Diyala Journal of Engineering Sciences

Journal homepage: https://djes.info/index.php/djes

ISSN: 1999-8716 (Print); 2616-6909 (Online)

Crime Activity Detection in Surveillance Videos Based on Developed Deep Learning Approach

Rasool Jamal Kolaib, Jumana Waleed *

Department of Computer Science, College of Science, University of Diyala, Iraq

ARTICLE INFO

Article history:
Received June 11, 2024
Revised July 30, 2024
Accepted August 7, 2024
Available online September 1, 2024

Keywords:
Crime Activity Detection
Convolutional Neural Network (CNN)
Developed Deep Learning Approach
Pre-trained CNN approach
Surveillance Videos

ABSTRACT

In modern communities, lots of offenders are prone to recidivism, hence, there is a requirement to inhibit such criminals, especially from impending socioeconomically disadvantaged and high-crime areas that experience elevated levels of criminal activity, involving drug-related offenses, violence, theft, and other forms of anti-social behavior. Consequently, surveillance cameras have been installed in relevant institutions, and further personnel have been provided to monitor videos using various surveillance apparatus. However, relying solely on monitoring with the naked eye and manual video processing falls short of accurately evaluating the footage acquired via such cameras. To handle the issues of conventional systems, there is a need for a system that is able to classify acquired images while supporting surveillance personnel actively. Therefore, in this paper, a deep-learning approach is developed to build a crime detection system. This developed approach includes various layers necessary to perform feature extraction and classification processes and make the system capable of efficiently and accurately detecting crime activities from surveillance video frames. Besides the proposed crime activity detection system, two deep-learning approaches (EfficientNet-B7, and MobileNet-V2) are trained and assessed on the popular UCF Crime and DCSASS datasets. Generally, the proposed detection system encompasses dataset preparation and pre-processing, splitting the pre-processed crime activity image dataset, and implementing the proposed deep learning approach and other pre-trained approaches. The experiments uncover that the proposed system achieved outstanding results and outscored the other deep-learning approaches and relevant state-of-the-art with an accuracy of 99.48%, precision of 99.47%, sensitivity of 99.41%, and F1 score of 99.44% using the UCF crime dataset, and accuracy of 89%, precision of 89%, sensitivity of 88%, and F1 score of 88% using DCSASS dataset. Hence, this system is effectively capable of tracking criminals' trails and detecting crime events.

1. Introduction

Severe crime activities such as theft, kidnapping, molestation, killing, vandalism, and intensive attacks represent a serious hazard to protecting the public and developing society. The count of criminal activities declared within a single day is growing considerably. Of these activities, only significant or crucial events are noted. However, attention should be paid to all activities, as even insignificant activities can lead to devastating consequences with time passing. Subsequently, governments worldwide are constantly searching for diverse approaches to detect and prevent crimes [1]. Surveillance cameras are increasingly installed in smart cities; including schools, shopping centers, hospitals, and other institutions, as a dissuasion or detection of criminals after committing a crime. This expanding count of cameras

* Corresponding author.
E-mail address: jumanawaleed@uodiyala.edu.iq
DOI: 10.24237/djes.2024.17307

This work is licensed under a Creative Commons Attribution 4.0 International License.

necessitates adequate personnel resources to manage the volume of generated videos, however, even if this requirement is achieved, the personnel responsible for managing and filtering videos will wind up with mental health issues owing to higher exposure to aggressive and criminal content [2]. Consequently, computer vision methods with the aid of artificial intelligent systems for video crime activity auto-detecting are becoming increasingly important [3].

Recently, video crime detection has acquired a lot of attention after realizing auspicious results by exploiting the incredible robustness of deep learning algorithms, especially Convolutional Neural networks (CNNs) [4]. The fundamental merit of CNNs is the hidden complex features extraction from high-dimensional data, which makes them appropriate feature extractors for image and video classification and detection tasks [5-7]. CNNs provide powerful tools for processing crime data complexities and detecting (or classifying) anomalies. By efficiently utilizing CNN approaches, crime activity detection systems can be more adaptive, accurate, and able to handle the progressing nature of criminal activities [8].

Concerning the factor of supervision, in general, crime activity detection tasks can be classified into several fundamental categories; weakly supervised, supervised, and unsupervised. The deep learning approach is trained in supervised crime activity detection using datasets containing labeled instances (of criminal and normal behaviors). Depending on the labeled data, this approach learns to discriminate between criminal and normal instances. In weakly supervised crime detection, the deep learning approach is implemented on labeled and non-labeled data together. The deep learning approach is trained on a dataset that principally involves labeled data (normal instances) and does not involve clear labels for crime instances. Depending on the labeled data, this approach learns the normal patterns, and depending on the highly deviant instances from these patterns, it identifies the criminal patterns. The last category is based solely on unlabeled data to flag highly deviant patterns without knowing what normal behavior constitutes. These categories hold various merits and demerits, and their selection is based on many aspects; data nature, labeled data availability, and other requirements for the crime detection task [9].

In numerous computer vision applications, CNN approaches were getting deeper, which made applying them in over-edge devices questionable. To conquer this issue and reach lightweight functionality, in this work, a supervision, lightweight CNN approach is proposed to classify crimes efficiently. The leading contribution of the proposed crime activity detection system is stated as follows:

1. Present end-to-end architecture to detect crime activity in surveillance videos.
2. Propose a deep learning approach to enhance the reliability and efficiency of video surveillance solutions. This approach enhances crime prevention abilities by providing automated anomaly detection.
3. Implement suitable pre-trained CNN approaches like EfficientNet-B7, and MobileNet-V2. These approaches align with the paper's aim, leading to selecting the best performance.
4. Perform extensive experiments using one of the benchmark crime-related datasets to detect various crime activity classes.

The remnant of this work is arranged as follows. Closely relevant works are presented in the next Section. The utilized crime-related dataset and the proposed crime activity detection system are clarified in Section 3. Experiments, evaluations, and discussions are presented in Section 4. The essential conclusions and future extents are pointed out in the last Section.

2. Related works

The emergence of crime detection systems is paramount, as its fundamental aim is to effectively anticipate and deter criminal activities. Recently, much research has increasingly exploited deep learning for crime activity detection, resulting in distinct

approaches and findings, and considerably increasing the knowledge reservoir. However, announced outcomes reveal prominent variability owing to variations in utilized approaches, settings, strategies, and complexities.

Ullah et al. [10], presented a lightweight CNN-based crime detection system that exploited a MobileNet-V2 approach to extract essential spatial features from surveillance video frames, and a residual attention-based Long Short Term Memory (LSTM) to classify anomalous activity sequences into four classes (explosion, assault, fighting, and normal). Experiments were accomplished on the real-world UCF crime dataset, and the attained results illustrated that the proposed system presented acceptable results. However, other deep learning approaches should be investigated to improve the system's performance. Furthermore, a generative method should be developed to detect more classes of crime activity.

Gulati et al. [11], presented a criminal actions detection system using improved CNN which encompasses various layers; two convolutional layers, Batch normalization, max pooling, global average pooling, one dense, Batch normalization, and the final dense layer with sigmoid function for binary classification. In this system, a rolling averaging scheme was used, where "n" window size is adopted to average the probabilities of predictions by the approach over "n" frames, and the highest probability prediction is adopted as the prediction for these "n" frames. The experiments were accomplished using several datasets such as Hockey Fight, Violent Flows, and DCSASS (which is a subset of the UCF crime dataset), and the attained results of accuracies were 97.94 %, 95.75 %, and 87.56 %, respectively.

