Received: April 12, 2023. Revised: June 3, 2023.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Utilizing Machine Learning and Computer Vision for the Detection of Abusive Behavior in IoT Systems

Suadad Zaidan Khalaf[1]* **Mohamed Ibrahim Shujaa[1]** **Ahmed Bahaaulddin A. Alwahhab[1]**

[1]*Department of Computer Techniques Engineering, Middle Technical University, Baghdad, Iraq*
**Correspondence:* bbc0069@mtu.edu.iq

Abstract: Law enforcement and civilian protection are increasingly dependent on surveillance technologies. To prevent social, economic, and environmental harm, video surveillance situations like those in train stations, schools, and hospitals must automatically detect aggressive and suspicious behavior. Intelligent video surveillance systems now enable human interaction monitoring. Despite its security benefits, crowds and the camera's vision make it difficult to distinguish regular from abnormal activity. Therefore, there is a great deal of investigation into the many methods of identifying violent behavior. The detection approach presented in this research can be broken down into three distinct subfields. The classification methods employed led to the formation of these classes. Classifications include Gradient Descent (SGD), Adaptive boosting (ADA), Naive Base (NB), and Logistic Regression (LR) for detecting violent acts with machine learning. Methods for feature extraction and violence detection are derived using Principal Component Analysis (PCA). Dataset learning- based models for violence detection can also be evaluated using the AIRTLab dataset, a model developed to test the robustness of algorithms against false positives. In order to determine if the proposed model's resistance to false positives, accuracy (ACC) metrics were computed on the proposed model and used to set a benchmark for the AIRTLab dataset. According to the research literature, the proposed models are accurate of 98.31% with the logistic regression classifier having superior generalization skills.

Keywords: Computer vision, Machine learning, Violence, Human behavior analysis.

1. Introduction

Closed-circuit television (CCTV) cameras have become increasingly affordable in recent years, leading to a rise in their use for both public and private video monitoring. These cameras are used for surveillance purposes in a variety of settings, including at home and in the workplace. It is common practice to save video frames in a database for later analysis in the event of an anomaly. Additionally, an operator is sometimes utilized to keep an eye on these cameras. Paying a person to constantly keep an eye on all of the numerous CCTV cameras would be prohibitively expensive, and the operator would inevitably make mistakes due to things like exhaustion and distraction. However, recent years have seen a proliferation of intelligent CCTV cameras, which evaluate behavior automatically and rapidly. In the fields of machine vision and AI, these kinds of systems are among the most dynamic research topics [1, 2]. The development of real-time intelligent systems for identifying anomalous behaviors and averting catastrophic mishaps is a major area of study at the moment. The range of human behavior can be broken down into three distinct tiers. Movements of a single body part, such as a raised hand or a thrown punch, are the basis of the first level of gesture recognition[3]. Next up are activities that require less effort and time, such as walking, running, and sitting [4].Interactions between many targets constituted of complicated behaviors like violence are at the pinnacle of human activity [5].

Each individual has a unique style of carrying out the same activity, making it difficult to generalize about how to detect violent behavior [6].On the other hand, it might be confusing to tell the difference between friendly and aggressive actions like hugging

and hand-shaking. Another thing is that activities can look different depending on the camera angle [7]. A similar movement seen from a different perspective, for instance, can take on a variety of appearances, all of which serve to raise the recognition system's complexity. A reliable recognition technique will be able to detect abnormalities in the presence of these factors.

There are both ground-level and high-level components in the study of human behavior. Machine vision techniques form the basis of the low-level analyses. This process involves the manual extraction of several feature descriptors employing texture [8, 9], motion [10–14], and forms [15–19]. Poor camera resolution [20], target shadows [21], picture noise [22], partial or complete target occlusion [23], and variations in illumination [24] all make it challenging to locate a single characteristic that adequately represents target behavior across all of these scenarios. However, advanced analysis makes use of machine learning methods, with models being trained to assign behaviors to predetermined classes. Support vector machine (SVM) [25], nearest neighbor search (NNS) [26], random forest (RF) [27], and deep learning (DL) [28–30] are just a few examples of the various categorization strategies available for separating violent from nonviolent behavior.

Our research aims to enhance the ACC and reliability of violence detection systems in machine learning. It focuses on the comprehensive analysis of spatial, temporal, and spatiotemporal components inherent in violent behavior. The objective is to discern distinctive features that allow for a more precise distinction between violent and non-violent behavior patterns. We argue that to adequately capture the complex dynamics of violent actions, an integrated approach encompassing various attributes is essential. Therefore, we propose discriminative features that are derived differentially and pivot on three key aspects: appearance, movement speed, and a representative image of actions. The latter is achieved by accumulating frames over time, creating an encompassing depiction of the action taking place. These features are designed not just to recognize individual elements of violence, but to comprehend the broader behavioral context in which they occur. We believe that incorporating these multifaceted characteristics into machine learning models can significantly enhance their violence detection capability. To that end, the problem this work addresses is the need for an improved violence detection system in machine learning, one that takes into account not just isolated instances, but the overall picture of violent behavior. It seeks to respond to this need by identifying and proposing discriminative features that offer a more comprehensive view of violent actions, based on differential analysis of appearance, speed, and an overall representative image of actions. The remainder of this paper is organized as follows: section 2 provides a description and comparison of related works of violence detection methods; section 3 introduces the proposed framework, including feature extraction and ML architecture; section 4 provides experimental results and discussion on dataset; and section 5 provides a conclusion and future works.

