

RESEARCH ARTICLE

Tuna Swarm Algorithm With Deep Learning Enabled Violence Detection in Smart Video Surveillance Systems

GHADAH ALDEHIM¹, MASHAEL M ASIRI², MOHAMMED ALJEBREEN³,
ABDULLAH MOHAMED⁴, MOHAMMED ASSIRI⁵, AND SARA SAADELDEEN IBRAHIM⁶

¹Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia

²Department of Computer Science, College of Science and Art at Mahayil, King Khalid University, Abha 61421, Saudi Arabia

³Department of Computer Science, Community College, King Saud University, P.O. Box 28095, Riyadh 11437, Saudi Arabia

⁴Research Centre, Future University in Egypt, New Cairo 11845, Egypt

⁵Department of Computer Science, College of Sciences and Humanities-Aflaj, Prince Sattam bin Abdulaziz University, Aflaj 16273, Saudi Arabia

⁶Department of Computer and Self Development, Preparatory Year Deanship, Prince Sattam bin Abdulaziz University, Al-Kharj 16278, Saudi Arabia

Corresponding author: Mohammed Assiri (m.assiri@psau.edu.sa)


The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through Large Groups Project under grant number (RGP2/65/44). Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R387), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. Research Supporting Project number (RSP2023R459), King Saud University, Riyadh, Saudi Arabia. This study is supported via funding from Prince Sattam bin Abdulaziz University project number (PSAU/2023/R/1444). This study is partially funded by the Future University in Egypt (FUE).

ABSTRACT In smart video surveillance systems, violence detection becomes challenging to ensure public safety and security. With the proliferation of surveillance cameras in public areas, there is an increasing need for automated algorithms that can accurately and efficiently detect violent behavior in real time. This article presents a Tuna Swarm Optimization with Deep Learning Enabled Violence Detection (TSODL-VD) technique to classify violent actions in surveillance videos. The TSODL-VD technique enables the recognition of violence and can be a measure to avoid chaotic situations. In the presented TSODL-VD technique, the residual-DenseNet model is applied for feature vector generation from the input video frames and then passed into the stacked autoencoder (SAE) classifier. The SAE model is enforced to recognize the events into violence and non-violence events. To improve the violence detection effectiveness of the TSODL-VD procedure, the TSO protocol is utilized as a hyperparameter optimizer for the residual-DenseNet model. The performance validation of the TSODL-VD procedure has experimented on a benchmark violence dataset. The experimental results demonstrate that the TSODL-VD technique accomplishes precise and rapid detection outcomes over the recent state-of-the-art approaches.

INDEX TERMS Violence detection, surveillance videos, public safety, deep learning, tuna swarm algorithm.

I. INTRODUCTION

The Smart City method is a hopeful resolution to the issues relevant to advanced urbanization. Its execution relies upon the ability to analyze and gather huge volumes of many live urban data [1]. It is gathered from private and public sensor networks run by several private bodies or agencies. Compared to other data kinds, video streams particularly offer valuable

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

data gathered straightly from the streets [2]. Smart city scrutiny includes a broad range of applications, including violence detection, urban traffic monitoring systems, disaster management and building structural damage detection [3]. Human operators may be simply overwhelmed with many video streams. So, a significant study was directed to develop techniques for the automatic processing of such to observe abnormal performance and to discard irrelevant data safely. Violence identification indicates a significant problem in smart city surveillance. Violence is a serious social problem.

There exist various reasons for the increase in violent actions in public places. Hatred is a Person's greed, frustration, and, along with that, economic and social insecurity [4]. Violence detection from surveillance videos was a type of activity detection. Many approaches and methodologies were advanced to find the other harmful patterns and brutally events in videos. In these procedures, various methods are modelled that operate with distinct input parameters [5]. The parameters were fundamentally different attributes of video, such as appearance, several accelerations, flow, and duration.

Despite being an alarming social problem, there are only a few various works indulged in the automation of violence recognition, action recognition, and protest recognition [6]; this domain of study has enormous applicability until social stability and security are concerned. Preventing violent activities and crime is impossible until brain signals can be detected and analyzed by the paradigm manifested in criminal thoughts in real time [7]. However, one can recognize aggressive actions in public places utilizing deep due because of learning-based computer vision. Cameras for surveillance were placed in private associations and public areas [8]. The potential violent identification method can aid the authorities or government in considering a fast and formalized method for finding the fierceness approach to thwart the devastation made to public property and human life, as everyone wants secure streets, areas, and work surrounding us [9]. Deep learning (DL) was superior to the machine learning (ML) method since it not required any feature engineering. There exist certain disadvantages, like large training datasets and high computing costs [10]. The technical aspects inspire us to advance a method that accomplishes and acquires lesser training periods and a modest number of training trials.

This article presents a Tuna Swarm Optimization with Deep Learning Enabled Violence Detection (TSODL-VD) procedure. The TSODL-VD technique enables to recognition of violence and can be a measure to avoid any chaotic situations. In the presented TSODL-VD technique, the residual-DenseNet model is applied for feature vector generation from the input video frames and then passes it into the stacked autoencoder (SAE) classifier. The SAE model is applied to recognize the events into violence and non-violence events. In order to improve the violence detection efficacious of the TSODL-VD procedure, the TSO algorithm is utilized as a hyperparameter optimization algorithm for the residual-DenseNet model. The performance validating process of the TSODL-VD procedure is investigated on a benchmark violence dataset. In short, the key contributions of the study is listed as follows.

- An automated violent detection model, named TSODL-VD technique comprising residual-DenseNet feature extractor, SAE classification, and TSO based hyperparameter tuning is presented for violence detection in the smart surveillance system. To the best of our knowledge, the TSODL-VD technique never existed in the literature.

- Residual-DenseNet leverages skip connections, allowing the direct flow of gradients through the network. This facilitates smoother and more efficient backpropagation, which helps address the vanishing gradient problem. As a result, the model can learn more effectively and converge faster during training.
- Hyperparameter tuning using TSO algorithm helps to improve the performance of the residual-DenseNet model. This fusion combines the advantages of swarm intelligence and deep learning, leading to enhanced accuracy and robustness in violence detection.

II. RELATED WORKS

Mohtavipour et al. [11] modelled a deep violence detection structure depending on the particular attributes extracted from handcrafted techniques. Such features are relevant to representative images, appearance, and speed of movement and are provided to Convolutional Neural Networks (CNNs) as spatiotemporal, spatial, and temporal streams. The spatial stream trained network with all frames in the videos for learning atmospheric paradigms. The temporal stream includes three sequential frames for learning motion paradigms of fierce performance with an altered differential magnitude of optical flows. In [12], the authors presented a method through the implementation of renowned ResNet50 for deriving indispensable attributes of all frames of input stream accompanied by a specific schema of recurrent neural networks (ConvLSTM) to find anomalous happenings in time-sequential data. Ehsan et al. [13] presented a new unsupervised network related to motion acceleration paradigms for abstracting discriminatory attributes from inputted trials. This network was built from an AE construction, and it was needed only to utilize normal trials in the training stage. The categorization was executed through one-class techniques for specifying normal and violent activities.