Thakare et al. [12], presented a multi-stream deep learning-based anomaly detection system in surveillance videos. This system was concentrated on attaining strong visual features via fusing spatial and motion information. Advanced classifiers based on multiple-instance learning have been proposed to handle feature variations. Moreover, a method of fuzzy aggregation was utilized for fusing diverse separate feature stream scores to enhance the accuracy of segment detection in videos with long duration. Eventually, a lightweight classifier of two classes was utilized for only classifying accident and fire activities. However, this classifier cannot be satisfactory for classifying human-related and robbery-related classes. Broad experiments were performed on this system using the UCF Crime dataset, and the classification accuracy was as high as 84.48%.

Khaire and Kumar [13], proposed a crime activity detection system based on the pre-trained CNN approach called Mobile-Net and Bidirectional Long Short-Term Memory to be utilized in critical environments (such as ATMs). In this system, the raw RGB and depth frames were exploited to extract rich visual features and to decrease the computational complexity of training and classification. The main merit of this system is that the training process is accomplished only on weakly labeled normal samples. The proposed system was assessed using the training and testing splits of the UCF crime2local dataset of six classes. This utilized dataset represents a subset of the UCF Crime dataset involving 300 RGB video samples. The accuracy of the proposed system was impressive at 91.1%.

Qasim and Verdu [14], suggested a deep learning system for identifying normal or anomaly behaviors specified in the Binary and UCF Crime datasets that involved violent and abnormal behaviors captured using cameras in various public positions. In this system, ResNet-18, ResNet-34, and ResNet-50 have been utilized as CNN approaches for extracting high-level spatial features from video frames, and the simple recurrent unit (SRU) has been utilized to collect temporal features. The combined ResNet-50 and SRU approaches outperformed the other combined models on the Binary and UCF-Crime datasets with 91.63% and 91.25% accuracy, respectively. However, the classification of anomalies especially in the UCF Crime dataset should be improved.

Park et al. [15], presented an initial deep learning-based anomaly detection framework in surveillance videos that utilizes an absorbing Markov Chain (MC) to filter the noisy

predictions prompted via weak supervision. This MC was integrated into the training process of a deep learning network via a graph convolutional network adoption. Gaussian mixture-based pseudo-labeling model was utilized to enhance the segment-wise labels. This presented framework provided good performance with accuracies of 95.61% and 84.94% for ShanghaiTech and UCF Crime datasets, respectively.

Patwal et al. [16], attempted to detect anomalies like property destruction, violence, burglary, assault, theft, blast, etc., in addition to detecting crowd irregularities. The presented system utilized one of the effective pre-trained CNNs named DenseNet-121 as an extractor of features and a classifier. This system achieved an accuracy of 86.63% using the UCF Crime dataset. However, this system should be improved and increase the classification accuracy using other CNN approaches.

3. Proposed detection system

The task of the crime activity detection system is to classify the type of crime into various classes reliably. Many images of criminal activities (including normal videos, vandalism, stealing, shoplifting, shooting, robbery, road accidents, fighting, explosion, burglary, assault, arson, arrest', and abuse) exist in the utilized database to train and test the approach.

The proposed crime activity detection system (as demonstrated in Figure 1) encompasses several main phases; dataset preparation and pre-processing, splitting pre-processed crime activity images into test and train corpora, implementing the proposed deep learning approach, and evaluating the system performance utilizing various assessment measures.

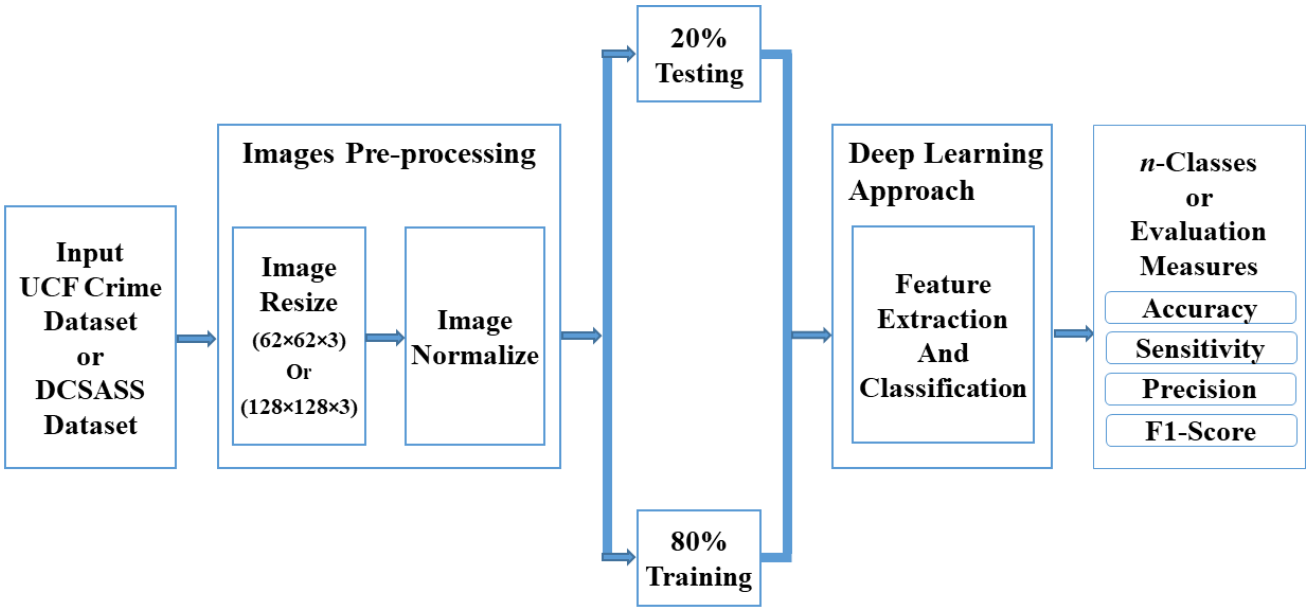


Figure 1. Architecture of crime activity detection system

3.1 Dataset preparation and pre-processing

In this phase, the data is prepared and pre-processed to adapt the utilized datasets to the input for the proposed CNN approach before training. Surveillance videos were first transformed into consecutive frames and then resized into (62×62) pixels in the UCF Crime dataset and resized into (128×128) pixels in the DCSASS dataset (with a frame rate of 30 frames per second). This pre-process ensures the resizing to a uniform size and can enhance the generalization and performance of the approaches. After that, resized video frames were normalized within [0, 1] using min-max normalization method to ensure that frames held the same scales. In the min-max normalization, every pixel intensity in the resized image (it's

value ranging from 0 to 255) is processed individually, in which the normalized value can be attained via subtracting the pixel value from the minimum value in the non-criminal or criminal image divided by the subtraction of the maximum value from the minimum value, resulting to the scaled image. This normalization technique accomplished intensity range scaling of the images, and attained the intensity distribution of images adequately smoother, with a bigger range. Additionally, this pre-process guarantees that the deep learning approaches become less sensitive to pixel intensity variations, contributing to enhanced convergence throughout training. After this phase, the frame's information could be fed into the specified deep-learning approach.

The dataset of images is then separated into testing and training data with a ratio of 20:80. The training data is employed to train the proposed CNN approach in the training process, and the testing data is employed to perform the testing process.

3.2 Proposed CNN approach implementation

The proposed deep CNN approach encompasses two processes; extracting features and classification, and the proposed approach architecture is depicted in Figure 2. The description of the layers, their output shapes, and parameters for the proposed approach are depicted in Table 1 and Table 2.