2. Related work

DL models, specifically three-dimensional convolutional neural networks (3D CNNs) and convolutional long short-term memory (3D ConvLSTMs), have recently shown promising outcomes in the realm of violence detection, surpassing the results obtained by previous methods [31]. The RWF-2000 dataset, introduced by Cheng et al. [32], was designed to overcome certain limitations by incorporating 2000 segments of surveillance camera footage sourced from YouTube. However, its distinguishing element, the inclusion of non-violent but rapid and potentially misleading movements (hugs, high-fives, applauding, exulting, and gesticulating), presents an increased risk of false positives.

Despite these advancements, each method has its shortcomings.

Ding et al.'s [33] suggested 9-layer 3D CNN runs under the presumption that 40 frames with a resolution of 60 x 90 pixels may be processed simultaneously and obtained an astounding 91% ACC on the Hockey Fight Dataset. This processing constraint may limit the model's scalability and versatility in handling videos with different resolutions or lengths.

Song et al.'s [34] The enhanced C3D architecture still falls short since it involves training the 3D CNN from scratch, which can be computationally expensive and time-consuming. This upgraded C3D architecture obtained 99 percent ACC on the Hockey Fight Dataset and 94.3 percent on the Crowd Violence.

Li et al.'s [35] 10-layer 3D CNN, despite its high ACC scores of 98.3% and 97.2% on the Hockey Fight and Crowd Violence datasets respectively, employed an architecture comprising alternating dense and transitional layers following a convolutional layer. This design can increase the model's complexity, potentially leading to overfitting issues.

Ullah et al.'s [36] approach leveraged transfer learning with pre-trained models, yielding good

results on both the hockey fight and the crowd violence datasets. However, the limitation of this method is its dependency on pre-trained models, which might not always align with the specific feature characteristics of the violence detection task.Sudhakaran and Lanz's [37] architecture combined ConvLSTM with 2D CNNs, achieving an ACC of 97.1% on the Hockey Fight dataset and 94.5% on the Crowd Violence dataset. However, the usage of 2D CNNs might not fully capture the spatiotemporal information in the videos, leading to potential detection errors. Lastly, Hanson et al.'s [38] strategy integrated VGG13 CNN with a ConvLSTM layer, although achieving high ACC on the datasets, it is predicated on a specific model, VGG13, thus limiting the generalizability of their approach.

Given these identified limitations, our work aims to provide a solution that both enhances the ACC of violence detection and improves upon the weaknesses of these conventional techniques. Our proposed approach considers the spatial, temporal, and spatiotemporal characteristics of violent behavior, offering a more comprehensive and efficient solution for violence detection in machine learning.

3. Proposed model

This section is describing the proposed model for violence recognition using ML (see Fig. 1). The processes involved in image processing are discussed initially. Following that is a discussion of the characteristics of extraction-based principal component analysis (PCA), followed by an illustration of each of the four machine learning methods.

3.1 Image Processing:

Before a picture is delivered to the model, it must go through five distinct processes. Specifically, the five steps of operation are:

a) Data set

The 350 clips in the dataset are all videos in the MP4 format (using the H.264 codec), with an average length of 5.63 seconds (ranging from 2 seconds to 14 seconds). All of the videos have a resolution of 1920 x 1080 and a frame rate of 30 frames per second. The information is organized into two primary folders, "non-violent" and "violent," with each folder containing videos labeled as depicting either violent or non-violent actions. The folder structure consists of two levels, cam1 and cam2:

- There are sixty clips in the non-violent/cam1 folder, all of which depict actions that are not violent.
- The 60 videos in "non-violent/cam2" are identical to those in "non-violent/cam1," except they were shot from a different angle.
- There are 115 clips in the violent/cam1 folder that depict violent actions.
- All 115 clips in "violent/cam2" are identical to those in "violent/cam1" but they were shot from a different camera angle.

Two cameras were set up in two different locations in the same room to capture all the footage in natural light (the top left corner in front of the room door, and the top right corner on the door side) [39]. Fig. 2 displays the
original images.

b) Convert RGB images to grayscale images

Grayscale is the simplest color scheme since it defines hue by lightness rather than hue. 0 (total blackness) to 255 (full brightness) represents light intensity (complete whiteness). Grayscale photos lack information compared to RGB images. Grayscale images are used in image processing due to their space and processing performance advantages.

The final value should increase for green while decreasing for blue as a result of averaging the portions with varying weights. The researchers reached a conclusion found in Eq. (1) [40]. Following much testing and investigation, the authors conclude Photographs taken after this procedure are shown in Fig. 3.

$$New\ grayscale\ image = ((0.3 \times R) + (0.59 \times G) + (0.11 \times B)) \quad (1)$$

Figure 3. Converting color images to grayscale.

c) Histogram Equalization Method

Histogram equalization is a common technique that can be used to enhance the quality of an image. The operation is quite similar to a histogram stretch, although it often produces more pleasing results across a wider range of images. The histogram of the final image will be as flat as practically possible when employing this method, while the general form of the histogram will be preserved when utilizing histogram stretching. Although the black pixels in a photograph cannot appear any darker than they currently are, the

