International Journal of Safety and Security Engineering
Vol. 13, No. 1, February, 2023, pp. 51-58
Journal homepage: http://iieta.org/journals/ijsse

Normalized Attention Neural Network with Adaptive Feature Recalibration for Detecting the Unusual Activities Using Video Surveillance Camera

Vijay Kumar Damera[1], Ramesh Vatambeti[2*], M S Mekala[3], Alok Kumar Pani[4], Chinthakunta Manjunath[4]

[1] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad 500090, India
[2] School of Computer Science and Engineering, VIT-AP University, Vijayawada 522237, India
[3] School of Communication Engineering, Yeungnam University, Republic of Korea, Gyeongsan 38541, Korea
[4] Department of Computer Science and Engineering, CHRIST (Deemed to be University), Bangalore 560074, Karnataka, India

Corresponding Author Email: ramesh.v@vitap.ac.in

https://doi.org/10.18280/ijsse.130106

Received: 9 November 2022
Accepted: 11 January 2023

Keywords:
surveillance data, unusual activities, adaptive feature recalibration, normalized attention network

ABSTRACT

Over the past few years, surveillance cameras have become common in many homes and businesses. Many businesses still employ a human monitor of their cameras, despite the fact that this individual is more probable to miss some anomalous occurrences in the video feeds owing to the inherent limitations of human perception. Numerous scholars have investigated surveillance data and offered several strategies for automatically identifying anomalous occurrences. Therefore, it is important to build a model for identifying unusual occurrences in the live stream from the security cameras. Recognizing potentially dangerous situations automatically so that appropriate action may be taken is crucial and can be of great assistance to law enforcement. In this research work, starting with an MRCNN for feature extraction and AFR for fine-tuning, this architecture has a number of key components (AFR). To increase the quality of the features extracted by the MRCNN, the AFR replicas the inter-dependencies among the features to enhance the quality of the low- and high-frequency features extracted. Then, a normalized attention network (NAN) is used to learn the relationships between channels, which used to identify the violence and speeds up the convergence process for training a perfect. Furthermore, the dataset took real-time security camera feeds from a variety of subjects and situations, as opposed to the hand-crafted datasets utilized in prior efforts. We also demonstrate the method's capability of assigning the correct category to each anomaly by classifying normal and abnormal occurrences. The method divided the information gathered into three primary groups: those in need of fire protection, those experiencing theft or violence, and everyone else. The study applied the proposed approach to the UCF-Crime dataset, where it outperformed other models on the same dataset.

1. INTRODUCTION

Surveillance systems based on video cameras are a cost-effective and efficient way to keep goods, property, and people safe. Video cameras are employed for long-distance, weather-independent observation of dynamic situations where light, temperature, and visibility are all subject to continual change. There is a global proliferation of public video surveillance systems, which may yield useful and detailed data for a wide range of safety-related tasks [1]. Crime and violence prevention rely heavily on video monitoring, but the necessity of studying hours of material diminishes the opportunity to make quick choices [2]. Several research have been published on the topic of automatically detecting violent incidents in films in an effort to relieve authorities of the load of having to view hours of footage in order to identify occurrences that only last seconds. Recent publications [3-5] emphasised the accuracy of systems in violence discovery, while earlier efforts [6-8] relied on custom-built features and flow forms are hallmarks of outdated approaches of identifying actions. Extraction of spatial-temporal features from films, or features that capture both the spatial information in a single frame and the gesture info in a series of edges, has been demonstrated to be possible with the use of deep learning approaches [9].

Surveillance in a "smart city" may be used for a wide variety of tasks, including as keeping tabs on city traffic [10], identifying structural damage to buildings [11], identifying instances of violence [12], and managing the aftermath of natural catastrophes [13]. The sheer volume of video feeds might potentially overwhelm human operators. For this reason, much work was put into discovering means of automatically processing such data for the purposes of monitoring for anomalous activity and discarding securely unneeded data. In the context of smart city monitoring, the ability to identify acts of violence is a key concern. With a wireless sensor network infrastructure, we can solve the problem at a reasonable cost. It's important to note that such a solution entail in contrast to current methods, the work here presents a distributed algorithm that can operate on sensor nodes and identify violent behaviour based on camera input while also reducing the communication load on nodes with limited CPU resources and bandwidth.

In contrast to current methods, the work presented here provides a distributed, sensor-level algorithm that can identify violent behaviour from camera input while simultaneously

reducing communication overhead by simply transmitting alerts. The most advanced algorithms for this task rely heavily on computer vision techniques including segmentation, tracking, and action identification, all of which need a lot of processing power [14]. The study of identifying violent acts is a special example of human action recognition [15]. Recognizing an action takes considerable effort due to factors including occlusion, size, and busy backgrounds. Furthermore, there are unique challenges associated with violence detection, the first of which is establishing a workable definition of violence. Humans have a hard difficulty differentiating between violent and nonviolent scenarios. Certain physical and social behaviours, such as jogging, dancing, and interacting with other people, might superficially resemble aggressive ones.