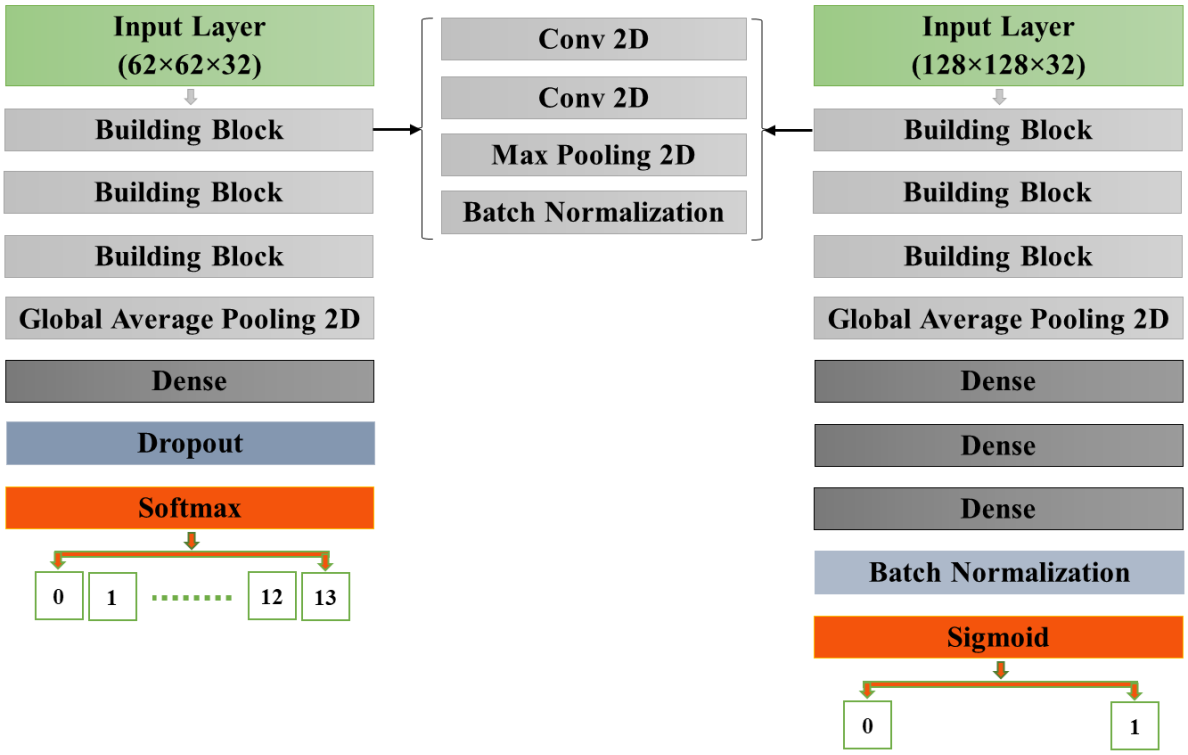


Figure 2. The architecture of the proposed CNN approach; (on the left side) using the UCF Crime dataset, and (on the right side) using the DCSASS dataset

In the feature extraction process, the most relevant image features are determined to attain a sequence of feature vectors. This process is accomplished using three building blocks. Each block encompasses several layers;

- Two layers of convolution: These layers perform the fundamental operations in the proposed CNN for extracting essential features via implementing a fixed count of filters to the pre-processed image to produce output 2D-feature maps. Each of these layers passed to the Rectified Linear Unit (ReLU) activation function for speeding up the training operation and converging more rapidly.
- One MaxPooling: This layer aims to down-sampling the resulting feature maps, i.e. it reduces the feature maps by

picking the biggest value within the (2×2) regions in the image. Max pooling hinders the issue of overfitting via the abstraction of the image's visual representation.

- One Batch normalization: It considerably enhanced the effectiveness of CNN training. Batch normalization helps avoid fading and bursting gradients by making activations more stable over training, and training thus becomes less of a condition for preparation. Moreover, it represents a regularizer, it accordingly repeals the need for utilizing dropouts. The utilization of dropout represents a significant solution for handling overfitting in CNN, permitting it to generalize unseen data and improve performance.

While, in the classification process, the numerical image features are analyzed (or classified) into 14 classes for crime activity using the UCF Crime dataset and 2 classes for the DCSASS dataset.

The classification process is accomplished for the UCF crime dataset using the following:

- One Global Average Pooling: This type of pooling provides various merits that strengthen the approach's performance. It chooses the most eminent from the input feature map to obtain the output, enabling the approach to catch and learn the principal features inside the image, hence enhancing the approach's accuracy.
- Dense layer involving fully connected neurons: Fully connected converts the received feature maps acquired from Global Average Pooling into a 1D-vector that affects the whole feature information.
- Dropout: It works on randomly chosen neurons to be neglected during training.
- Softmax layer for classification: SoftMax function guarantees that probability summation throughout entire classes is equivalent to one, ensuring a normalized output appropriate for the task of image classification. This attribute of normalization could improve classification reliability and stability, specifically during the handling of imbalanced datasets.

Concerning the DCSASS dataset, this process is accomplished using one Global Average Pooling, three dense layers followed by batch normalization, and the last dense layer connected to a sigmoid classifier for binary classification.

Table 1: The layers' names and parameters of the proposed CNN approach using the UCF crime dataset

	Layer	Output Shape	Parameters
Building Block	Conv2D	(None, 62, 62, 32)	896
	Conv2D	(None, 60, 60, 64)	18496
	MaxPooling2D	(None, 30, 30, 64)	0
	Batch Normalization	(None, 30, 30, 64)	256
Building Block	Conv2D	(None, 28, 28, 64)	36928
	Conv2D	(None, 26, 26, 64)	36928
	MaxPooling2D	(None, 13, 13, 64)	0
	Batch Normalization	(None, 13, 13, 64)	256
Building Block	Conv2D	(None, 11, 11, 128)	73856
	Conv2D	(None, 9, 9, 128)	147584
	MaxPooling2D	(None, 4, 4, 128)	0
	Batch Normalization	(None, 4, 4, 128)	512
	Global Average Pooling2D	(None, 128)	0
	Dense	(None, 128)	16512
	Dropout	(None, 128)	0
	Classification (Dense)	(None, 14)	1806
	Total parameters: 334,030 (1.27 MB)		
	Trainable parameters: 333,518 (1.27 MB)		
	Non-trainable parameters: 512 (2.00 KB)		

Table 2: The layers' names and parameters of the proposed CNN approach using the DCSASS dataset

	Layer	Output Shape	Parameters
Building Block	Conv2D	(None, 128, 128, 32)	896
	Conv2D	(None, 124, 124, 32)	9,248
	MaxPooling2D	(None, 62, 62, 32)	0
	Batch Normalization	(None, 62, 62, 32)	128
Building Block	Conv2D	(None, 60, 60, 64)	18,496
	Conv2D	(None, 58, 58, 64)	36,928
	MaxPooling2D	(None, 29, 29, 64)	0
	Batch Normalization	(None, 29, 29, 64)	256
Building Block	Conv2D	(None, 27, 27, 128)	73,856
	Conv2D	(None, 25, 25, 128)	147,584
	MaxPooling2D	(None, 12, 12, 128)	0
	Batch Normalization	(None, 12, 12, 128)	512
	Global Average Pooling2D	(None, 128)	0
	Dense	(None, 512)	66,048
	Dense	(None, 256)	131,328
	Dense	(None, 128)	32,896
	Batch Normalization	(None, 128)	512
	Classification (Dense)	(None, 1)	129
	Total parameters: 518,817 (1.98 MB)		
	Trainable parameters: 518,113 (1.98 MB)		
	Non-trainable parameters: 704 (2.75 KB)		

3.3 Pre-trained CNN approach implementation

Recently, diverse pre-trained CNNs have proven their effectiveness in diverse image classification tasks, involving crime activity detection in surveillance videos. To efficiently extract features and classify images for criminal activity detection, the proposed system implemented two of the most efficient pre-trained CNN approaches (EfficientNet-B7, and MobileNet-V2).