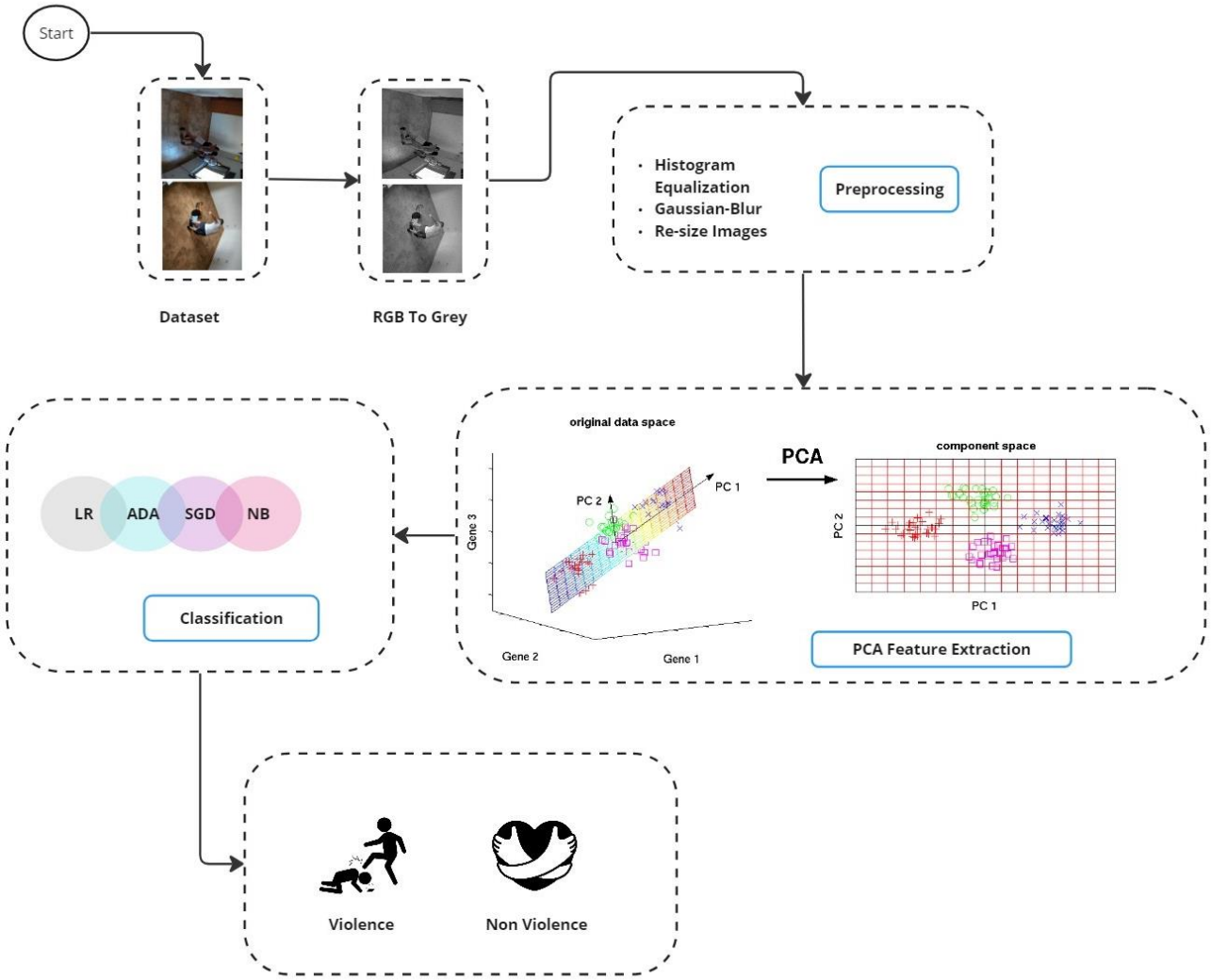

Figure 1. Block diagram of proposed work

Figure. 2 Original images dataset

histogram can be spread or flattened to make the dark pixels appear darker and the light pixels appear lighter (the important word being "appear").Instead, if the somewhat brighter ones become substantially lighter, the black pixels will appear darker [41], [42]. There are four phases to the histogram equalization procedure for digital images:

Step1: Calculate the cumulative histogram values.

Step 2: Divide the values from step 1 by the total number of pixels in order to normalize them.

Step 3: Multiply the numbers from Step 2 by the highest possible gray level, and round.

Step 4: Use a one-to-one mapping of grayscale values to the output of step 3 correspondence.

At this step, Histogram Equalization is determined for every image in the dataset. The histogram was built in the following way using Eq. (2):

$$h[i] = \sum_{x=1}^{N} \sum_{y=1}^{M} \begin{Bmatrix} 0 & if\ f[x.y] = i \\ 1 & otherwise \end{Bmatrix} \quad (2)$$

where, $h[i]$: the result value of the histogram equalization, M and N: grayscale image dimension [42]. Histogram-induced image distortions are depicted in Fig. 4.

Figure 3. Converting color images to grayscale.

Figure 4. The Histogram's effect on grayscale images

Figure 5. Gaussian distribution.