Using a supervised learning framework, this research suggests recasting. Our definition of an anomalous occurrence determines whether or not we should be concerned about a wide variety of odd behaviours in the actual world. However, in this paper, we emphasis on the UCF-Crime dataset, which contains a great deal of aberrant, unlawful, and violent behaviour caught on public surveillance cameras and which can cause serious issues for both individuals and the population as a whole. For feature extraction in our suggested model, we turned to MRCNN with AFR. Next, we include the video dataset into the model architecture by including an NAN, that is well-suited to this type of data. After that, the model reports back on whether or not the given video contains any criminal activity. This approach has the potential to boost the efficiency with which humans can detect permanent harm and reduce associated costs. In this research work, starting with an MRCNN for feature extraction and AFR for fine-tuning, this architecture has a number of key components (AFR). To increase the quality of the features extracted by the MRCNN, the AFR replicas the inter-dependencies among the features to enhance the quality of the low- and high-frequency features extracted. Then, a normalized attention network (NAN) is used to learn the relationships between channels, which used to identify the violence and speeds up the convergence process for training a perfect. Furthermore, the dataset took real-time security camera feeds from a variety of subjects and situations, as opposed to the hand-crafted datasets utilized in prior efforts. We also demonstrate the method's capability of assigning the correct category to each anomaly by classifying normal and abnormal occurrences. The method divided the information gathered into three primary groups: those in need of fire protection, those experiencing theft or violence, and everyone else. The study applied the proposed approach to the UCF-Crime dataset, where it outperformed other models on the same dataset. The following is a brief summary of the method's key contribution.

• MRCNN-ARF and NAN are used together to identify anomalies in video surveillance.

• For this study, we employed the UCF-Crime dataset, which is comprised of 13 types of aberrant events depicted in natural settings captured by security cameras.

• We established two by separating the typical scenarios from the out-of-the-ordinary ones in order to gain a deeper comprehension of each anomaly type.

Section 2 of this paper discusses how other studies in the field of anomaly detection in security cameras employ a variety of models, each of which is a sub-model of the overall concept. After that, in Section 3, we detail our suggested model. Section 4 presents to critically assess our work via a series of experiments. Section 5 discusses the conclusion and future scope.

2. RELATED WORKS

Vijeikis et al. [16] introduce an innovative framework for identifying violent incidents captured by security cameras. Using MobileNet V2 as an encoder and LSTM for classification, our proposed model is a U-Net-like network for extracting spatial features. In spite of its computational efficiency, the suggested model performs admirably. Using a complicated safety camera film RWF-2000, the studies demonstrated an regular accuracy of 0.82 2% and an regular precision of 0.81 3%.

Using deep learning architectures, Sahay et al. [17] projected a new method for real-time violence identification in crime scene videos. The determination of this study is to gather real-time crime scene footage from a shadowing scheme and extract features-based classification method. Features have been retrieved and identified from the processed and transformed video frames that were originally the input footage. Its goal is to identify potentially dangerous situations in real time and separate out outliers from the average. Our method is trained and tested using the extensive UCF Crime anomaly dataset to ensure its efficacy. In terms of accuracy (98%), precision (96%), recall (80%t), and F-1 score (78%), the experimental findings show that the projected method performs well in datasets.

One such model is the IVADC-FDRL model presented by Mansour et al. [18], which detects and classifies anomalies in videos combining faster RCNN and deep learning. Anomaly detection and classification are the two main mechanisms of the IVADC-FDRL model described. To begin, we apply the Faster RCNN model as an object sensor using the Residual Network as the baseline model, which finds the outliers as objects. In addition, a DRL model based on deep Q-learning (DQL) is used to categorise the abnormalities that have been found. Extensive testing was dataset to confirm the effective anomaly finding and organization capabilities of the IVADC-FDRL perfect. With a maximum dataset, respectively, the IVADC-FDRL model was shown to outperform the other approaches in the experiments.

Using optical flow and a generative adversarial network (GAN), Alafif et al. [19] offer a method for detecting anomalous behaviour. The article makes three primary contributions. Optical flows are used to first extract the model's dynamic characteristics. Experimental evidence confirms the characteristics' efficacy. We then develop a GAN-based optical flow architecture and employ a transfer learning technique to identify behavioural outliers in large-scale crowd situations. U-Net and Flownet are utilised by the outline to produce and differentiate between typical and atypical crowd actions. In the end, we gather and evaluate a sample of pilgrimage movies showcasing unusual behaviour in a variety of settings. The accuracies for the first, second, and third scenes in UMN and for UCSD range from 99.4 to 97.1% and 97.2 to 89.26%. When applied to the Abnormal Behaviors HAJJ dataset, it reaches a detection accuracy of 79.63 percent in films of massive crowds.

Kim et al. [20] offer a closed-circuit television (CCTV) environment-optimized method and system for real-time surveillance of suspicious activity. Using a combination of pedestrian recognition and monitoring, the suggested system

is able to gather data about pedestrians in real time and identify anomalous behaviours including incursion, loitering, falling down, and aggression. First, it uses an object's coordinates to identify incursion or loitering, and then it uses the object's behaviour pattern to identify a downturn or violent incident. The suggested solution is tested with data from a smart CCTV system provided by the Korea Internet and Security Agency (KISA).

Narejo et al. [21] used our own dataset to train the YOLO V3 "You Only Look Once" object identification perfect. According to the training data, YOLO V3 is superior to both its predecessor, YOLO V2, and the standard CNN. Moreover, since we working transfer learning to train our model. By incorporating this approach into our monitoring system, we may work to prevent senseless killing and hopefully save lives. Our suggested approach isn't only limited to use in ground-breaking security and surveillance robots, though; it can also be integrated into airborne drones to help spot threats like weapons or hazardous materials before they endanger humans.