MobileNet-V2 [17] MobileNet-V2 represents an efficient, speedy, and less memory usage approach. This deep learning approach encompasses the utilization of depthwise separable convolutions presented in the first version, which works on splitting the normal convolutions into depthwise and pointwise convolutions to reduce size and approach computation considerably. It utilizes a sequence of the inverted residuals and linear bottlenecks to provide an outstanding balance between efficiency and accuracy, in which the channels are expanded (with a convolution of 1x1), and depthwise convolution of 3x3 is then applied, after that, the channels are projected back to less dimension (with another convolution of 1x1). The last convolution of 1x1 in the inverted residual block utilizes a linear activation (linear bottlenecks) to preserve more information.

EfficientNet-B7 [18] represents an accurate approach while maintaining efficiency regarding computational resources. It is the greatest and most powerful release in the series of EfficientNets, providing advanced performance on various image classification benchmarks. In this approach, a novel method of compound scaling was introduced in which the count of layers (depth), count of channels (width), and input image size (resolution) were uniformly scaled by utilizing a collection of determined scaling coefficients, thus the whole network dimensions are scaling instead of one dimension at a time, resulting in a superior performance with fewer resources. Analogous to MobileNet, EfficientNet-B7 also utilizes depthwise separable convolutions to minimize the cost of computation and enhance efficiency without compromising accuracy. Furthermore, mobile inverted bottleneck convolution blocks are exploited in this approach to improve efficiency by utilizing fewer operations and parameters while preserving performance, and these blocks are equivalent to the inverted residual blocks in MobileNet-V2. Generally, EfficientNet-B7 encompasses diverse layers

initiated with convolution (64-filter, 2-stride, and 3x3 kernel size), batch normalization, and activation of Swish. Several blocks of Mobile Inverted Bottleneck Convolutional (MBConv) follow this initiation and after these blocks, convolution of 1x1 (2048-filter), batch normalization, and activation of Swish are implemented. Finally, a global average pooling and fully connected layers are utilized for the purpose of classification.

4. Experimental results and discussion

As the system encompasses implementing several CNN approaches, the attained results and performance of these approaches have been underscored. The description of the utilized dataset, assessment measures, and approaches' validation are highlighted in the subsequent sub-sections to assess the proposed detection system.

4.1 Dataset description

Typically, the datasets utilized for developing the implemented deep learning approaches involve training, calibration (or validation), and testing sets. The calibration set works on monitoring the system performance through training and comparing the implemented CNN architectures. When the last approach is specified and its whole hyper-parameters are determined, the system performance is assessed using the testing set.

To investigate the system's performance various experiments have been conducted on the UCF Crime and DCSASS datasets. The UCF Crime dataset [19] involves 1900 surveillance videos (of the real world) associated with crime scenes and anomalies in public locations. This dataset holds one normal class and thirty classes of crime activity (a total of forty classes). In every full-length video, each tenth frame is extracted and merged for each video in that class. The count of extracted PNG images (frames) is 1,377,653 of size 64×64.

In order to make the dataset more balanced, 900,000 images are removed from normal videos. The image distribution of the balanced UCF Crime dataset is demonstrated in Figure 3.

The second utilized dataset is the DCSASS Dataset [20] which involves surveillance videos encompassing criminal and non-criminal behaviors and labeled as abnormal (1) or normal (0). The construction of this dataset was based on the UCF Crime dataset which includes thirteen crime classes. The total count of crime videos was 16853, and the video distribution of the DCSASS dataset is demonstrated in Figure 4. Samples of four frames for each class for the UCF crime and DCSASS datasets are demonstrated in Figure 5.