In [14], many key difficulties were incorporated with prevailing work. Initially, violent substances cannot be described manually, and the system should deal with ambiguity. The next stage was the accessibility of the tagged dataset, as physical annotation video was a labour-intensive task and expensive. The CNN techniques were assessed with the presented MobileNet method. The MobileNet technique was contrasted with GoogleNet, AlexNet, and VGG-16 techniques. Mumtaz et al. [15] explored deep representative techniques utilizing transfer learning (TL) for managing the problem of unexpected movement of the camera. Subsequently, a new Deep Multi-Net (DMN) structure related to GoogleNet and AlexNet is modelled for detecting violence in videos. GoogleNet and AlexNet were high-ranked priority trained methods for image categorization having different prior learnt efficient attributes. The combination of such methods may have a better outcome.

Ullah et al. [16] devised a triple-staged end-wise DL violence detection structure. Firstly, people were identified in the streaming of the scrutiny video through the light-weight

CNN method for reducing and solving the voluminous process of unusable frames. Secondly, a series of sixteen frames with recognized persons are sent to 3D CNN, in which spatio-temporal attributes of such series are derived and given to the Softmax method. Additionally, the authors maximized the 3D CNN method utilizing an open visual conclusion and NNs precipitation equipment kit formulated by Intel, which converted the trained method into in-between representations and adjusted it for enhanced implementation at the end platform for the ultimate estimation of aggressive actions. In [17], the authors explored spatio-temporal autocorrelations of gradient-related attributes to proficiently detect violent actions in crowded scenes. A discriminatory was utilized for detecting violent activities in videos.

In [18], intelligent and automated schemes are executed, which attempts to overcome this utilizing DL approaches. The violence from the video was identified utilizing frames, and accuracy was measured. A threat was identified by the method in the video frame, dependent upon that the scheme for condition creates an alert. Qu et al. [19] examine the retrieval and place of violence from long-time series videos. Aiming at the minimal accuracy of violence detection in long-time series video, a 2-stages violence time series place model dependent upon DC3D network method was presented. Mahmoodi et al. [20] present a novel 3D ConvNet together with a process to extract interest frames. During this manner, the 16 video frames with lesser SSIM are assumed that dominant motion frames that are then sent to 3D-CNN for classification. AlDahoul et al. [21] examine a novel structure of end-to-end CNN-LSTM (Long Short-Term Memory) approach, which is run on low-cost Internet of Things (IoT) devices.

Sahay et al. [22] presented a new approach in identifying crime scene video surveillance method in realtime violence detection utilizing DL structures. Yildiz et al. [23] examined a novel automatic audio violence detection (AVD) approach for filling this gap. In [24], a solution has been presented to employ a realtime violence detection approach by DL on UAVs. Ye et al. [25] introduces a physical violence identifying system depends on distributed surveillance cameras. In [8], an AI allowed IIoT-based structure with VD-Network (VD-Net) has been presented. Initially, the input video frames can passed to light-weight CNN approach for essential data gathered comprising humans or suspicious objects like knives/guns. Wang et al. [26] drive of this work is to analysis models of brute force recognition and face detection depends on DL. Febin et al. [27] examine a cascaded approach of violence recognition dependent upon motion boundary SIFT (MoBSIFT) and movement filter. In [28], a new deep NeuralNet approach has been presented for the task of Violence Detection by extraction of motion features in RGB Dynamic Images (DI). Deepak et al. [29] introduce a novel statistical feature descriptor for detecting violent human actions in real-time surveillance videos. Ehsan et al. [30] presented a novel Vi-Net structure dependent upon the deep CNN for detecting activities with abnormal velocity. Optical

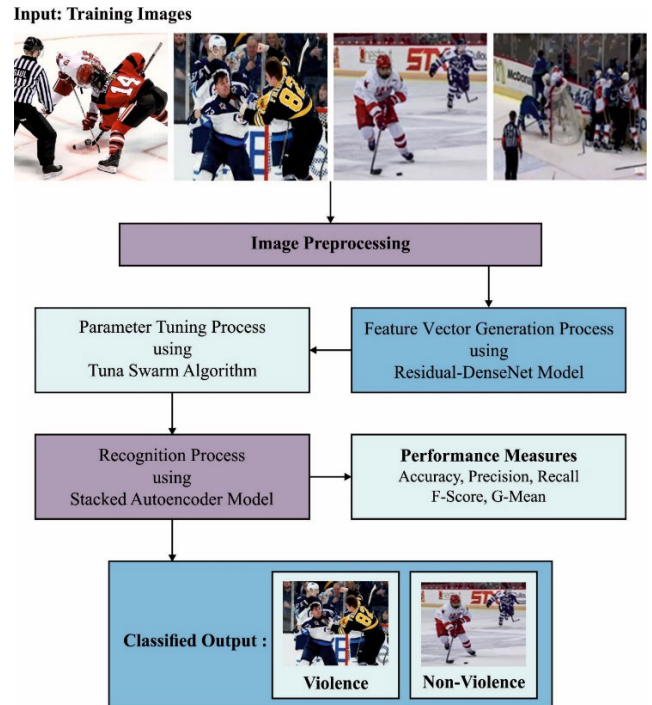


FIGURE 1. The overall workflow of the TSODL-VD approach.

flow vectors to train the Vi-Net network estimation motion designs of targets from the video.

Though several violence detection models are available in the literature, it is still needed to improve the classification performance. Most of the DL models does not concentrate on hyperparameter tuning process, which greatly affects the detection performance. Since manual hyperparameter selection follows a tedious trial and error procedure, metaheuristic algorithms find useful. Therefore, the TSO algorithm is used for the hyperparameter tuning process.

III. THE PROPOSED MODEL

In this research, we have introduced a novel TSODL-VD technique for automated violence identification in surveillance videos. It helps to automatically and accurately recognize violence and can be a measure to avoid any chaotic situations. Fig. 1 demonstrates the comprehensive workflow of the TSODL-VD approach.

A. FEATURE EXTRACTION MODULE

In the presented TSODL-VD technique, the Residual-DenseNet model is applied for feature vector generation. An input of Residual-DenseNet has medicinal images, and the outcome is the vigorous Feature Vector (FV) of images [31]. The Residual-DenseNet is separated into two parts the backbone Network utilized for extracting image feature mapping previously Feature Output Element and the feature output element that procedures the feature mapping resultant by backbone networking.

Backbone Network: According to the DenseNet121 method pre-training on ImageNet, it is more an enduring infrastructure at Dense Block4 for obtaining the Backbone Network of Residual-DenseNet. Input is an image, and the outcome is the feature mapping extraction in the medicinal images.