Mohtavipour et al. [22] offer a unique deep violence detection framework using characteristics extracted by manual labour. A convolutional neural network (CNN) receives these properties as streams of data in three different dimensions: space, time, and space-time. Every video frame sent into the spatial stream educated the network on the environment's patterns. Three successive images were included in the time sequence for studying the effects of differential optical flow modification on aggressive behaviour in motion. Additionally, a unique differential motion energy picture was included as a discriminative feature in the spatio-temporal stream to better describe violent acts. By combination data from several sources, this technique can provide light on various facets of aggressive conduct. The proposed CNN network is trained using a combination of labelled and unlabelled images from the Hockey, Movie, and ViF datasets, which were collected under both packed and empty conditions. In terms of processing time, the testing findings demonstrated that the suggested deep violence finding technique surpassed state-of-the-art research.

3. PROPOSED METHODOLOGY

3.1 Dataset

In this study, we apply the suggested model to the UCF-Crime dataset [23], which contains a large amount of footage from public surveillance cameras documenting anomalous, unlawful, and violent behaviour in settings as diverse as schools, businesses, and streets. This dataset was chosen because its events are representative of those that occur often and in a variety of settings [24-26]. Furthermore, these aberrant behaviours can cause significant issues for both people and society. Several articles depend on artificially constructed datasets or datasets that have a common context and history (such as hockey battle datasets and movie datasets), neither of which is representative of real-world experience. This dataset contains extensive, unedited footage from 13 different types of anomalous occurrences and one type of normal event [27]. In Figure 1, sample images of UCF-Crime dataset is depicted. In our trials, we alienated the data into a training set of 75% and a testing set of 25% so that we could fairly compare our results to those of other researchers in the area.

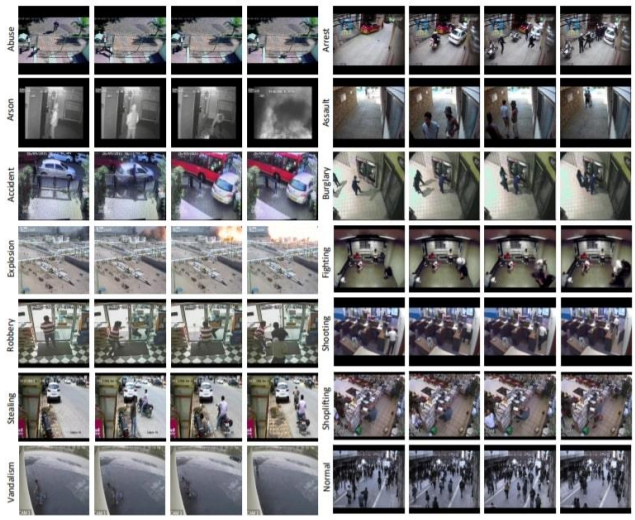

Figure 1. Examples of UCF-Crime dataset

We used the UCF-Crime dataset in its four different iterations: UCF-Crimes, Binary, 4MajorCat, and NREF. Our starting point is a 14-class dataset called UCF-Crimes. Within the confines of the Binary, we group together the 13 outlier occurrences into a single category. We divided the usual occurrences into one category and the anomalous ones into three major ones for the 4MajorCat. Theft, Vandalism, and Violence are the three categories that encompass these actions. Another example of skewed data, the NREF isolates just three outliers.

Table 1. Sum of videos for Dual and Ucfcrimes datasets

Binary	No. Videos	No. Videos	Ucfcrimes
Shoplifting	50	50	Shoplifting
Stealing	100	50	Stealing
Abuse	50	50	Abuse
Assault	50	50	Assault
Explosion	50	50	Explosion
Fighting	50	50	Fighting
Robbery	150	50	Robbery
Shooting	50	50	Shooting
Arrest	50	50	Arrest
Arson	50	50	Arson
Normal	950	50	Normal
Total	1900	700	Total

Table 2. Sum of videos for 4MajCat and NREF datasets

4MajCat	No. Videos	No. Videos	NREF
Theft (Burglary, Theft, Shoplifting, Stealing)	150	30	RoadAccident
Vandalism (Arson, Explosion, RoadAccident, Vandalism)	150	50	Explosion
Ferocity behaviours (Misuse, Arrest,Assault, Fighting, Shooting)	150	70	Fighting
Normal	150	150	Normal
Total	600	300	Total

In this dataset, aberrant films have portions that feature abnormalities judged abnormal, rather than utilising predetermined abnormal and normal videos. On the other hand, the rest of the video is classified as neutral. Also, we

condensed all NREF videos into 10-second clips so that we could extract the most relevant images from each source. Table 1 shows the distribution of video counts across categories in the Ucfcrimes and Binary datasets. In addition, the total sum of videos in the 4MajCat and NREF databases is listed in Table 2.