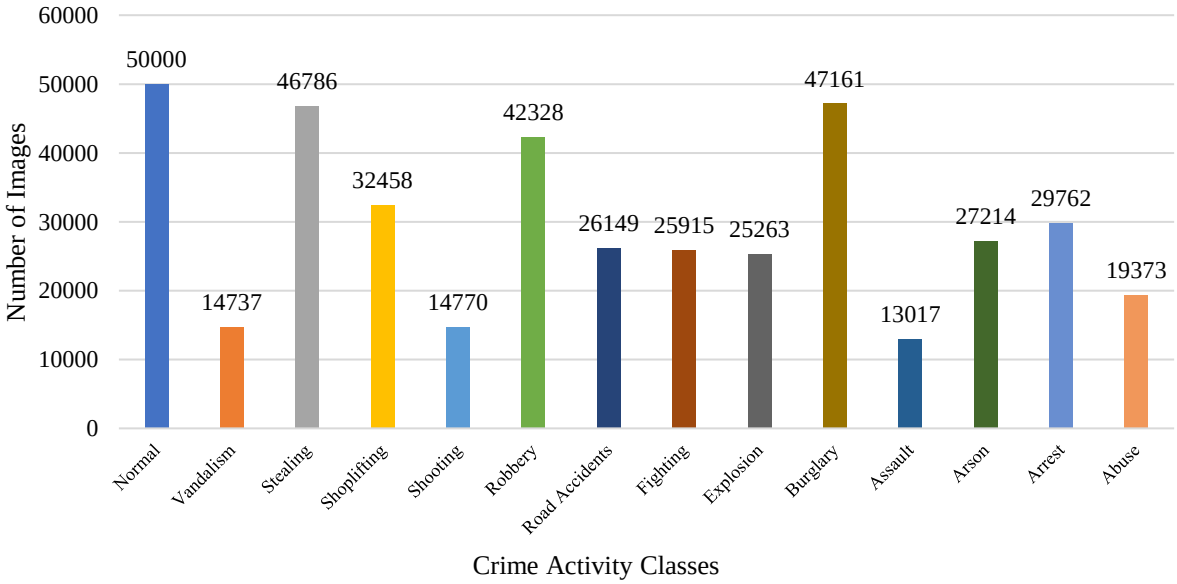


Figure 3. Distribution of the balanced UCF Crime Dataset

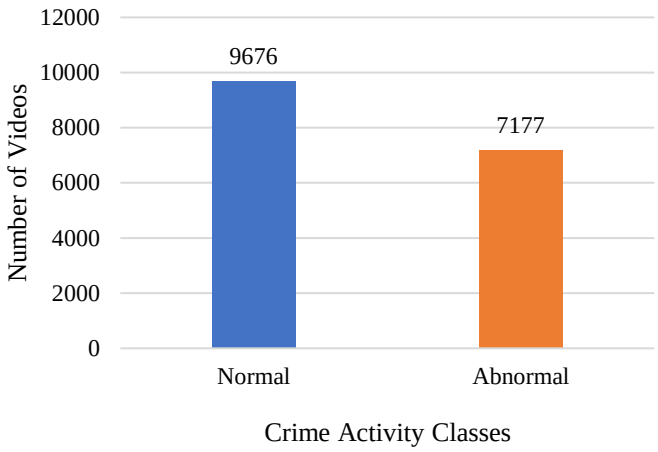


Figure 4. Distribution of the DCSASS Dataset

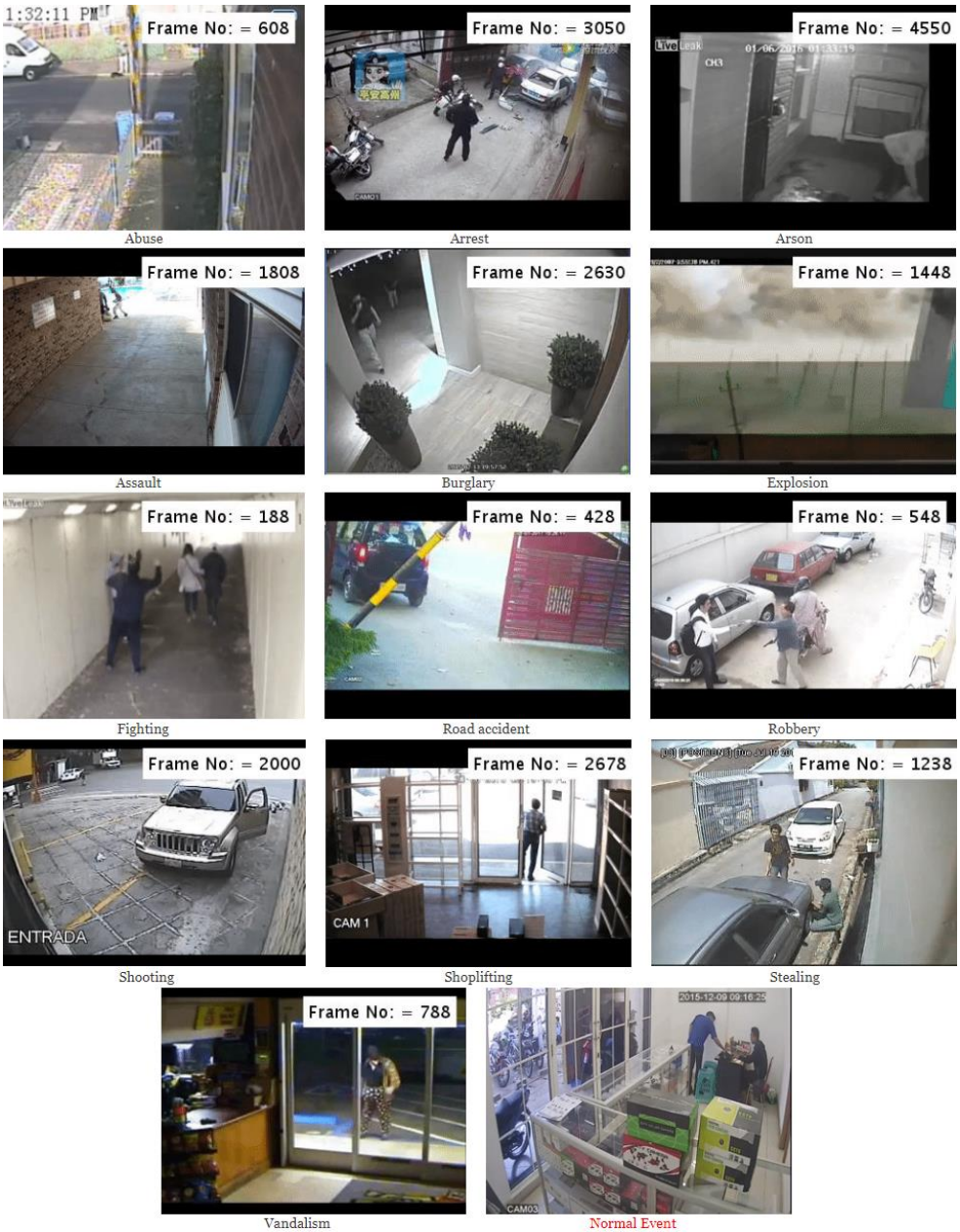


Figure 5. Samples of various crime activities from surveillance videos in the UCF Crime and DCSASS datasets

4.2 Assessment measures

The efficient manner to assess the performance of the proposed classification CNN approach is by checking its confusion matrix [21] [22]. This matrix depicts the variance between the ground truth of the utilized dataset and the approach's predictions, and it depends on the following aspects; True Positive (T_rP) depicts the count of negative crime activity images identified rightly as negative crime images, True Negative (T_rN) depicts the count of positive crime activity images identified rightly as positive crime activity images, False Positive (F_aP) depicts the count of positive crime images identified rightly as negative, and False Negative (F_aN) depicts the count of negative crime activity images wrongly identified as positive crime activity images. Figure 6 and Figure 7 depict the confusion matrices for the proposed CNN and other pre-trained approaches using the balanced UCF Crime and the DCSASS datasets, respectively.

Generally, video crime detection systems are assessed by calculating the accuracy of crime activity classification for individual surveillance videos. Nevertheless, this measure is not capable of completely explaining the systems' performance. Other measures are also implemented that are more concise and can supply more precise information concerning the assessment of the classifier's performance.