Feature Output Element: It comprises a 2D Conv layer (kernel size: 1×1), a global average pooling layer, and a 1-D Conv layer. Next, the feature mapping extraction with Backbone Network is dealt with by the Feature Output element, and an FV with 64 lengths can be gained. It can be a robust FV $PFV(i)$ of images.

In Residual-DenseNet, it employs the excellent extraction feature capability of DenseNet121 for extraction feature mapping of distinct scales in the images. Related to low-level aspects, utilizing skip connection for connecting *DenseBlock3* and *DenseBlock4* completely mines the deep semantic data betwixt image, and these higher-level features demonstrated firm vigorousness. The basic concept of the zero-watermarking technique is to connect the imagery feature with the watermark, and the vigorousness of the image extraction feature with the process directly defines the robustness of the zero-watermark technique. Therefore, during the Feature Output element, 1×1 Conv was utilized for reducing the count of feature mapping, global average pooling was executed for reducing dimensional, and lastly, a vigorous length of 64 FV $PFV(i)$ was reached using 1-D Conv. Implement mean binarization operators on the robust FV $PFV(i)$ extracting by Residual-DenseNet for obtaining a vigorous hash vector $FV(i)$ as expressed in Eq. (1).

$$FV(i) = \begin{cases} 1, & PFV(i) \geq \mu \\ 0, & \text{otherwise,} \end{cases} \quad \mu = \frac{1}{64} \sum_{i=0}^{63} PFV(i) \quad (1)$$

B. HYPERPARAMETER TUNING MODULE

To improve the violence detection efficiency of the TSODL-VD technique, the TSO model is utilized as a hyperparameter enhancer for the residual-DenseNet model. The TSO is a bio-inspired optimization algorithm that mimics the collective behavior of tuna fish. It make use of social interaction and self-organization principles for searching optimum solution. With the utilization of the TSO algorithm to detect violent actions, the TSO algorithm explores the vast solution space to effectively identify violent events in video streams. The advantages of the TSA include bio-inspired optimization, efficient search and optimization, global and local exploration, robustness to noise and uncertainty, fewer control parameters, parallel and distributed implementation, versatility and applicability, as well as interpretability and visualization. These qualities make the TSO a promising optimization algorithm on violence detection.

Tuna is identified to be the lead hunter in the ocean [32]. While the swimming of tuna is very rapid, any little victim is further stretchy than tuna. Thus, during the predation

progression, tuna frequently select group cooperation for capturing prey. The tuna swarm (TS) takes two effectual predatory approaches, spiral and parabolic foraging strategies. If the TS utilizes the parabolic foraging approach, all the tuna carry out the prior individual meticulously. The TS procedures a parabola for surrounding the prey. Once the TS implements the spiral foraging approach, the TS would be combined as spiral shapes and effort prey for shallowing water regions. The prey was highly possibly captured. After observing these two foraging performances of TS, the research workers presented a novel, SI optimized named TSA. There are NP tunas from the TS. At the initialized swarm stage, the TSA technique arbitrarily creates the primary swarm from the searching space. The mathematical models to initialize tuna individuals as:

$$X_i^{int} = rand \cdot (ub - lb) + lb$$

$$[x_i^1 x_i^2 \dots x_i^j] = \begin{cases} i = 1, 2, \dots, NP \\ j = 1, 2, \dots, Dim \end{cases} \quad (2)$$

whereas X_i^{int} refers to the i^{th} tuna, ub and lb refer to the upper as well as lesser restrictions of the tuna range searches, and $rand$ denotes the arbitrary variable with uniform distribution from zero to one. Particularly, all the individuals, X_i^{int} , during the TS signifies the candidate's outcome for TSA. All the individuals' tuna comprises a group of Dim -dimensional numbers.

Herring and eel are the essential food resources of tunas. If it encounters a predator, it can utilize its speed benefit to always modify its swimming direction. It is extremely complex for predators to catch them. While the tuna has lesser agile than its prey, the TS drive takes a cooperative scheme for attacking the prey. The TS can utilize the victim as a source point to maintain the chase of the victim. In this hunting, all the tunas follow the prior individual, and the entire TS procedures a parabola for surrounding the prey. Also, the TS utilizes a spiral foraging scheme.

Considering that the possibility of TSs selecting both approaches is 50%, the scientific process of parabolic scavenging of TS is:

$$X_i^{t+1} = \begin{cases} X_{best}^t + rand \cdot (X_{best}^t - X_i^t) + TF \cdot p^2 \cdot (X_{best}^t - X_i^t), & \text{if } rand < 0.5 \\ TP \cdot p^2 \cdot X_i^t, & \text{if } rand \geq 0.5 \end{cases} \quad (3)$$

$$p = \left(1 - \frac{t}{t_{max}}\right)^{t \left(\frac{t}{t_{max}}\right)} \quad (4)$$

whereas t demonstrates that t^{th} iteration has presently run, and t_{max} signifies the maximal count of iterations preset. TP signifies the random value of 1 or -1 .

In addition, to the parabolic scavenging scheme, there exists another effectual cooperative scavenging scheme named the spiral scavenging scheme. But chasing the victim, most tuna could not select the right direction; however,

a smaller count of tunas guided the swarm to swim in the correct way. If the smaller groups of tuna begin to chase the victim, the neighbouring tuna carry out this smaller ensemble of entities. Finally, the whole TSs are procedure a spiral development for catching the prey. Once the TS implements a spiral foraging scheme, individuals interchange data with optimum in carrying out the individuals or neighbouring individuals from the swarm. The tuna then chooses an arbitrary individual from the swarm to follow. The scientific equation of the spiral scavenging approach is as follows (5), shown at the bottom of the next page, whereas X_i^{t+1} implies the i^{th} tuna from the $t + 1$ iteration. The present optimum individual is X_{best}^t . X_{rand}^t refers to the reference point arbitrarily chosen from the TS. α_1 denotes the trend weighted co-efficient for controlling the tuna swim individual for optimum individual or arbitrarily chosen neighbouring individuals. α_2 stands for the trend weighted co-efficient for controlling the tuna swim individual to the individual facing each other. τ demonstrates the distance parameter, which controls the distance betwixt the tuna individual as well as the optimum individual or an arbitrarily chosen reference individual. The mathematical computation process is as follows:

$$\alpha_1 = a + (1 - a) \cdot \frac{t}{t_{max}} \quad (6)$$

$$\alpha_2 = (1 - a) - (1 - a) \cdot \frac{t}{t_{max}} \quad (7)$$

$$\tau = e^{bl} \cdot \cos(2\pi b) \quad (8)$$

$$l = e^{3\cos((t_{max}+1/t)-1)\pi} \quad (9)$$

In which a signifies the constant for measuring the degree of tunas following, and b represents the arbitrary number uniformly distributing from the range of zero and one. During the iterative procedure of the TSA technique, all the tunas are arbitrarily selected for performing both the spiral and parabolic foraging approaches. Tuna also creates novel individuals from the searching range based on probability Z . Thus, the TSA select various approaches based on Z if creating a novel individual position. In the implementation of the TSA technique, every tuna individual from the population is constantly upgraded; still, the count of iterations gains a pre-defined value. Eventually, the TSA technique returns a better individual from the population and their better value. The following benefits of TSA are realized in Algorithm 1: (i) The TSA technique has some modifiable parameters that are helpful to the execution of the algorithm. (ii) This algorithm keeps the location of the optimum tuna individual from all the iterations. At the same time, the quality of candidate solutions reduces, and it cannot affect the place of optimum value. (iii) The TSA technique saves the balance betwixt exploitation as well as exploration by choosing two foraging approaches.

