) at the University Hassan II Mohammedia-Casablanca Morocco (FSTM). He is a member of the laboratory of Materials, Energy, and Smart Systems (LMESI) . He managed

several doctoral theses. He is a co author of several national and international publications.



ILHAM EL MOURABIT holder of a doctoral degree in Electronics and Telecommunication Systems from Hassan II University. She received her M.Sc. degree in Electronic and Automatic Systems Engineering (Telecommunication and Information Technologies specialty) from the Faculty of Sciences and Technology of Mohammedia, Morocco. She is currently working as a Professor at the FSTM. She is a member of the laboratory of Engineering Sciences and Biosciences (LSIB),

at Hassan II University Casablanca. Her main research areas include geolocation technologies in wireless networks, image processing, computer vision, digital signal processing, and vehicular communications.

...