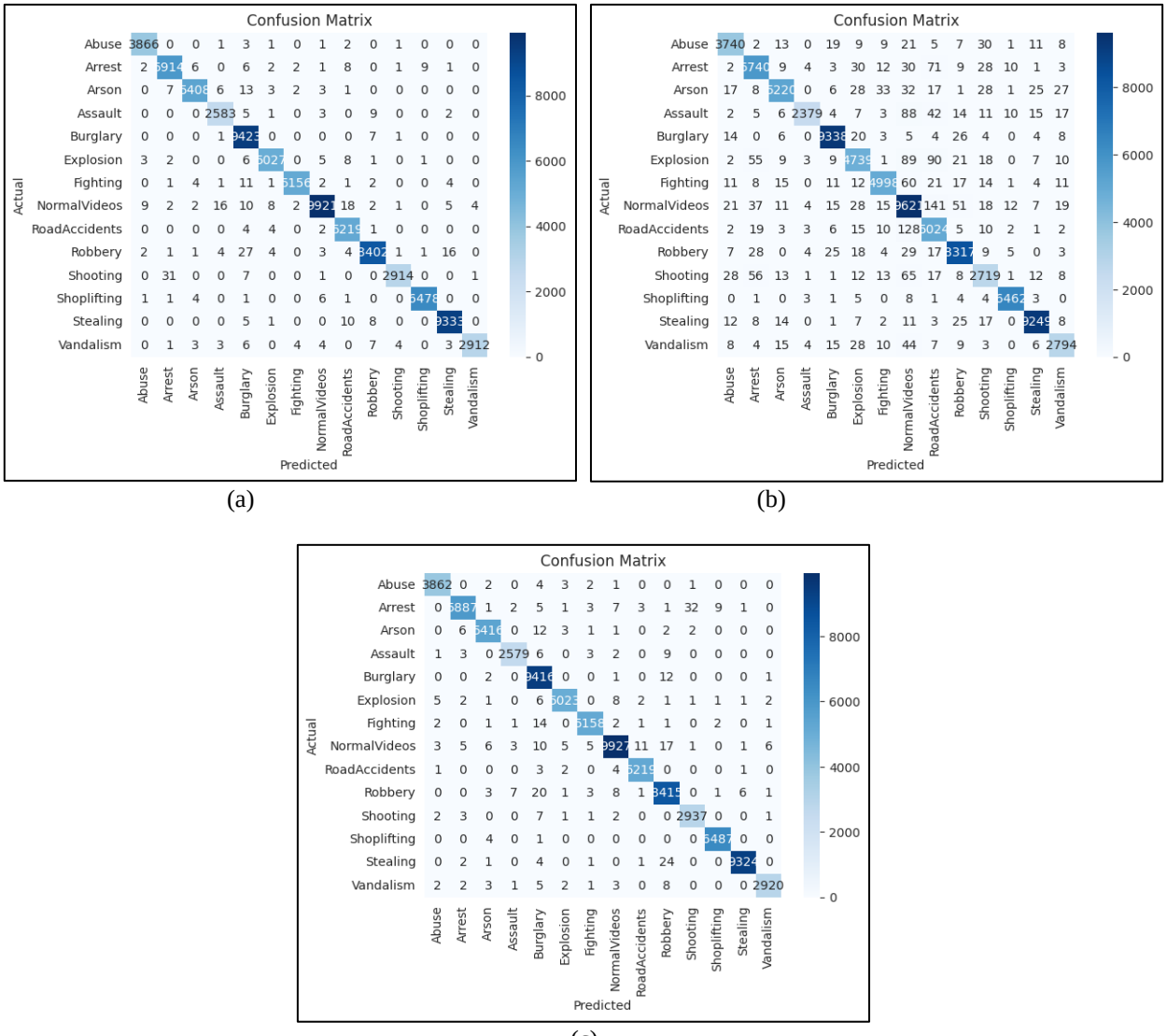


Figure 6. Confusion matrix using UCF Crime dataset for (a) The proposed detection system, (b) EfficientNet-B7, and (c) MobileNet-V2

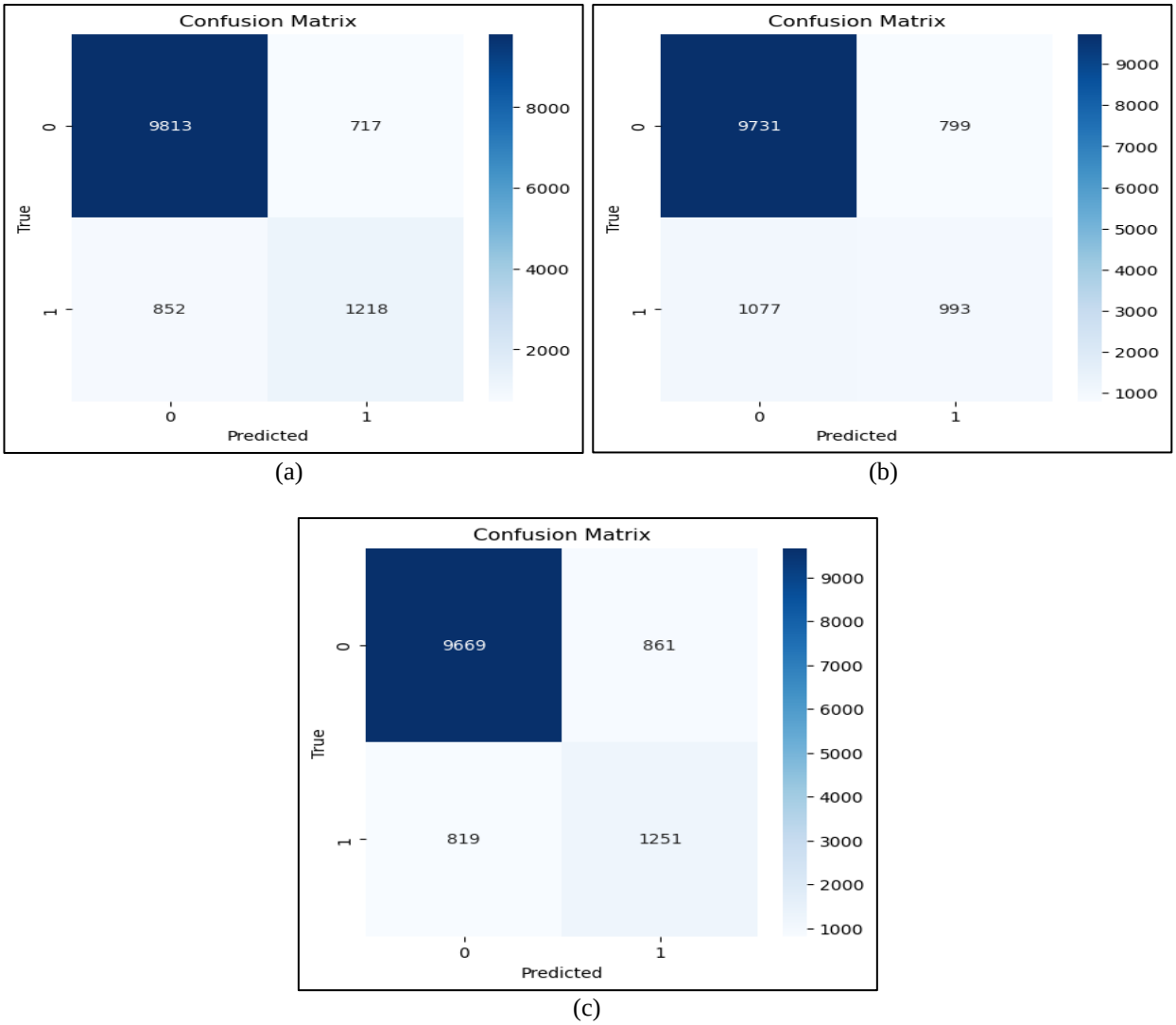


Figure 7. Confusion matrix using DCSASS dataset for (a) The proposed detection system, (b) EfficientNet-B7, and (c) MobileNet-V2

The classification accuracy (A_c) of the proposed approach can be computed using the following formula:

$$A_c = \frac{T_rN+T_rP}{F_aN+F_aP+T_rN+T_rP} \times 100\% \quad (1)$$

Sensitivity (S_e) also known as classifier recall depicts a measure for the count of crime images classified correctly as such. It can be computed using the following formula:

$$S_e = \frac{T_rP}{F_aN+T_rP} \times 100\% \quad (2)$$

The precision (P_r) measure depicts the right predictions of positive observations, while the F_1 measure depicts total observation that takes sensitivity and precision into consideration. These measures can be computed using the following formulas:

$$P_r = \frac{T_rP}{F_aP+T_rP} \times 100\% \quad (3)$$

$$F_1 = 2 \times \frac{P_r \times S_e}{P_r+S_e} \times 100\% \quad (4)$$

The attained results for the utilized measures are exhibited in Table 3 and Table 4, and all these measures provided outstanding findings. This pointed out that the proposed CNN approach was carried out as expected, with low rates of false positives and negatives.