The TSO method derived a fitness function from having enhanced categorizing results. It defined positive values for designating higher derivatives of the candidate resolutions. In this article, the mitigated categorizer fault rate was studied

as the fitness function, as presented in Eq. (10).

$$\begin{aligned} fitness(x_i) &= ClassifierErrorRate(x_i) \\ &= \frac{numberofmisclassifiedsamples}{Totalnumberofsamples} * 100 \quad (10) \end{aligned}$$

C. VIOLENCE CLASSIFICATION MODULE

In this study, the SAE model is enforced for the classification of events into violence and non-violence events. AE is a typical Feed-Forward Neural Network (FFNN). The networking input to Hidden Layer (HL) is assumed to be the encoder approach. Additionally, the HL to the resultant layer was regarded as the decoder method [33]. The encoder data can be rebuilt to novel data with the decoding approach. The AE minimization the MSE of inputting and outputting with trained networking for achieving data extraction features. The extraction aspect of AE generally offers three distinct manifestations: Primarily, the node in HL are lesser than in input as well as output layers, and the extraction aspects are compressed dimension mitigation representations of trained data. Secondly, the HL node is greater than the input as well as output layers, and the extraction factors are the higher-dimension depiction of the trained record. At last, the HL node has equivalent to the input node, and the feature was the equivalent-dimension representation of trained data. Encode and decode approaches of AE are:

$$\begin{cases} h = \sigma(W_1x + b_1) \\ \hat{x} = \sigma(W_2h + b_2) \end{cases} \quad (11)$$

In which $x = [x_1, x_2, \dots, x_m]^T \in R^m$ and $\hat{x} = [\hat{x}_1, \hat{x}_2, \dots, \hat{x}_v]^T \in R^m$ implies the inputting and outputting layers of AE networks correspondingly. $h = [h_1, h_2, \dots, h_v] \in R^v$ refers the HL of networks. $W_1 \in R^{v \times m}$ and $b_1 \in R^v$ denote the weighted matrix and offset vector of HLs. Also, $W_2 \in R^{v \times m}$ and $b_2 \in R^v$ indicate the weighted matrix and bias vector of the resultant layer, correspondingly. σ stands for the neuron activation function, mostly utilizing tanh and sigmoid roles. Pic. 2 demonstrates the framework of SAE.

SAE is a DL technique stacked by AEs. It removes the decoder part, afterwards trains the AE and feeds the HL parameter achieved by primary AE as secondary AE to train for achieving a novel feature representation. Then repeat the above steps many times, and the preferred SAE technique is gained. Then the network pre-trained approach, the equivalent HLs of all the AEs, can be stacked for the procedure of a deep AE network with several HLs. Compared with typical AE, the parameter of networks can be fine-tuned utilizing the back-propagation (BP) technique by computing the error betwixt the network output and input layers. While an archetypal DL network, SAE is superior extraction feature capability than AE, and its pre-trained stages avoid problems like over-fitting of a network trained.

The hyperparameter tuning procedure of the SAE method can be made by Adam enhancer. It can be a type of common Stochastic Gradient Descent (SGD) technique to upgrade networking weights in trained records [34]. It is used to

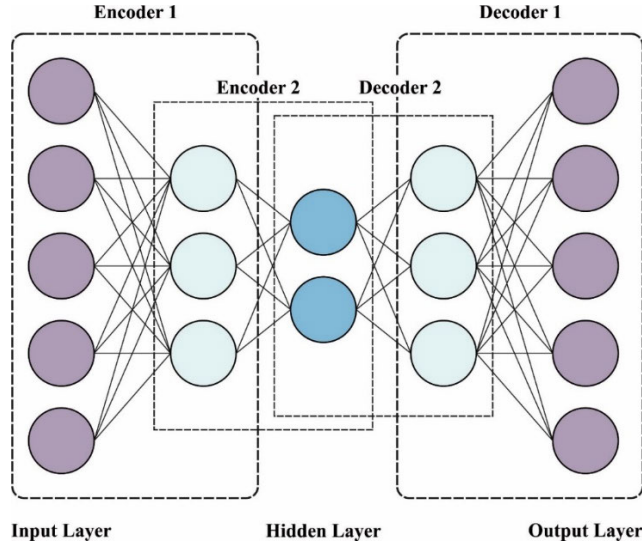


FIGURE 2. The architecture of SAE.

perform optimization and was the best optimizer. Adam proceeds in adagrad, and it is an additional flexible manner. Adagrad and momentum together are known as Adam.

Parameters $w^{(t)}$ and $L^{(t)}$, where index t designates the currently trained repetition, Parameter enhancement in Adam is provided below:

$$m_w^{(t+1)} \leftarrow \beta_1 m_w^{(t)} + (1 - \beta_1) \nabla_w L^{(t)} \quad (12)$$

$$v_w^{(t+1)} \leftarrow \beta_2 v_w^{(t)} + (1 - \beta_2) (\nabla_w L^{(t)})^2 \quad (13)$$

$$\hat{m}_w = \frac{m_w^{(t+1)}}{1 - (\beta_1)^{(t+1)}} \quad (14)$$

$$V_w = \frac{v_w^{(t+1)}}{1 - (\beta_2)^{(t+1)}} \quad (15)$$

$$w^{t+1} \leftarrow w^t - \eta \frac{\hat{m}}{\sqrt{\hat{v}_w} + \epsilon} \quad (16)$$

Here β_2 and β_1 designate the second moment of gradients and gradient forgetting features. In Eq. (16), ϵ denotes the smaller scalar used to prevent division by 0.

IV. PERFORMANCE VALIDATION

The violence recognition achievement of the TSODL-VD technique is inspected on two datasets [35]: the hockey fights

TABLE 1. Details of dataset.

Class	No. of Instances	
	Hockey Fights Dataset	Movies Dataset
Violence	500	100
Non-Violence	500	100
Total No. of Instances	1000	200

dataset and the movies dataset. The particulars associated with the datasets are represented in Table 1. The Hockey Fights dataset comprises clips from ice-hockey matches. The dataset has 500 violent clips and 500 non-violent clips of average duration of 1 s. The clips had a similar background and subjects. Every clip has of 50 frames of 720×576 pixels and is manually labeled as “fight” or “non-fight”. Next, the Movies dataset contains clips from different movies for action sequence whereas the non-fight sequences consist of clips from action recognition datasets. The dataset has 100 violent clips and 100 non-violent clips of average duration of 1 s. Unlike the Hockey Fights dataset, the clips of movies have different backgrounds and subjects.