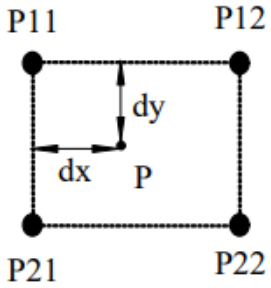

Figure. 6 The 4-dimensional surrounding area of a point 'p' in a 2-dimensional picture space,

d) Gaussian-blur

One technique used to enhance multiscale image structures is the Gaussian blur. Mathematically speaking, blurring an image with a Gaussian filter of size (5×5) is the same as applying a Gaussian function to each individual pixel in the image and then convolving the result. The formula for a one-dimensional Gaussian function is:

$$G(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \quad (3)$$

The standard deviation of a Gaussian distribution can be represented by the horizontal distance x from the center. Fig. 5 helps illustrate the result.

e) Re-size images

Bilinear interpolation resizes images often. Interpolated values are formed by averaging four surrounding pixels, resulting in a smoother image. This method is more accurate than nearest-neighbor interpolation. [43, 44]. Pixels $x_o, x_1, y_o,$ and y_1 are shown in the bilinear interpolar as:

$$y = y_0 \left(1 - \frac{x - x_0}{x_1 - x_0}\right) + y_1 \left(1 - \frac{x_1 - x}{x_1 - x_0}\right) \quad (4)$$

Then the pixels used to make interpolar objects will have their values modified correspondingly. The value between pixels x_o and x_1 is shown as the interpolated value of y, and similarly for pixels $y_o,$ and y_1. Keep in mind that the y-values cannot exceed the $x_o, x_1, y_o,$ and y_1 values. A thermal camera's color scheme is derived from this equation. As can be seen in Fig. 6, the interpolar bilinear coefficient is determined by the horizontal and vertical distances between neighboring pixels, exposing the image to be composed of white and black color patterns separated by grayscale transitions [44].

3.2 Features extraction

a) Principal component analysis

Pattern discovery in high-dimensional data is a popular use of Principal component analysis (PCA). PCA uses fewer Eigenobject feature images to represent common and unknown behavior. PCA's use in recognition technologies has been shown to detect and validate face features. A two-dimensional (2D) matrix of facial photographs must be converted into a one-dimensional (1D) vector for PCA. The row or column orientation of a one-dimensional vector has no effect. [45]. The PCA space can be condensed by picking the first l principle component axes, q1 through ql. After then, data can be superimposed onto this l-dimensional area.

Eigenvectors which is relying on the PCA techniques the strategy of the eigenvectors method consists of extracting the characteristic features from the frame and representing the violence or non-violence shape in question as a linear combination of the so called ‘eigenvector’ obtained from the feature extraction process [46].

$$Av = \frac{1}{M} \sum_{n=1}^{M} Training\ images\ (n) \quad (5)$$

$$Cov = \sum_{n=1}^{M} sub(n)\, sub^{T}(n) \quad (6)$$

Where Av is the average of training images frames,
Cov represents the covariance of the images
M: is the Training set of total images
μ: represent the average Mean
Sub: represent the subtracted image from the average μ.

3.3 Classification model

Machine learning is a method that can be used to automate the procedure of generating analytical models for data analysis. Subset of AI based on the premise that computers can learn independently from data, recognize patterns, and make judgments with minimum human involvement. This study employs four algorithms—the NB, SGD, LR, and ADA—and the following sections elaborate on each:

Naive Bayes NB: That's a supervised learning algorithm. developing a set of probabilities by counting occurrences of certain data and the quantity of relevant collections. To calculate the posterior probability that a given document d belongs to class c using the naive bayes method, we can use the following Eqs. (7) and (8) [45].

$$P(c|d) = \frac{P(d|c)\, P(c)}{p(d)} \quad (7)$$

$$P(c|d) = \frac{P(w1,w2,\ldots wn|c)\, p(c)}{p(d)} \quad (8)$$

Where: P(d|c) is the likelihood which is the probability of the predictor given class.

P (wi |c) is the qualified probability of term wi taking place in document d of class c. P (wi |c) denotes to a measure of how much wi contributes, and c is the accurate class. (w1, w2, ..., wn) are the tokens in document d and are part of the vocabulary used for classification, and n is the number of such tokens in document d.

Logistic Regression LR: It is categorized as a log-linear or exponential classifier. Naive bayes tends to support LR. After a log-linear classifier is applied to the input data, a set of weighted highlights is obtained, and the logs are combined linearly. Logistic regression is defined as:

$$y = \frac{1}{(1 + e^{-(b_0 + b_1 \times x_1 + b_2 \times x_2)})} \quad (9)$$

Where b_0, b_1 and b_2 are the coefficients, x_1 and x_2 are the features (input value) [47].

Stochastic gradient descent (SGD): The model is good for linear classifier learning. One can approximate the gradient with a less exact estimate. The stochastic (or "operational") gradient descent approach approximates the cost function gradient by giving each learning element a gradient. To account for expected shifts, many parameter adjustments were made. Its parameters were adjusted whenever training data was added. When used to massive datasets, stochastic gradient descent outperforms the standard technique.[48]. There is a high degree of success with this kind of facilitation. Simple-to-understand SGD revisions are shown in Eq. (10) as follows:

$$\theta^{(t+1)} = \theta^{(t)} - \alpha_t \nabla l_i\left(\theta^{(t)}\right) \quad (10)$$

The number of iterations is t, and and reflect the size of the learning set used to fine-tune the parameters. Each iteration will randomly assign a new value to index i. Actually, randomizing samples before assessing them is standard practice.[49] .

Adaptive boosting Utilizing an adaptive boosting ML approach improved classification results. When coupled with other learning algorithms, meta-algorithms can improve performance. It adapts to misclassified cases by modifying subsequent classifiers. Thus, ADA uses a weak classifier and adjusts each example's weight with each call. In this method, the misclassified samples will be weighted higher than the correctly classified ones, prompting the new classifier to favor them [50, 51].