Table 3: Results of the proposed system using the UCF crime dataset

Classes	P_r	S_e	F_1	Support
Abuse	0.995622	0.997677	0.996649	3875
Arrest	0.992282	0.993616	0.992948	5952
Arson	0.996315	0.99357	0.994941	5443
Assault	0.987763	0.992317	0.990034	2603
Burglary	0.989085	0.999152	0.994093	9432
Explosion	0.995051	0.994855	0.994953	5053
Fighting	0.998064	0.994791	0.996425	5183
Normal Videos	0.996885	0.992	0.994436	10000
Road Accidents	0.989947	0.997897	0.993906	5230
Robbery	0.995616	0.99244	0.994025	8466
Shooting	0.996921	0.986459	0.991662	2954
Shoplifting	0.998305	0.997843	0.998074	6492
Stealing	0.996689	0.997435	0.997062	9357
Vandalism	0.998286	0.988124	0.993179	2947
A_c		**0.994806**		82987
macro avg	0.994774	0.994155	0.994456	82987
weighted avg	0.994821	0.994806	0.994806	82987

Table 4: Results of the proposed system using the DCSASS dataset

Classes	P_r	S_e	F_1	Support
0	0.92	0.93	0.93	10530
1	0.63	0.65	0.64	2070
A_c		**0.89**		12600
macro avg	0.77	0.79	0.78	12600
weighted avg	0.89	0.88	0.88	12600

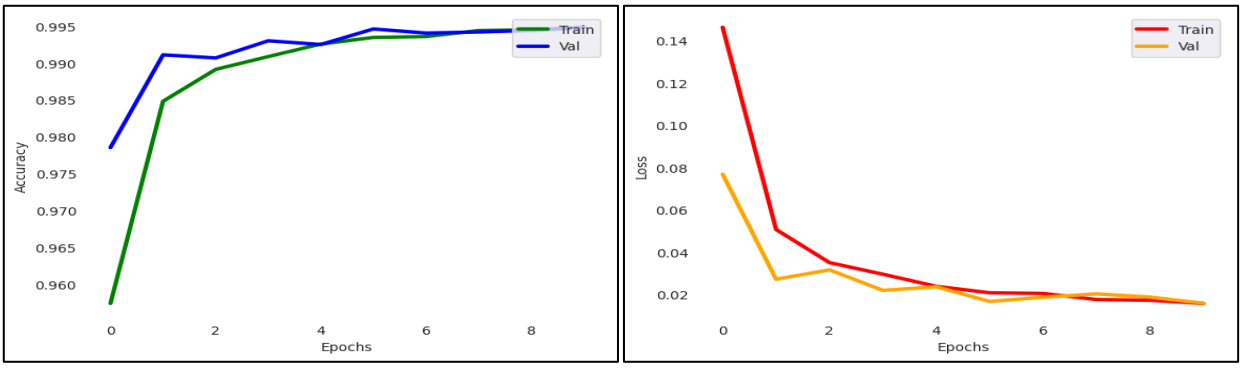
The proposed crime activity detection system was run at 10-15 epochs (using UCF crime dataset) and 20-50 epochs (using DCSASS dataset), 32-batch size, and 0.00003 learning rate. The system performance could be specified with the count of epochs needed to obtain the highest achieved accuracy. The final validation/training accuracies alteration using the proposed CNN approach were observed in epoch number 9, and their values were 0.9948 and 0.9938 for the UCF crime dataset, and the final validation/training accuracies alteration were observed in epoch number 18, and their values were 0.9321 and 0.8943 for the DCSASS dataset.

It is potential to utilize the comparison between two curves to specify the approaches' performance. The approach is over-fitted when the validation loss is decreased and increased once again, and it is under-fitted when the validation loss becomes extremely high. While the ideally lined up curve depicts the CNN approach performance is ideal. The final validation/training losses alteration were observed in epoch number 9, and their values were 0.0159 and 0.0156 for the UCF crime dataset, and the final validation/training losses alteration were observed in epoch number 9, and their values were 0.3004 and 0.3833 for the DCSASS dataset, Figure 8 and Figure 9 exhibit these alterations until the best epochs were attained.

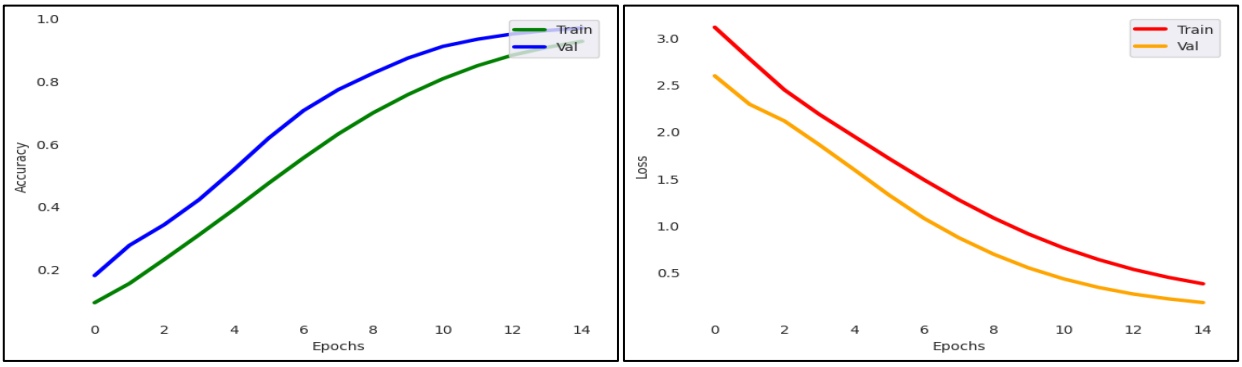
Concerning the proposed CNN approach, it can be noticed that the training approach performed slightly better than the validation approach. In the validation approach, the accuracy increased and loss decreased since it

referred to how the actual results were turned from the predicted results.

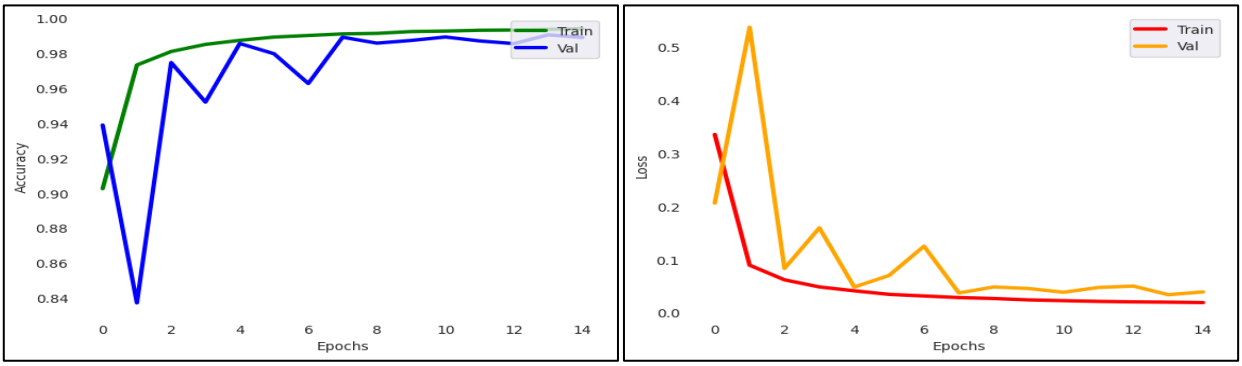
It is noticeable that the MobileNet-V2 achieved higher accuracy than EfficientNet-B7 because the UCF Crime and DCSASS datasets contain simpler patterns efficiently learned by a smaller model. Additionally, the datasets used do not contain enough variance, thus EfficientNet-B7 could not generalize well enough, resulting in lower accuracy. Furthermore, MobileNet-V2 had better pre-trained weights on the datasets used, therefore, it provided higher accuracy.