The confusion matrix of the TSODL-VD technique on the violence detection process is demonstrated in Fig. 3. The results signify that the TSODL-VD technique can accurately recognize the violence and non-violence events in both movies and hockey fights datasets.

In Table 2 and Pic. 4, the violence detection outcomes of the TSODL-VD procedure on 70:30 of TRS/TSS under the hockey fights dataset are stated. The attained values indicate the effectual recognition accomplishment of the TSODL-VD procedure. For instance, on 70% of TRS, the TSODL-VD technique reaches an average $accu_{bal}$ of 98.72%, $prec_n$ of 98.71%, $reca_l$ of 98.72%, F_{score} of 98.71%, and G_{mean} of 98.72%. Meanwhile, on 30% of TSS, the TSODL-VD method reaches an average $accu_{bal}$ of 98.65%, $prec_n$ of 98.69%, $reca_l$ of 98.65%, F_{score} of 98.67%, and G_{mean} of 98.65%.

The training accuracy (TACC) and validation accuracy (VACC) of the TSODL-VD technique under hockey fights dataset accomplishment in Pic. 5. The results designated that the TSODL-VD algorithm has enhanced accomplishment

$$X_i^{t+1} = \begin{cases} \alpha_1 \cdot (X_{rand}^t + \tau \cdot |X_{rand}^t - X_i^t| + \alpha_2 \cdot X_i^t), & \\ \quad i = 1 & \\ \alpha_1 \cdot (X_{rand}^t + \tau \cdot |X_{rand}^t - X_i^t| + \alpha_2 \cdot X_{i-1}^t), & \text{if } rand < \frac{t}{t_{max}} \\ \quad i = 2, 3, \dots, NP & \\ \alpha_1 \cdot (X_{best}^t + \tau \cdot |X_{best}^t - X_i^t| + \alpha_2 \cdot X_i^t) & \\ \quad i = 1 & \\ \alpha_1 \cdot (X_{best}^t + \tau \cdot |X_{best}^t - X_i^t| + \alpha_2 \cdot X_{i-1}^t) & \text{if } rand \geq \frac{t}{t_{max}} \\ \quad i = 2, 3, \dots, NP & \end{cases} \quad (5)$$

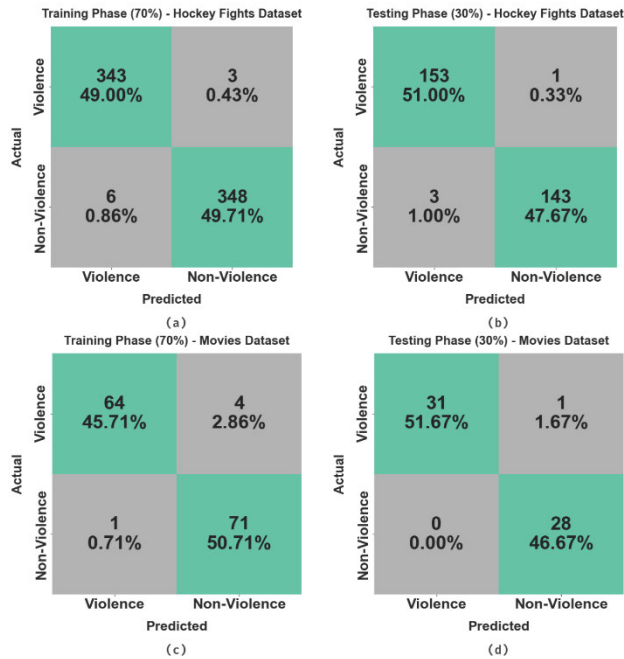


FIGURE 3. Confusion matrices of TSODL-VD structure (a-b) 70:30 of TRS/TSS under hockey fights dataset and (c-d) 70:30 of TRS/TSS under movies dataset.

TABLE 2. Violence detection outcome of TSODL-VD approach on 70:30 of TRS/TSS under hockey fights dataset.

Hockey Fights Dataset					
Class	Accuracy _{bal}	Precision	Recall	F-Score	G-Mean
Training Phase (70%)					
Violence	99.13	98.28	99.13	98.71	98.72
Non-Violence	98.31	99.15	98.31	98.72	98.72
Average	98.72	98.71	98.72	98.71	98.72
Evaluation Stage (30%)					
Violence	99.35	98.08	99.35	98.71	98.65
Non-Violence	97.95	99.31	97.95	98.62	98.65
Average	98.65	98.69	98.65	98.67	98.65

with improved values of TACC and VACC. Especially the TSODL-VD procedure has reached the utmost TACC results.

The training loss (TLS) and validation loss (VLS) of the TSODL-VD procedure under hockey fight dataset accomplishment in Pic. 6. The results shows the TSODL-VD technique has exhibited better accomplishment with lesser values of TLS and VLS.

In Table 3 and Pic. 7, the violence recognition results of the TSODL-VD method on 70:30 of TRS/TSS under the movies dataset are reported. The acquired values specify the effectual recognition performance of the TSODL-VD process. For example, on 70% of TRS, the TSODL-VD method reaches

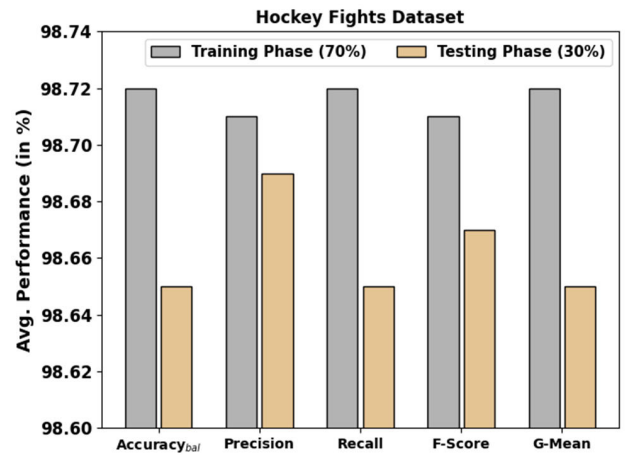


FIGURE 4. The average outcome of the TSODL-VD approach under the hockey fights dataset.