4. Experimental work

In order to identify which of the proposed methods for violence detection was most effective, researchers conducted scientific experiments. A 70% training set and a 30% test set are used to evaluate the data. It is a major challenge in today's digital world to manage violence detection. For the problem of violence detection systems, numerous categorization algorithms have been utilized. There is a multi-step process involved in determining whether or not violence has occurred. Classification classifiers were evaluated based on models trained with each dataset. The desired model relies on a machine learning classifier that was trained on a wide variety of methods and large data sets. An additional python model is developed using the ML library.

5. Performance evaluations of classification algorithm

At this phase, the model's classification ACC must be determined. Comparing predicted and observed class labels lets us assess the classifier's

performance. Classification ACC can be assessed by counting the number of accurately identified class instances (true positives), well-recognized examples that were not class-relevant (true negatives), and examples that were either mistakenly categorized (false positives) or not classified at all (false negatives). Quantitative analysis involves these steps [52]:

A. Accuracy: is defined as the fraction of predictions that came true, as indicated in Eq. (11), and is also used to define the model's success.

$$Accurcy = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

B. Precision: stands for the fraction of positive results that can be sorted into the true positives and false positives categories. For example, in Eq. (12). True positives (TP) and false positives (FP) are abbreviations for these two types of results.

$$Precision = \frac{TP}{TP+FP} \quad (12)$$

C. Recall: The number of supposedly critical and relevant data points relied upon by this model determines how reliable it actually is. In this context, "true positives" (TP) are the desired outcome, whereas "false negatives" (FN) are the unwanted side effects.

$$Recall = \frac{TP}{TP+FN} \quad (13)$$

D. F1 Score: The F1 metric measures how quickly ACC and recall are in step with one another.

$$F_1 = 2 \times \frac{precision \times recall}{precision+recall} = \frac{2TP}{2TP+FP+FN} \quad (14)$$

If F1 is high, the system is generally running smoothly.

E. Specificity: It is the portion of negatives that are correctly identified over all the available negatives.

$$Specificity = \frac{TN}{TN+FP} \quad (15)$$

6. Results and discussion

We present in this subsection the metrics grown by the developed model on the AIRTLab dataset. The results of using the model and the other algorithms covered earlier are demonstrated below. When components like feature extraction, feature processing, and violence categorization are incorporated, a better result with more accurate output is generated. The SGD and LR methods are also the most precise among several algorithms. Using the additional attributes detailed in Table 1.

Fig. (7) demonstrate a comparison from prespective of the accuracy. However, the experimental study found that out of the four ML algorithms used to anticipate the violence actions from non-violent, Results showed that the suggested model of LR accurately predicted violence detection on the AIRTLab dataset with an accuracy of 98.31%. As we see in Table 1, our proposed LR method outperforms other machine learning algorithms such as NB, SGD, and ADA on key measures including ACC, precision, recall, and F1-score. This is largely due to LR's ability to estimate the probabilities using a logistic function, which can be particularly effective in binary classification problems like violence detection. However, an in-depth discussion is warranted to fully understand the scientific contribution of this research. While the NB algorithm shows relatively lower performance in comparison, it's important to note that this may be attributed to its underlying assumption of feature independence, which can be a strong assumption for real-world applications like violence detection. SGD, on the other hand, closely matches the performance of the LR model. However, SGD may not always converge to the optimum and its performance heavily depends on the careful tuning of hyperparameters, unlike LR which is simpler to optimize.

In the case of ADA, although it performs reasonably well, it is still outperformed by Looking at Table 2, our proposed method also excels when compared to existing techniques such as ConvLSTM [53], BrutNet [54], and a combination of CNN and Bi-LSTM [55]. However, it's critical to understand the principles underlying these techniques to comprehend why our proposed method offers superior results.

ConvLSTM [53] incorporates the advantages of CNN in image processing and the capability of LSTM in sequence prediction. Despite its strong performance in capturing spatiotemporal features, its complexity can increase computational cost

Table 1. Measurements taken for the AIRTLab dataset across various ML classifiers

Algorithm	Accuracy	Precision	Recall	F1-score	Mean Sensitivity	Mean Specificity	AUC
LR	0.9831	0.98	0.98	0.98	0.9832	0.98	0.98
NB	0.7007	0.73	0.70	0.70	0.7067	0.70	0.71
SGD	0.9803	0.98	0.98	0.98	0.9790	0.97	0.98
ADA	0.8925	0.89	0.89	0.89	0.89	0.89	0.89

Table 2. Accuracy measurements taken from the AIRTLab dataset and compared across various architectures

Ref.	Algorithms	Accuracy	F1-Score
Sernani et al,[53]	ConvLSTM	97.15	97.89
Haque et al,[54]	BrutNet	97.22	N/A
Siddique et al,[55]	CNN Bi-LSTM	83.33	0.89
Proposed	LR	98.31	0.98

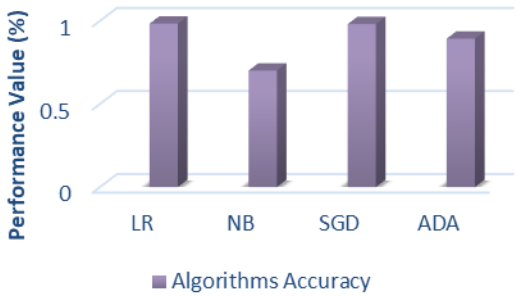

Figure. 7 Accuracy measurement for classifier performance evaluation

BrutNet [54] represents an ensemble of several neural networks, where each member contributes to the final decision, achieving robust performance. However, the absence of reported F1-Score makes it difficult to compare its performance on a balanced measure.