3.2 Pre-processing

First, we extract individual frames from each video. By counting how many frames in a video file should be ignored, we may determine how many frames can be used to create a subset of n frames. Therefore, if a video file is 60 seconds long, and the video format is set to 30 frames per second, then the total sum of video frames is m=1800. Let's pretend for a moment that n=30 and that, from a total of 1800, you've had to pick only 30. Therefore, once 60 frames have been skipped, we need to pick one. In order to account for spatial mobility in each input, we first selected the frames and then computed the difference between each pair of frames. One set of datasets was pre-processed differently from the others. From the UCF-Crime dataset, we chose to focus on three subsets. We identified the out-of-the-ordinary moment in each video and marked it as a "Anomaly," while the other footage was classified as "Normal." Next, cut all of the videos into equal-sized pieces. Thus, n frames will be chosen from m frames, as was the case before, but now the sum of m is also predetermined. So, in the prior work's agreement, we simply paid attention to the timestamps at which unexpected occurrences occur. In addition, both the anomalous event and the normal set come from the same data source. This implies that in contrast to the initial UCF- dataset, the only difference between the two is the acting, therefore the algorithm is better able to spot out-of-the-ordinary occurrences.

3.3 Feature extraction

In Figure 2, we can see the MRCNN and AFR modules used to extract topographies from the pre-processed frames.

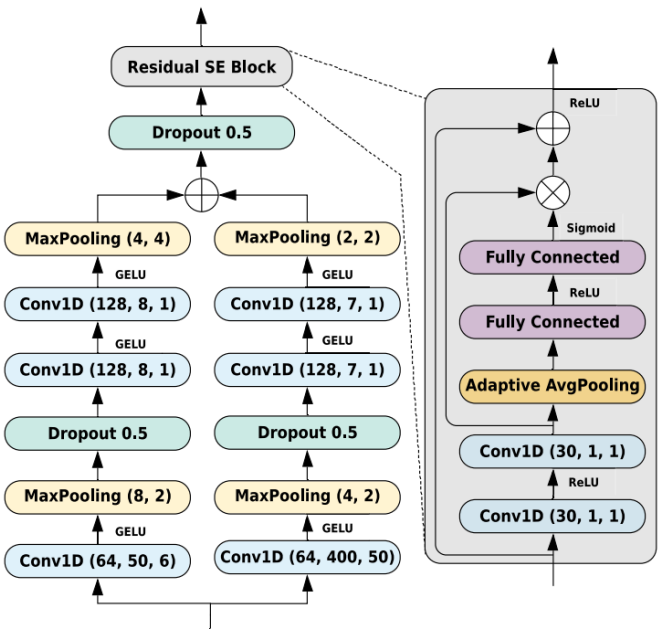

Figure 2. The MRCNN and AFR components for feature extraction

1) Multi-Resolution CNN: In Figure 2, we can see that we've implemented a multi-resolution CNN architecture to facilitate the extraction of a wide variety of features. Specifically, we use two forks of convolutional layers with varying kernel sizes, with the goal of probing various frequency regions based on the sampling rate of the frames. This is motivated by prior efforts that employed several CNN kernel dimensions to extract low- and high-frequency features. It is becoming more crucial to address multiple frequency bands in order to improve the retrieved characteristics, as different phases are characterised by diverse ranges of frequencies. This is why we employ a variety of kernel sizes to imprisonment a wide variety of timesteps and, by extension, characteristics from a wide variety of frames. First, the 400-kernel wide collects 4-second timesteps, allowing for the collection of a complete cycle of a sinusoidal signal at frequencies as low as 0.25 Hz (T=1/F). In this region, you'll find the delta band. Second, the data corresponds to the alpha and theta bands since each convolution window catches 50 samples (0.5 second) with a kernel of 50.

As can be seen in Figure 2, each of the four branches is made up of three max-pooling layers, with the convolution layers employing a batch normalisation layer and a Gaussian Error Linear function. More specifically, the configuration shown in Figure 2 as Conv1D (64, 50, 6) employs a 1D convolution layer with 64 filters, a 50-by-50-pixel kernel, and a 6-pixel stride. In a similar vein, a maxpooling layer with a kernel size of 8 and a stride of 2 is denoted by the notation MaxPooling (8, 2). Overfitting is mitigated by introducing a dropout step following the initial maxpooling in each of the two branches and again following the concatenation of the two branches, as illustrated in Figure 2.

2) Adaptive Feature Recalibration (AFR): To that end, AFR seeks to adjust the MRCNN-learned features for the purpose of enhancing their performance. Through a residual squeeze and excitation block, the AFR replicas the relationships between the features and adaptively chooses the most discriminative ones. The SE block provides a context-aware technique that aids the network's lower levels in making better use of contextual information from beyond their immediate receptive field. We use two Conv1D (30,1,1) convolutions in the residual SE block, setting both the kernel and stride size to 1 and the activation function to ReLU. The MRCNN-learned feature map I R(Ld) is convolved with itself twice, as shown by F=Conv2(Conv1(I)), to get F. where $F = \{F1, \ldots, FN\} \in R^{N \times d}$, N is the total sum of features, d is the distance of $F_i (1 \leq i \leq N)$, and Conv1 and Conv2 are procedures in AFR module.