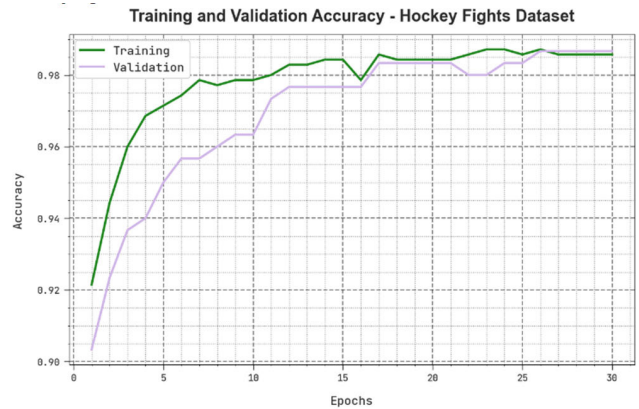


FIGURE 5. TACC and VACC result of TSODL-VD approach under hockey fights dataset.

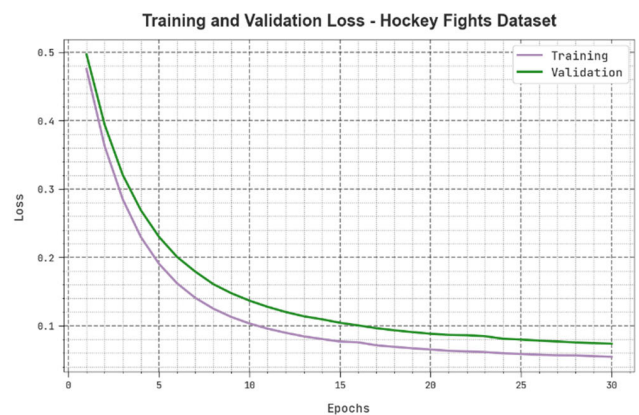


FIGURE 6. TLS and VLS outcome of TSODL-VD approach under hockey fights dataset.

an average $accu_{bal}$ of 96.36%, $prec_n$ of 96.56%, $reca_l$ of 96.36%, F_{score} of 96.42%, and G_{mean} of 96.34%. In the meantime, on 30% of TSS, the TSODL-VD approach reaches an average $accu_{bal}$ of 98.44%, $prec_n$ of 98.28%, $reca_l$ of 98.44%, F_{score} of 98.33%, and G_{mean} of 98.43%.

TABLE 3. Violence recognition result of TSODL-VD procedure on 70:30 of TRS/TSS under movies dataset.

Movies Dataset					
Class	Bal. Accuracy	Precision	Recall	F-Score	G-Mean
Training Phase (70%)					
Violence	94.12	98.46	94.12	96.24	96.34
Non-Violence	98.61	94.67	98.61	96.60	96.34
Average	96.36	96.56	96.36	96.42	96.34
Testing Phase (30%)					
Violence	96.88	100.00	96.88	98.41	98.43
Non-Violence	100.00	96.55	100.00	98.25	98.43
Average	98.44	98.28	98.44	98.33	98.43

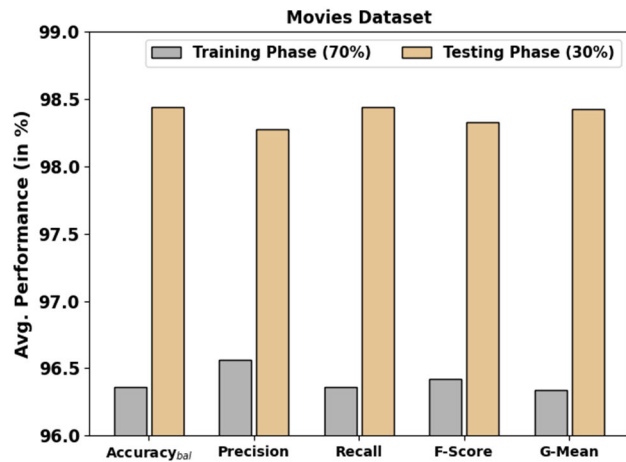


FIGURE 7. The average outcome of the TSODL-VD approach under the movies dataset.

The TACC and VACC of the TSODL-VD algorithm under movies dataset accomplishment in Pic. 8. The picture demonstrates the TSODL-VD procedure has depicted enhanced accomplishment with enhanced values of TACC and VACC. Remarkably, the TSODL-VD technique has reached supreme TACC results.

The TLS and VLS of the TSODL-VD procedure under movies dataset accomplishment in Fig. 9. The picture exhibited that the TSODL-VD procedure has exposed better accomplishment with the least values of TLS and VLS. Apparently, the TSODL-VD approach has given an outcome in mitigated VLS results.

Fig. 10 demonstrates the classifier results of the TSODL-VD technique under the hockey flights and movies dataset. Figs. 10a-10b demonstrates the PR analysis of the TSODL-VD technique under the hockey flights and movies dataset. The pictures specified that the TSODL-VD approach had gained greater PR accomplishment under all categories. Finally, Figs. 10c-10d exemplifies the ROC research of the TSODL-VD technique under the hockey flights and movies dataset. The picture shows that the TSODL-VD approach has given an outcome in expert outcomes with greater ROC values under different class tags.



FIGURE 8. TACC and VACC outcome of TSODL-VD approach under movies dataset.

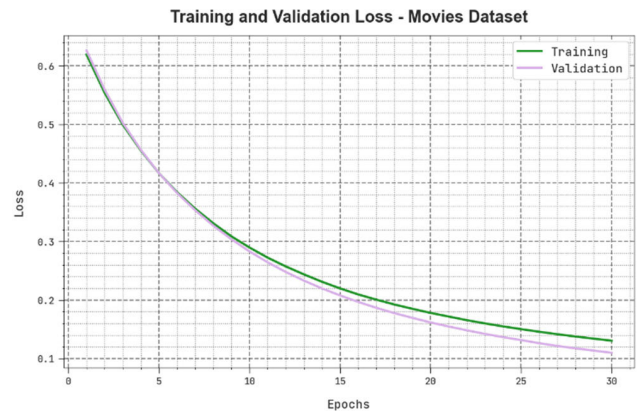


FIGURE 9. TLS and VLS outcome of TSODL-VD approach under movies dataset.

To demonstrate the better violence classification results of the TSODL-VD technique, a widespread comparison research is given in Table 4 [2], [36]. The table values inferred that the TSODL-VD technique reaches increasing values of $accu_y$ on both datasets. For instance, on the hockey flights dataset, the TSODL-VD technique attains a higher $accu_y$ of 98.72%. In contrast, the CNN-BiLSTM, Motion-IWLD, MobileNet, Inception-ResNet, SVM, and HOG3D-KELM techniques obtain decreasing $accu_y$ of 95.32%, 97.42%, 93.83%, 91.32%, 92.71%, and 93.54% respectively. Furthermore, on the movies dataset, the TSODL-VD method achieves a higher $accu_y$ of 98.44% while the CNN-BiLSTM, Motion-IWLD, MobileNet, Inception-ResNet, SVM, and HOG3D-KELM methods obtain decreasing $accu_y$ of 92.19%, 96.63%, 91.36%, 94.98%, 95%, and 97.52% correspondingly. These results ensured that the TSODL-VD technique has proficiently recognized the violence in the surveillance videos.