The combination of CNN and Bi-LSTM [55] offers an approach to capture both spatial and temporal information. However, their ACC and F1-score are lower than our proposed method, possibly due to the inherent complexity of integrating these two architectures which can make the model prone to overfitting and harder to train.

The introduction and comparison of these existing techniques can seem abrupt without a proper theoretical background provided in previous sections. Hence, it's essential to delve into the fundamental differences in their methodology to truly appreciate the superior performance of our proposed LR method in the context of violence detection.

The ML technique performs fairly well considering its precision, recall, and F1-score to construct the model. While the outcomes of using the proposed methodologies are encouraging, a more in-depth analysis reveals some interesting and significant statistics that might be used to make better decisions about which ML algorithms to use and how effective those choices are. Insightful analysis and helpful details are included below:

[1] The proposed model was found to provide the best all-around performance in terms of ACC, precision, recall, and F1-score when compared to other methods.

[2] Out of four different machine learning methods—Bayesian modelling techniques, stochastic gradient descent, adaptive boosting, and logistic regression—the SGD and LR algorithms performed the best. It's important to highlight the fact that the selected ML performance is extremely efficient at violence detection in computer vision when employing ML techniques.

[3] To evaluate the efficacy of ML algorithms, the researchers focused on just four metrics: ACC, precision, recall, and F1-score. Yet, the model's precision and development time are essential. In terms of the specified violence detection, the utilized ML methods all carry out admirably.

7. Conclusion

The use of data mining algorithms is absolutely necessary for the classification of enormous datasets in an effective manner. This study looks at a total of four different machine learning approaches, all with the intention of identifying instances of violent behavior. When comparing the outcomes obtained using SGD and LR, it would appear that they are on par with one another. In contrast, the ACC achieved by the ADA algorithm is superior than that achieved by the NB technique. The research concluded that the

proposed model had a precision of 98 % when it came to accurately predicting the presence of violence by making use of the AIRTLab dataset. The fact that various methods of machine learning produce such a wide range of results is suggestive of the idea that the ACC of the model could be improved. This argument is given more weight by the fact that there is a disparity between the score on the Recall test and the score on the F1 test. The researchers are looking forward to broadening the scope of the study in the future in the hopes of developing a hybrid detector that is able to identify instances of violent behavior.

Conflicts of interest

Authors have no conflict of interest to declare.

Author contributions

S. Z. Khalaf has made a paper conceptualization, methodology, investigation, writing-original draft preparation, and system development. **M. I. Shujaa** made a work supervision, project administration, and final copy revision. **A. A. Alwahhab** was responsible to do model validation, formal analysis, resources preparation, and results visualization.

References

[1] S. Caraiman, A. Morar, M. Owczarek, A. Burlacu, D. Rzeszotarski, N. Botezatu, P. Herghelegiu, F. Moldoveanu, P. Strumillo, and A. Moldoveanu, "Computer vision for the visually impaired: the sound of vision system", In: *Proc. of the IEEE International Conf. on Computer Vision Workshops (ICCVW)*, Venice, Italy, pp. 1480-1489, 2017.

[2] D. R. Beddiar, B. Nini, M. Sabokrou, and A. Hadid, "Vision-based human activity recognition: A survey", *Multimedia Tools and Applications*, Vol.79, pp. 30509-30555, 2020.

[3] B. Bansal, "Gesture recognition: A survey", *International Journal of Computers Applications*, Vol.139, No. 2, pp. 0975-8887, 2016.

[4] G. Varol, I. Laptev, and C. Schmid, "Long-Term Temporal Convolutions for Action Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40, No. 6, pp. 1510-1517, 2018.

[5] C. Finn, I. Goodfellow, and S. Levine, "Unsupervised learning for physical interaction through video prediction", In: *Proc. of International Conf. on Neural Information Processing Systems (NIPS)*, Barcelona, Spain, pp. 64–72, 2016.

[6] R. K. Tripathi, A. S. Jalal, and S. C. Agrawal, "Suspicious human activity recognition: A review", *Artificial Intelligence Review*, Vol. 50, No. 2, pp.283-339, 2018.

[7] T. Hao, D. Wu, Q. Wang, and J. S. Sun, "Multi-view representation learning for multi-view action recognition", *Journal of Visual Communication and Image Representation*, Vol. 48, pp.453-460, 2017.

[8] R. Nayak, U. C. Pati, and S. K. Das, "A comprehensive review on deep learning-based methods for video anomaly detection", *Image and Vision Computing*, Vol. 106, p. 104078, 2021.

[9] A. B. Mabrouk, and E. Zagrouba, "Abnormal behavior recognition for intelligent video surveil lance systems: A review", *Expert Systems with Applications*, Vol. 91, pp. 480-491, 2018.

[10] C. Mu, J. Xie, W. Yan, T. Liu, and P. Li, "A fast recognition algorithm for suspicious behavior in high-definition videos", *Multimedia Systems*, Vol. 22, No. 3, pp. 275-285, 2016.

[11] Y. Feng, Y. Yuan, and X. Lu, "Learning deep event models for crowd anomaly detection", *Neurocomputing*, Vol. 219, pp. 548-556, 2017.

[12] S. Zhu, J. Hu, and Z. Shi, "Local abnormal behavior detection based on optical flow and spatio-temporal gradient", *Multimedia Tools and Applications*, Vol. 75, No. 15, pp. 9445-9459, 2016.