Then, we use a technique called adaptive average pooling to reduce the size of the world's spatial data. $F \in R^{N \times d}$ to $s=\{s_1, \ldots, s_N\}$, where si is the mean value of the d data points. Two FC layers are then functional to the assembled data for utilisation. Specifically, after the first layer, a ReLU activation function is used to do dimensionality reduction (as mentioned in Equation 1), and after that, layers are utilised to utilise the gathered information. In Eq. (1), for instance, we see that the first layer is shadowed by a ReLU activation function for dimensionality reduction, while the second layer is shadowed by a smoothing sigmoid activation function for dimensionality cumulative.

$$e = \sigma\left(W_2\left(\delta\left(W_1(s)\right)\right)\right) \in R^{N \times d} \quad (1)$$

where, σ and δ mention to sigmoid and ReLU activation purposes correspondingly, and W_1 and W_2 signify the two FC layers in AFR. Then, the feature by e as shadows:

$$O = F \otimes e \in R^{N \times d} \quad (2)$$

where, $\otimes$ means that F was multiplied by e on a point-by-point basis. The upgraded chosen characteristics learnt in the residual SE block are likewise merged with the initial input I via a new, streamlined link. The AFR module's end product is:

$$X = I + O \in R^{N \times d} \quad (3)$$

Please note that the GELU activation function is used in the MRCNN module since it permits certain negative weights of the input. The subsequent AFR module might be affected by these negative weights and make different choices as a result. Given that ReLU converts all negative weights to zero, the AFR module won't be able to use them, GELU is expected to perform better. In the AFR module, however, we employ ReLU to ensure that the gradient does not explode or vanish, as well as to speed up the calculations and make them more reliably converge. While other activation functions, such as Leaky-ReLU and PReLU, also return negative values, GELU may be preferable due to its more flexible implementation. This is because strong negative activations might have an undesired effect on the total of activations feeding the subsequent layers when using these activation functions. However, GELU differs in that it demonstrates more control to limit the impact of these detrimental activations. By using these extracted features, violence or non-violence are detected by using the proposed classifier.

3.4 Classification

To better understand the connection between channels, we propose a normalised attention network that incorporates 1D normalisation techniques into the already-powerful SENet. We also present the normalised attention network for CNN denoising, wherein each channel is given a certain gain and channels have distinct functions in the subsequent convolution. Assuming that the batch size is N, we will create N random values from the noise level range to use as noise standard deviations during training.

The output Y_R is the quantity of the input X_R of features and output Z:

$$Y_R = X_R + Z \quad (4)$$

Eq. (4) can be reformed as $Z=Y_R-X_R$, The effort required to learn Z is referred to as remaining learning. The computational cost of training can be reduced by avoiding the vanishing gradient problem and by the use of learned residues. In the following, we'll discuss NAN Block's structure in further depth.

3.4.1 NAN

Squeeze uses an average pooling function to transform each channel in the input X_N into a point, assuming that there are 96 channels in the X_N. The following is a mathematical formulation of the squeeze operation for each channel X_N^l:

$$X_N^l = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_N^l(i,j) \quad (5)$$

where, $X_N^l(i,j)$ is the amplitude at position (i,j), x_N^l is the typical X_N^l value in a channel. 96 neurons make up the fully linked FC layer. Convergence may be sped up and the vanishing gradient problem solved with 1DBN and Relu layers. If batchsize is set to k, then k samples will be used in each training procedure. In this section, we will discuss the 1DBN.:

$$\tilde{x}_{bn} = \frac{x_{bn} - E(x_{bn})}{\sqrt{var(x_{bn})}} \quad (6)$$

$$y_{bn} = \gamma \tilde{x}_{bn} + \beta \quad (7)$$

where, x_{bn} and y_{bn} are back propagation variables that are changed to reflect changes in the input and output vectors of the BN block. To convert the input to the interval [0,1], the sigmoid function is used on the last layer. Sigmoid function output is denoted by the letter s and is a 96-dimensional vector. as (s_1, s_2, $\cdots$, s_{n-1}, s_n), n is 96. For each channel, the scale process is labeled as shadows:

$$Y_N^m = X_N^m . s_m, \ \ m = 1,2, \dots, n \quad (8)$$

where, Y_N^m and X^m are the mth stations of Y_N and X_N, respectively. s_m is the mth element of s. From Eq. (8), we find every channel/frame in XN is increased by the conforming element of s, which is used to identify the frame is violence or non-violence.

4. RESULTS AND DISCUSSION

The suggested model MRCNN-AFR with NAN was tested with Keras toolkit. Our model has a learning rate of 104. Although the epochs are set to 50, the code terminates when the loss function converges and does not improve when it is let to run further.

4.1 Performance metrics

There has to be an understanding of the underlying truth value in order to make a correct assessment of the various statistical measures. The standard for binary classification was a collection of connection registers that were either typical or indicative of an assault. Where L and A are the total number of normal and attack logs in the test dataset, respectively, these expressions will be used to evaluate the quality of the classification model. To make meaningful comparisons, every statistical indicator needs an underlying truth value. The reality behind binary classification was built on a set of connection registers that were either safe or vulnerable to assault. For the purpose of this evaluation, let's use the notation L and A to stand in for the test dataset's normal and attack logs, respectively.

True Positive (TP) - the quantity of connection records appropriately categorised to the Usual class.

False Negative (FN) - the total quantity of misclassified Attack connections added to the total number of regular connections.

True Negative (TN) - the total quantity of links that have been identified as Attacks.

False Positive (FP) - incorrectly grouped as an Attack connecting record when added to the aggregate of Normal linking records.

Accuracy: An estimate is made of how many records in the

test dataset have predictable connection information compared to the total number of records. The quality of an ML model increases with increasing accuracy.

Precision: The percentage of attachment logs that were successfully detected is estimated as a percentage of total attachment logs. If the ML model has a greater level of accuracy, it is better.

False Positive Rate (FPR): The ratio of standard connection records to regular linking records is used to determine the severity of assaults. Model performance for ML will increase with the reduced FPR.