In summary, the TSODL-VD technique exhibits better performance with maximum $accu_y$ of 98.72% and 98.44% on Hockey fights and movie datasets respectively. The enhanced performance of the proposed model is due to the incorporation of the residual-DenseNet and TSO based

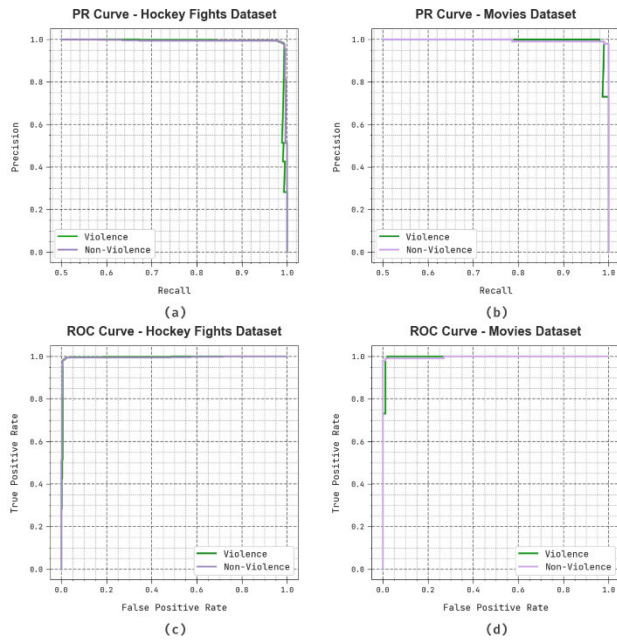


FIGURE 10. (a-c) PR and ROC curves under the hockey flights dataset and (b-d) PR and ROC curves under the movies dataset.

TABLE 4. Relative examination of the TSODL-VD system with other recent procedures [2], [36].

Accuracy (%)		
Methods	Hockey Fights Dataset	Movies Dataset
TSODL-VD	98.72	98.44
CNN-BiLSTM	95.32	92.19
Motion-IWLD	97.42	96.63
MobileNet	93.83	91.36
Inception-ResNet	91.32	94.98
SVM	92.71	95.00
HOG3D-KELM	93.54	97.52

hyperparameter tuning. The dense connectivity pattern in Residual-DenseNet encourages feature reuse throughout the network. In addition, it ensures that information can propagate more directly and rapidly throughout the network. This allows the model to preserve and propagate valuable information across layers, enabling effective feature learning even in deeper networks. As a result, Residual-DenseNet can capture both low-level and high-level features more efficiently, leading to improved representation learning. The advantages of the residual-DenseNet model include improved gradient flow, feature reuse, reduced parameters and memory usage, enhanced information flow, mitigation of overfitting, and compatibility make it a powerful and efficient deep learning architecture for various computer vision tasks. On the other hand, the TSO chooses the optimal values for the hyperparameters of a given residual-DenseNet model. Hyperparameters are settings that are not learned during training, but must be set prior to training. They can have

a significant impact on the performance of the model, and selecting the optimal values can lead to better accuracy. The TSA facilitates efficient search and optimization, guiding the deep learning models to identify relevant features associated with violence. This fusion of swarm intelligence and DL offers a synergistic effect, resulting in improved accuracy and robustness in violence detection. These results ensured the improved performance of the TSODL-VD technique over other existing techniques.

V. CONCLUSION

In this research, we have introduced a novel TSODL-VD method for automated violence recognition in surveillance videos. It helps to automatically and accurately recognize violence and can be a measure to avoid any chaotic situations. In the presented TSODL-VD technique, the Residual-DenseNet model is applied for feature vector generation, and the SAE model is applied for the categorization of events into fierceness and non-fierceness events. To improve the violence detection effectiveness of the TSODL-VD procedure, the TSO protocol is utilized as a hyperparameter enhancer for the residual-DenseNet model. The performance validation of the TSODL-VD procedure is examined on the benchmark violence dataset. The experimental results demonstrate that the TSODL-VD technique accomplishes precise and rapid detection outcomes over the recent state of the art approaches.

ACKNOWLEDGMENT

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through Large Groups Project under grant number (RGP2/65/44). Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R387), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. Research Supporting Project number (RSP2023R459), King Saud University, Riyadh, Saudi Arabia. This study is supported via funding from Prince Sattam bin Abdulaziz University project number (PSAU/2023/R/1444).

REFERENCES

- [1] M. Sharma and R. Baghel, "Video surveillance for violence detection using deep learning," in *Advances in Data Science and Management*. Singapore: Springer, 2020, pp. 411–420.
- [2] F. U. M. Ullah, M. S. Obaidat, K. Muhammad, A. Ullah, S. W. Baik, F. Cuzzolin, J. J. P. C. Rodrigues, and V. H. C. de Albuquerque, "An intelligent system for complex violence pattern analysis and detection," *Int. J. Intell. Syst.*, vol. 37, no. 12, pp. 10400–10422, Dec. 2022.
- [3] G. Sreenu and M. A. S. Durai, "Intelligent video surveillance: A review through deep learning techniques for crowd analysis," *J. Big Data*, vol. 6, no. 1, pp. 1–27, Dec. 2019.
- [4] S. Roshan, G. Srivathsan, K. Deepak, and S. Chandrakala, "Violence detection in automated video surveillance: Recent trends and comparative studies," in *The Cognitive Approach in Cloud Computing and Internet of Things Technologies for Surveillance Tracking Systems*, 2020, pp. 157–171.
- [5] M. Baba, V. Gui, C. Cernazanu, and D. Pescaru, "A sensor network approach for violence detection in smart cities using deep learning," *Sensors*, vol. 19, no. 7, p. 1676, Apr. 2019.