[13] R. Sharma and A. Sungheetha, "An efficient dimension reduction-based fusion of CNN and SVM model for detection of abnormal incident in video surveillance", *Journal of Soft Computing Paradigm (JSCP)*, Vol. 03, No. 02, pp. 55-69, 2021.

[14] Z. Xia, J. Xing, and X. Li, "Gesture tracking and recognition algorithm for dynamic human motion using multimodal deep learning", *Security and Communication Networks*, Vol. 2022, p. 11, 2022.

[15] Y. Wang, S. Yang, F. Li, Y. Wu, and Yu. Wang, "FallViewer: A fine-grained indoor fall detection system with ubiquitous Wi-Fi devices", *IEEE Internet of Things Journal*, Vol. 8, No. 15, pp. 12455-12466, 2021.

[16] G. Sun and Z. Wang, "Fall detection algorithm for the elderly based on human posture estimation", In: *Proc. of Asia-Pacific Conf. on Image Processing, Electronics and Computers (IPEC)*, Dalian, China, pp. 172-176, 2020.

[17] D. K. Vishwakarma and C. Dhiman, "A unified model for human activity recognition using spatial distribution of gradients and difference of

Gaussian kernel", *The Visual Computer*, Vol. 35, No. 11, pp. 1595-1613, 2019.

[18] N. Zhu, G. Zhao, X. Zhang, and Z. Jin, "Falling motion detection algorithm based on deep learning", *IET Image Processing*, Vol. 16, No. 11, pp. 2845-2853, 2022.

[19] S. D. Bansod and A. V. Nandedkar, "Crowd anomaly detection and localization using histogram of magnitude and momentum", *The Visual Computer*, Vol. 36, No. 3, pp. 609-620, 2020.

[20] U. Demir, Y. S. Rawat, and M. Shah, "Tinyvirat: Low-resolution video action recognition", In: *Proc. of International Conf. on Pattern Recognition (ICPR)*, Milan, Italy, pp. 7387-7394, 2021.

[21] P. Mantini and S. K. Shah, "Multiple people tracking using contextual trajectory forecasting", In: *Proc. of IEEE conf. on Technologies for Homeland Security (HST), Massachusetts Area*, USA, pp. 1-6, 2016.

[22] C. G. Maldonado, S. H. Mendez, A. L. Solis, H. V. Figueroa, and A. M. Hernandez, "The Effects of Using a Noise Filter and Feature Selection in Action Recognition: An Empirical Study", In: *Proc. of International Conf. on Mechatronics, Electronics and Automotive Engineering (ICMEAE)*, Cuernavaca, Mexico, pp. 43-48, 2017.

[23] A. Dapogny, K. Bailly, and S. Dubuisson, "Confidence-Weighted Local Expression Predictions for Occlusion Handling in Expression Recognition and Action Unit Detection", *International Journal of Computer Vision*, Vol. 126, Nos. 2–4, pp. 255-271, 2018.

[24] G. Stratou, A. Ghosh, P. Debevec, and L. P. Morency, "Effect of illumination on automatic expression recognition: A novel 3D relight able facial database", In: *Proc. of International Conf. on Automatic Face and Gesture Recognition*, California, USA, pp. 611-618, 2011.

[25] A. Deshpnande and K. K. Warhade. "An Improved Model for Human Activity Recognition by Integrated feature Approach and Optimized SVM", In: *Proc. of International Conf. on Emerging Smart Computing and Informatics (ESCI)*, Pune, India, pp. 571-576, 2021.

[26] M. Shen, X. Jiang, and T. Sun, "Anomaly detection based on Nearest Neighbor search with Locality-Sensitive B-tree", *Neurocomputing*, Vol. 289, pp. 55-67, 2018.

[27] B. Kang and T. Q. Nguyen, "Random forest with learned representations for semantic segmentation", *IEEE Transactions on Image Processing*, Vol. 28, No. 7, pp. 3542-3555, 2019.

[28] S. J. Berlin and M. John, "Particle swarm optimization with deep learning for human action recognition", *Multimedia Tools Applications*, Vol. 79, Nos. 25–26, pp. 17349-17371, 2020.

[29] T. Z. Ehsan and S. M. Mohtavipour, "Vi-Net: A deep violent flow network for violence detection in video sequences", In: *Proc. of International Conf. on Information and Knowledge Technology (IKT)*, Tehran, Iran, pp. 88-92, 2020.

[30] N. Jaouedi, N. Boujnah, and M. S. Bouhlel, "A new hybrid deep learning model for human action recognition", *Journal of King Saud University-Computer and Information Sciences*, Vol. 32, No. 4, pp. 447-453, 2020.

[31] M. Ramzan, A. Abid, H. U. Khan, S. M. Awan, A. Ismail, M. Ahmed, M. Ilyas, and A. Mahmood, "A Review on State-of-the-Art Violence Detection Techniques", *IEEE Access*, Vol.7, pp. 107560-107575, 2019.

[32] M. Cheng, K. Cai, and M. Li, "RWF-2000: An open large scale video database for violence detection", In: *Proc. of International Conf. on Pattern Recognition*, Milan, Italy, pp. 4183-4190, 2020.

[33] C. Ding, S. Fan, M. Zhu, W. Feng, and B. Jia, "Violence detection in video by using 3D convolutional neural networks", In: *Proc of International Symposium on Visual Computing: Advances in Visual Computing*, Las Vegas, USA, pp. 551-558, 2014.