Table 3 offerings the comparative analysis of projected model for binary classification, where the existing techniques are implemented with this dataset and results are averaged.

Table 3. Binary UCG-Crimes datasets

Algorithm	Accuracy (%)	Recall (%)	Precision (%)	F-score (%)
SVM	87.70	85.21	99.41	91.82
RF	81.10	75.91	99.41	86.72
MobileNet	92.50	90.98	99.82	95.27
RCNN	92.90	99.78	91.52	95.41
GAN	92.50	99.63	91.38	95.18
CNN	92.70	99.90	91.93	95.32
YOLO	93.27	99.91	92.47	95.63
Proposed Model	**94.32**	**99.95**	**93.24**	**96.02**

In the above Table 3 signifies that the Binary classification on UCF-crimes datasets performance. We have evaluated the model by using different techniques such as MobileNet, CNN, RF, SVM and GAN etc. By this comparison analysis the projected model reached the better results than other comparing methods. For instance, the projected model achieved 94% of accuracy, YOLO achieved 93% of accuracy, MobileNet, RCNN, CNN and GAN achieved 92% of accuracy and ML techniques achieved only 86% of accuracy. In the analysis of recall, the proposed model achieved 93%, where existing DL techniques achieved nearly 90% to 91% and existing ML techniques achieved only 75% and 85%. The analysis of F-score describes that the proposed model achieved 96%, ML techniques achieved 86% and existing DL techniques achieved nearly 95%. Figures 3 and 4 present the graphical comparison.

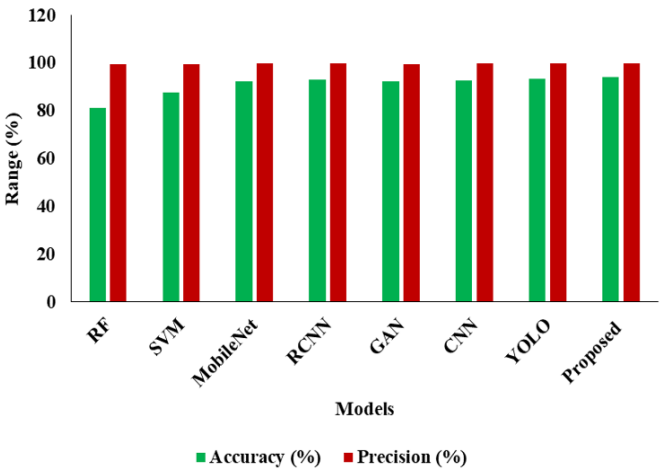

Figure 3. Accuracy and precision comparison

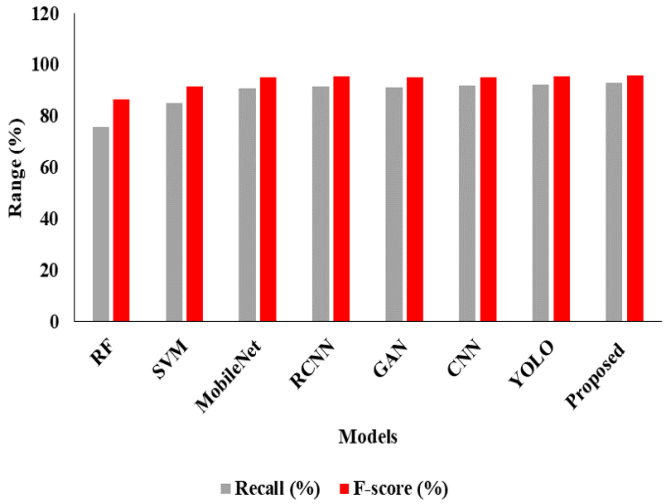

Figure 4. Recall and F-score comparison

Table 4. 4MajCat and NREF datasets

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)
RF	90.82	92.43	89.79	93.57
SVM	92.44	93.12	90.41	95.13
MobileNet	92.13	93.26	92.27	95.83
RCNN	94.63	96.16	92.66	97.27
GAN	93.12	95.31	92.55	96.90
CNN	94.82	97.82	93.50	97.49
YOLO	95.39	97.23	94.11	98.48
Proposed Model	**96.87**	**99.65**	**95.39**	**99.52**

In the above Table 4, it signifies that the 4MajCat and NREF datasets performance. We have evaluated the model by using different techniques. By this comparison analysis the proposed model reached the better results than other comparing methods. In this analysis, the proposed model achieved 96% of accuracy, 99% of precision, 95% of recall and 99% of F-score. However, the existing DL techniques achieved nearly 94% of accuracy, 96% of precision, 92% of recall and 97% of F-score and existing ML techniques achieved nearly 91% of accuracy, 93% of precision, 90% of recall and 94% of F-score. Figures 5 and 6 present the comparative analysis in graphs.