- [6] H. Yao and X. Hu, "A survey of video violence detection," *Cyber-Phys. Syst.*, vol. 9, no. 1, pp. 1–24, Jan. 2023.
- [7] P. Sernani, N. Falcionelli, S. Tomassini, P. Contardo, and A. F. Dragoni, "Deep learning for automatic violence detection: Tests on the AIRTLab dataset," *IEEE Access*, vol. 9, pp. 160580–160595, 2021.
- [8] F. U. M. Ullah, K. Muhammad, I. U. Haq, N. Khan, A. A. Heidari, S. W. Baik, and V. H. C. de Albuquerque, "AI-assisted edge vision for violence detection in IoT-based industrial surveillance networks," *IEEE Trans. Ind. Informat.*, vol. 18, no. 8, pp. 5359–5370, Aug. 2022.
- [9] K. Rezaee, S. M. Rezaekhani, M. R. Khosravi, and M. K. Moghimi, "A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance," *Pers. Ubiquitous Comput.*, pp. 1–17, Jun. 2021.
- [10] A. Srivastava, T. Badal, P. Saxena, A. Vidyarthi, and R. Singh, "UAV surveillance for violence detection and individual identification," *Automated Softw. Eng.*, vol. 29, no. 1, pp. 1–28, May 2022.
- [11] S. M. Mohtavipour, M. Saeidi, and A. Araborkhi, "A multi-stream CNN for deep violence detection in video sequences using handcrafted features," *Vis. Comput.*, vol. 38, no. 6, pp. 2057–2072, Jun. 2022.
- [12] S. Vosta and K.-C. Yow, "A CNN-RNN combined structure for real-world violence detection in surveillance cameras," *Appl. Sci.*, vol. 12, no. 3, p. 1021, Jan. 2022.
- [13] T. Z. Ehsan, M. Nahvi, and S. M. Mohtavipour, "DABA-Net: Deep acceleration-based AutoEncoder network for violence detection in surveillance cameras," in *Proc. Int. Conf. Mach. Vis. Image Process. (MVIP)*, Feb. 2022, pp. 1–6.
- [14] T. Hussain, A. Iqbal, B. Yang, and A. Hussain, "Real time violence detection in surveillance videos using convolutional neural networks," *Multimedia Tools Appl.*, vol. 81, pp. 38151–38173, Apr. 2022.
- [15] A. Mumtaz, A. B. Sargano, and Z. Habib, "Fast learning through deep multi-net CNN model for violence recognition in video surveillance," *Comput. J.*, vol. 65, no. 3, pp. 457–472, Mar. 2022.
- [16] F. U. M. Ullah, A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Violence detection using spatiotemporal features with 3D convolutional neural network," *Sensors*, vol. 19, no. 11, p. 2472, May 2019.
- [17] K. Deepak, L. K. P. Vignesh, and S. J. I. E. Chandrakala, "Autocorrelation of gradients based violence detection in surveillance videos," *ICT Exp.*, vol. 6, no. 3, pp. 155–159, Sep. 2020.
- [18] N. Suba, A. Verma, P. Baviskar, and S. Varma, "Violence detection for surveillance systems using lightweight CNN models," in *Proc. 7th Int. Conf. Comput. Eng. Technol. (ICCET)*, vol. 2022, Feb. 2022, pp. 23–29.
- [19] W. Qu, T. Zhu, J. Liu, and J. Li, "A time sequence location method of long video violence based on improved C3D network," *J. Supercomput.*, vol. 78, no. 18, pp. 19545–19565, Dec. 2022.
- [20] J. Mahmoodi, H. Nezamabadi-pour, and D. Abbasi-Moghadam, "Violence detection in videos using interest frame extraction and 3D convolutional neural network," *Multimedia Tools Appl.*, vol. 81, no. 15, pp. 20945–20961, Jun. 2022.
- [21] N. AlDahoul, H. A. Karim, R. Datta, S. Gupta, K. Agrawal, and A. Alburni, "Convolutional neural network–long short term memory based IoT node for violence detection," in *Proc. IEEE Int. Conf. Artif. Intell. Eng. Technol. (IICAIET)*, Sep. 2021, pp. 1–6.
- [22] K. B. Sahay, B. Balachander, B. Jagadeesh, G. A. Kumar, R. Kumar, and L. R. Parvathy, "A real time crime scene intelligent video surveillance systems in violence detection framework using deep learning techniques," *Comput. Electr. Eng.*, vol. 103, Oct. 2022, Art. no. 108319.
- [23] A. M. Yildiz, P. D. Barua, S. Dogan, M. Baygin, T. Tuncer, C. P. Ooi, H. Fujita, and U. Rajendra Acharya, "A novel tree pattern-based violence detection model using audio signals," *Expert Syst. Appl.*, vol. 224, Aug. 2023, Art. no. 120031.
- [24] H. H. Nguyen, Q. T. Le, V. Q. Nghiem, M. S. Hoang, and D. A. Pham, "A novel violence detection for drone surveillance system," in *Proc. Int. Conf. Commun., Circuits, Syst. (IC3S)*, May 2023, pp. 1–6.
- [25] L. Ye, S. Yan, J. Zhen, T. Han, H. Ferdinando, T. Seppänen, and E. Alasaarela, "Physical violence detection based on distributed surveillance cameras," *Mobile Netw. Appl.*, vol. 27, no. 4, pp. 1688–1699, Aug. 2022.
- [26] P. Wang, P. Wang, and E. Fan, "Violence detection and face recognition based on deep learning," *Pattern Recognit. Lett.*, vol. 142, pp. 20–24, Feb. 2021.
- [27] I. P. Febin, K. Jayasree, and P. T. Joy, "Violence detection in videos for an intelligent surveillance system using MoBSIFT and movement filtering algorithm," *Pattern Anal. Appl.*, vol. 23, no. 2, pp. 611–623, May 2020.
- [28] A. Jain and D. K. Vishwakarma, "Deep NeuralNet for violence detection using motion features from dynamic images," in *Proc. 3rd Int. Conf. Smart Syst. Inventive Technol. (ICSSIT)*, Aug. 2020, pp. 826–831.
- [29] K. Deepak, L. K. P. Vignesh, G. Srivathsan, S. Roshan, and S. Chandrakala, "Statistical features-based violence detection in surveillance videos," in *Cognitive Informatics and Soft Computing*. Singapore: Springer, 2020, pp. 197–203.
- [30] T. Z. Ehsan and S. M. Mohtavipour, "Vi-Net: A deep violent flow network for violence detection in video sequences," in *Proc. 11th Int. Conf. Inf. Knowl. Technol. (IKT)*, Dec. 2020, pp. 88–92.
- [31] C. Gong, J. Liu, M. Gong, J. Li, U. A. Bhatti, and J. Ma, "Robust medical zero-watermarking algorithm based on residual-DenseNet," *IET Biometrics*, vol. 11, no. 6, pp. 547–556, Nov. 2022.
- [32] W. Wang and J. Tian, "An improved nonlinear tuna swarm optimization algorithm based on circle chaos map and levy flight operator," *Electronics*, vol. 11, no. 22, p. 3678, Nov. 2022.
- [33] B. Liu, Y. Chai, Y. Jiang, and Y. Wang, "Industrial fault detection based on discriminant enhanced stacking auto-encoder model," *Electronics*, vol. 11, no. 23, p. 3993, Dec. 2022.
- [34] K. K. Chandriah and R. V. Naraganahalli, "RNN/LSTM with modified Adam optimizer in deep learning approach for automobile spare parts demand forecasting," *Multimedia Tools Appl.*, vol. 80, pp. 26145–26159, Apr. 2021.
- [35] E. B. Nieves, O. D. Suarez, G. B. Garcia, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Proc. Int. Conf. Comput. Anal. Images Patterns*. Berlin, Germany: Springer, 2011, pp. 332–339.
- [36] R. Halder and R. Chatterjee, "CNN-BiLSTM model for violence detection in smart surveillance," *Social Netw. Comput. Sci.*, vol. 1, no. 4, pp. 1–9, Jul. 2020.

• • •