[34] W. Song, D. Zhang, X. Zhao, J. Yu, R. Zheng, and A. Wang, "A Novel Violent Video Detection Scheme Based on Modified 3D Convolutional Neural Networks", *IEEE Access*, Vol. 7, pp. 39172-39179, 2019.

[35] J. Li, X. Jiang, T. Sun, and K. Xu, "Efficient violence detection using 3D convolutional neural networks", In: *Proc. of IEEE International Conf. on Advanced Video and Signal Based Surveillance (AVSS)*, Taipei, Taiwan, pp. 1-8, 2019.

[36] 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, 2019.

[37] S. Sudhakaran, and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory", In: *Proc. of IEEE International Conf. on Advanced Video and Signal Based Surveillance (AVSS)*, Lecce, Italy, pp. 1-6, 2017.

[38] A. Hanson, K. Pnvr, S. Krishnagopal, and L. Davis, "Bidirectional convolutional LSTM for the detection of violence in videos", In: *Proc. of European Conf. on Computer Vision*, Munich, Germany, pp. 0-0, 2018.

[39] M. Bianculli, N. Falcionelli, P. Sernani, S. Tomassini, P. Contardo, M. Lombardi, and A. F. Dragoni "A dataset for automatic violence detection in videos", *Data Brief*, Vol. 33, p. 106587, 2020.

[40] S. A. Alrubaie and A. H. Hameed, "Dynamic Weights Equations for Converting Grayscale Image to RGB Image", *Journal of University of Babylon for Pure and Applied Sciences*, Vol. 26, No. 8, pp. 122-129, 2018.

[41] B. Oktavianto and T. W. Purboyo, "A Study of Histogram Equalization Techniques for Image Enhancement", *International Journal of Applied Engineering Research*, Vol. 13, No. 2, pp. 1165-1170, 2018.

[42] S. Dubey and M. Dixit, "Image Enhancement Techniques: An Exhaustive Review", In: *Proc. of International Conf. on Sustainable and Innovative Solutions for Current Challenges in Engineering Technology: Intelligent Computing Applications for Sustainable Real-World Systems*, Gwalior, India, pp. 363-375, 2020.

[43] P. Parsania, and P. V. Virparia, "A comparative analysis of image interpolation algorithms", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 5, No. 1, pp. 29-34, 2016.

[44] I. W. Adiyasa, A. P. Prasetyono, A. Yudianto, P. P. W. Begawan, and D. Sultantyo, "Bilinear interpolation method on 8x8 pixel thermal camera for temperature instrument of combustion engine", In: *Proc. of International Conf. on Vocational Education of Mechanical and Automotive Technology*, Yogyakarta, Indonesia, p. 012076, 2020.

[45] S. Phauk and T. Okazaki, "Hybrid machine learning algorithms for predicting academic performance", *International Journal of Advanced Computer Science and Applications*, Vol. 11, No. 1, pp. 32–41, 2020.

[46] S. Karamizadeh, S. M. Abdullah, A. A. Manaf, M. Zamani, and A. Hooman. "An overview of principal component analysis", *Journal of Signal and Information Processing*, Vol. 4, No.3B, pp. 173-175, 2013.

[47] H. R. Pourghasemi, A. Gayen, S. Park, C. W. Lee, and S. Lee, "Assessment of landslide-prone areas and their zonation using logistic regression, LogitBoost, and naïvebayes machine-learning algorithms", *Sustainability*, Vol. 10, No. 10, p. 3697, 2018.

[48] Y. Zhang, A. M. Saxe, M. S. Advani, and A. A. Lee, "Energy–entropy competition and the effectiveness of stochastic gradient descent in machine learning", *Molecular Phsicys*, Vol. 116, Nos. 21–22, pp. 3214–3223, 2018.

[49] A. Razaque and A. M. Alajlan, "Supervised machine learning model-based approach for performance prediction of students", *Journal of Computer Science*, Vol. 16, No. 8, pp. 1150–1162, 2020.

[50] A. Taherkhani, G. Cosma, and T. M. McGinnity, "AdaBoost-CNN: An adaptive boosting algorithm for convolutional neural networks to classify multi-class imbalanced datasets using transfer learning", *Neurocomputing*, Vol. 404, pp. 351–366, 2020.

[51] A. Shahraki, M. Abbasi, and Ø. Haugen, "Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost", *Engineering Applications of Artificial Intelligence*, Vol. 94, p. 103770, 2020.

[52] H. M. Fadhil, M. N. Abdullah, and M. I. Younis, "A Framework for Predicting Airfare Prices Using Machine Learning", *Iraqi Journal of Computers*, Vol. 22, No. 3, 2022.

[53] 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.

[54] M. Haque, S. Afsha, and H. Nyeem, "Developing BrutNet: A New Deep CNN Model with GRU for Realtime Violence Detection," In: *Proc. of 2022 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, Chattogram, Bangladesh, pp. 390–395, 2022.

[55] L. A. Siddique, R. Junhai, T. Reza, S. S. Khan, and T. Rahman, "Analysis of Real-Time Hostile Activitiy Detection from Spatiotemporal Features Using Time Distributed Deep CNNs, RNNs and Attention-Based Mechanisms", In: *Proc. of IEEE International Conf. on Image Processing Applications and Systems (IPAS)*, Genova, Italy, pp. 1-6, 2022.