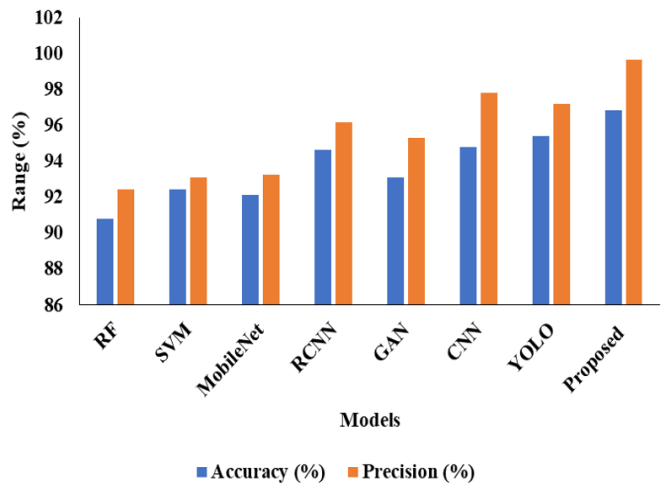

Figure 5. Accuracy and precision comparison

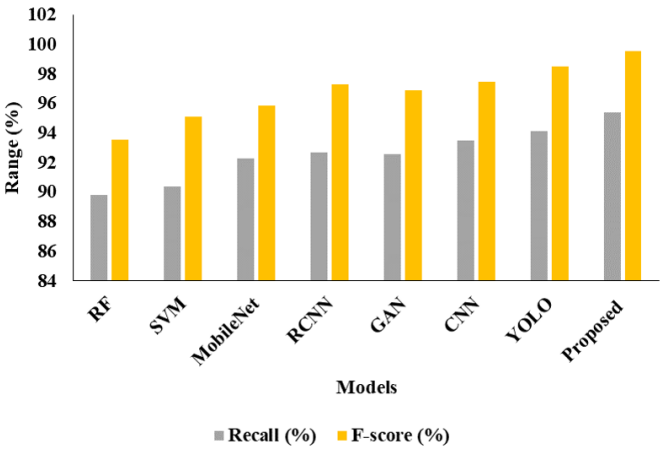

Figure 6. Recall and F-score comparison

5. CONCLUSIONS

In order to identify out-of-the-ordinary actions in the UCF-Crime dataset, this research builds a unique framework that combines MRCNN-AFR with NAN. In this research work, starting with an MRCNN for feature extraction and AFR for fine-tuning, this architecture has a number of key components (AFR). To increase the quality of the features extracted by the MRCNN, the AFR replicas the inter-dependencies among the features to enhance the quality of the low- and high-frequency features extracted. Then, a normalized attention network (NAN) is used to learn the relationships between channels, which used to identify the violence and speeds up the convergence process for training a perfect. Furthermore, the dataset took real-time security camera feeds from a variety of subjects and situations, as opposed to the hand-crafted datasets utilized in prior efforts. We also demonstrate the method's capability of assigning the correct category to each anomaly by classifying normal and abnormal occurrences. The method divided the information gathered into three primary groups: those in need of fire protection, those experiencing theft or violence, and everyone else. The study applied the proposed approach to the UCF-Crime dataset, where it outperformed other models on the same dataset. Our collection includes conditions with varying levels of light, movement, and human participants. As an example, the film experienced various irregularities, and in other clips, we didn't even see any people (i.e., car accidents). Additionally, we have another dataset restriction that has to be addressed. Abnormal occurrences may only last a few seconds, and even in 10-second recordings, normal behaviour is shown for more than 80% of the time. We used all 14 UCF-Crime categories, binary classification, and division into four main groups; we also edited the original footage of three aberrant instances. Both extraordinary and typical occurrences coexist with the same setting and props. We used MRCNN-AFR, a widely used network, to find the most important elements in each video frame. Next, a NAN structure is applied to the MRCNN-AFR output to investigate the out-of-the-ordinary occurrence over several images. At last, we employed classifiers for each dataset to discover how the classical correctly identifies the appropriate category for each input video. Our method outperformed the competition in experiments, but we'd still like to get better at categorising the whole range of irregularities present in the UCF-Crime dataset. In the future, one of the methods that we will employ is to add a layer of focus to the structure. Thus, CNN structure and/or NAN can benefit from the addition of this attention layer. This allows the model to zero in on the actual abnormalities present in the video data.

REFERENCES

[1] Xu, Z., Hu, C., Mei, L. (2016). Video structured description technology based intelligence analysis of surveillance videos for public security applications. Multimedia Tools and Applications, 75: 12155-12172. https://doi.org/10.1007/s11042-015-3112-5

[2] Castillo, A., Tabik, S., Pérez, F., Olmos, R., Herrera, F. (2019). Brightness guided preprocessing for automatic cold steel weapon detection in surveillance videos with deep learning. Neurocomputing, 330: 151-161. https://doi.org/10.1016/j.neucom.2018.10.076

[3] Hassner, T., Itcher, Y., Kliper-Gross, O. (2012). Violent flows: Real-time detection of violent crowd behavior. In 2012 IEEE computer society conference on computer vision and pattern recognition workshops, pp. 1-6. https://doi.org/10.1109/CVPRW.2012.6239348

[4] Xu, L., Gong, C., Yang, J., Wu, Q., Yao, L. (2014). Violent video detection based on MoSIFT feature and sparse coding. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3538-3542. https://doi.org/10.1109/ICASSP.2014.6854259

[5] Gao, Y., Liu, H., Sun, X., Wang, C., Liu, Y. (2016). Violence detection using oriented violent flows. Image and Vision Computing, 48: 37-41. https://doi.org/10.1016/j.imavis.2016.01.006

[6] Meng, Z., Yuan, J., Li, Z. (2017). Trajectory-pooled deep convolutional networks for violence detection in videos. In Computer Vision Systems: 11th International Conference, ICVS 2017, Shenzhen, China, July 10-13, 2017, Revised Selected Papers 11, pp. 437-447. https://doi.org/10.1007/978-3-319-68345-4_39

[7] Fenil, E., Manogaran, G., Vivekananda, G.N., Thanjaivadivel, T., Jeeva, S., Ahilan, A.J.C.N. (2019). Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM. Computer Networks, 151: 191-200. https://doi.org/10.1016/j.comnet.2019.01.028

[8] Ullah, F.U.M., Ullah, A., Muhammad, K., Haq, I.U., Baik, S.W. (2019). Violence detection using spatiotemporal features with 3D convolutional neural network. Sensors, 19(11): 2472. https://doi.org/10.3390/s19112472

[9] Zhang, L., Zhu, G., Shen, P., Song, J., Afaq Shah, S., Bennamoun, M. (2017). Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition. In Proceedings of the IEEE international conference on computer vision workshops, pp. 3120-3128.

[10] Engel, J.I., Martin, J., Barco, R. (2016). A low-complexity vision-based system for real-time traffic monitoring. IEEE Transactions on Intelligent Transportation Systems, 18(5): 1279-1288. https://doi.org/10.1109/TITS.2016.2603069

[11] Jahanshahi, M.R., Masri, S.F. (2012). Adaptive vision-based crack detection using 3D scene reconstruction for condition assessment of structures. Automation in

Construction, 22: 567-576. https://doi.org/10.1016/j.autcon.2011.11.018
[12] Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G., Sukthankar, R. (2011). 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, pp. 332-339. https://doi.org/10.1007/978-3-642-23678-5_39
[13] Ray, P.P., Mukherjee, M., Shu, L. (2017). Internet of things for disaster management: State-of-the-art and prospects. IEEE Access, 5: 18818-18835. https://doi.org/10.1109/ACCESS.2017.2752174
[14] Mabrouk, A.B., Zagrouba, E. (2018). Abnormal behavior recognition for intelligent video surveillance systems: A review. Expert Systems with Applications, 91: 480-491. https://doi.org/10.1016/j.eswa.2017.09.029
[15] Zhang, T., Jia, W., Yang, B., Yang, J., He, X., Zheng, Z. (2017). MoWLD: A robust motion image descriptor for violence detection. Multimedia Tools and Applications, 76: 1419-1438. https://doi.org/10.1007/s11042-015-3133-0
[16] Vijeikis, R., Raudonis, V., Dervinis, G. (2022). Efficient violence detection in surveillance. Sensors, 22(6): 2216. https://doi.org/10.3390/s22062216
[17] Sahay, K.B., Balachander, B., Jagadeesh, B., Kumar, G.A., Kumar, R., Parvathy, L.R. (2022). A real time crime scene intelligent video surveillance systems in violence detection framework using deep learning techniques. Computers and Electrical Engineering, 103, 108319.
[18] Mansour, R.F., Escorcia-Gutierrez, J., Gamarra, M., Villanueva, J.A., Leal, N. (2021). Intelligent video anomaly detection and classification using faster RCNN with deep reinforcement learning model. Image and Vision Computing, 112: 104229. https://doi.org/10.1016/j.compeleceng.2022.108319
[19] Alafif, T., Alzahrani, B., Cao, Y., Alotaibi, R., Barnawi, A., Chen, M. (2022). Generative adversarial network based abnormal behavior detection in massive crowd videos: A Hajj case study. Journal of Ambient Intelligence and Humanized Computing, 13(8): 4077-4088.
[20] Kim, D., Kim, H., Mok, Y., Paik, J. (2021). Real-time surveillance system for analyzing abnormal behavior of pedestrians. Applied Sciences, 11(13): 6153. https://doi.org/10.3390/app11136153
[21] Narejo, S., Pandey, B., Esenarro Vargas, D., Rodriguez, C., Anjum, M. R. (2021). Weapon detection using YOLO V3 for smart surveillance system. Mathematical Problems in Engineering, 2021: 9975700. https://doi.org/10.1155/2021/9975700
[22] Mohtavipour, S.M., Saeidi, M., Arabsorkhi, A. (2022). A multi-stream CNN for deep violence detection in video sequences using handcrafted features. The Visual Computer, 1-16. https://doi.org/10.1007/s00371-021-02266-4
[23] Mohtavipour, S. M., Saeidi, M., & Arabsorkhi, A. (2022). A multi-stream CNN for deep violence detection in video sequences using handcrafted features. The Visual Computer, 38: 2057–2072. https://doi.org/10.1007/s00371-021-02266-4
[24] Rehman, A., Butt, M.A., Zaman, M. (2022). Liver lesion segmentation using deep learning models. Acadlore Trans. Mach. Learn., 1(1): 61-67. https://doi.org/10.56578/ataiml010108
[25] Basysyar, F.M., Dwilestari, G. (2022). House price prediction using exploratory data analysis and machine learning with feature selection. Acadlore Trans. Mach. Learn., 1(1): 11-21. https://doi.org/10.56578/ataiml010103
[26] Rajasab, N., Rafi, M. (2022). A deep learning approach for biometric security in video surveillance system using gait. International Journal of Safety and Security Engineering, 12(4): 491-499. https://doi.org/10.18280/ijsse.120410
[27] Cheng, Y.K., Shi, Z.W., Zu, F.J. (2020). An evaluation model of subgrade stability based on artificial neural network. International Journal of Safety and Security Engineering, 10(5): 679-688. https://doi.org/10.18280/ijsse.100513