(a)



(b)



(c)

Figure 8. Validation/training accuracy & loss using UCF crime dataset for (a) The proposed detection system, (b) EfficientNet-B7, and (c) MobileNet-V2

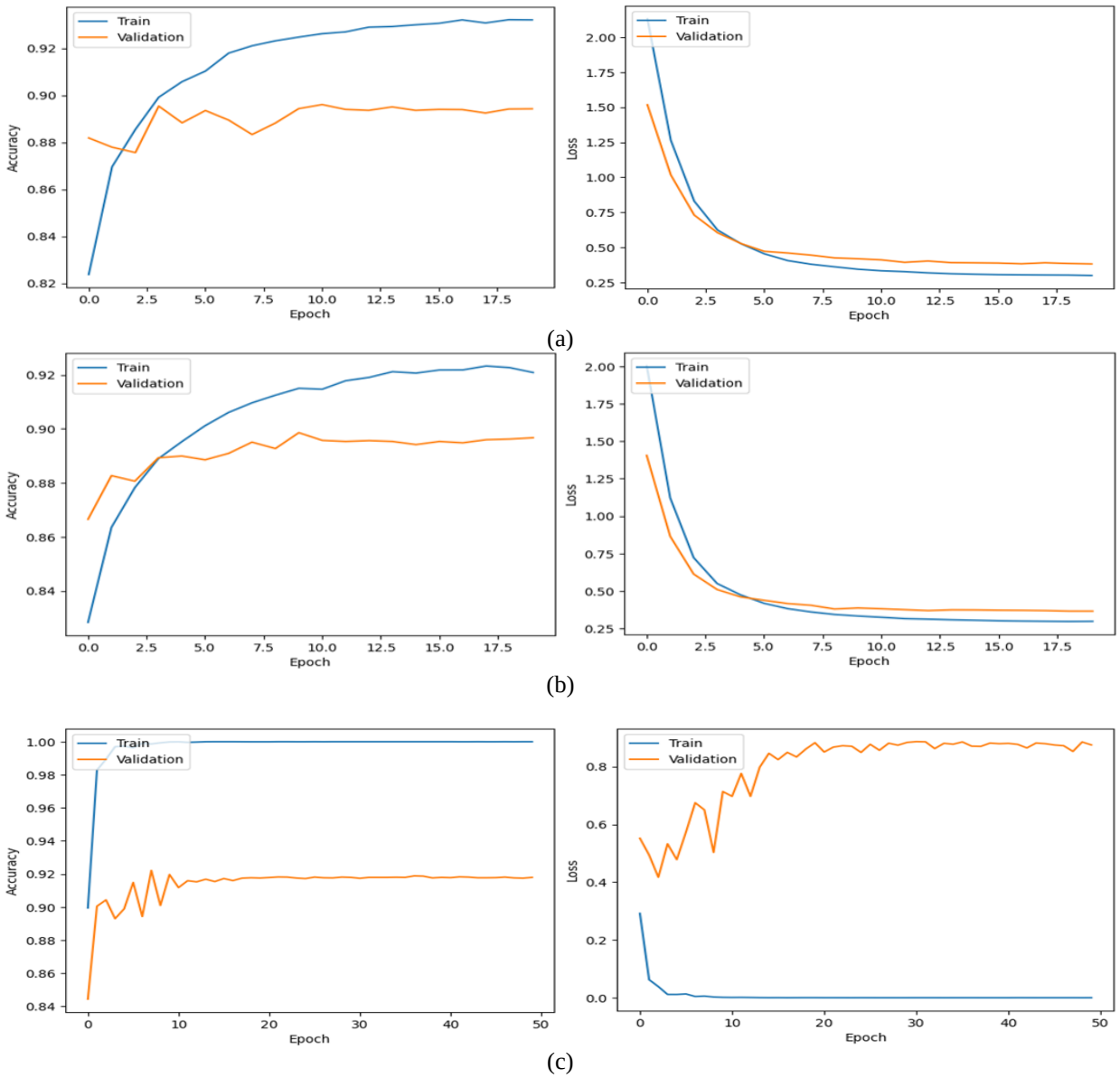


Figure 9. Validation/training accuracy & loss using DCSASS dataset for (a) The proposed detection system, (b) EfficientNet-B7, and (c) MobileNet-V2

The developed deep learning approach was compared with the MobileNet-V2 and EfficientNet-B7 approaches using the same count and size of frames, and the experiments were accomplished using the same conditions, the count of epochs and batch size. Additionally, we compared the proposed CNN approach with the closely relevant works. These comparisons are stated in detail in Table 5.

The MobileNet-V2 approach provided a balanced solution for the crime detection system where computational efficiency, quick image processing, and limited storage capacity are given priority. Nevertheless, its weaknesses in addressing the need for deep feature extraction and realizing the highest accuracy in classifying varied criminal activities should be taken into account. While the EfficientNet-B7 requires high computation and has a slow inference speed making it less preferable for real-time crime activity classification on edge devices. Compared with EfficientNet-B7, the MobileNet-V2 approach provided higher performance.

Table 5: Comparison with closely relevant crime activity classification systems using diverse assessment measures

Author/(s), Year, Ref.	Utilized deep learning Models	Utilized Dataset	Attained Results			
			P_r	S_e	F_1	Accuracy
Ullah et al., 2021, [10]	MobileNet-V2 and a residual attention-based LSTM	UCF Crime dataset	87%	78%	81%	78.43%
Gulati et al., 2021, [11]	Developed CNN	DCSASS dataset	-	-	-	87.56%
Thakare et al., 2022, [12]	Multi-stream deep learning with fuzzy aggregation	UCF Crime dataset	89%	86%	81%	84.48%
Qasim and Verdu, 2023, [14]	ResNet-50 and SRU models	UCF Crime dataset	91.54%	-	91.94%	91.25%
Park et al., 2023, [15]	Deep neural network with absorbing MC	UCF Crime dataset	-	-	-	84.94%
Patwal et al., 2023, [16]	DenseNet-121	UCF Crime dataset	-	-	-	86.63%
Implemented Approaches	MobileNet-V2	UCF Crime dataset	99%	98%	98%	98%
	EfficientNet-B7		97%	96%	96%	97%
	Proposed Approach		99.47%	99.41%	99.44%	99.48%
	MobileNet-V2	DCSASS dataset	87%	87%	87%	87%
	EfficientNet-B7		84%	85%	85%	85%
	Proposed Approach		89%	88%	88%	89%

Concerning the UCF crime dataset, the accuracy achieved by the EfficientNet-B7 approach was 97%, precision was 97%, sensitivity was 96%, and F1 score was 96%. While the accuracy achieved by the MobileNet-V2 approach was 98%, precision was 99%, sensitivity was 98%, and F1 score was 98%. Furthermore, concerning the DCSASS dataset, the accuracy achieved by the EfficientNet-B7 approach was 85%, precision was 84%, sensitivity was 85%, and F1 score was 85%. While the accuracy achieved by the MobileNet-V2 approach was 87%, precision was 87%, sensitivity was 87%, and F1 score was 87%.

The proposed system achieved the highest results and outscored the other implemented deep-learning approaches and relevant state-of-the-art. The accuracy achieved by the proposed CNN approach using the UCF crime dataset was 99.48%, precision was 99.47%, sensitivity was 99.41%, and F1 score was 99.44%. The accuracy achieved by the proposed CNN approach using the DCSASS dataset was 89%, precision was 89%, sensitivity was 88%, and F1 score was 88%. These findings confirm that the proposed system is proper for crime activity classification.

5. Conclusions

Crime detection systems in surveillance videos utilizing deep learning approaches represent an area of attention-grabbing research for mitigating criminal activities and hence developing a peaceful community. Many systems have previously evolved to reduce levels of crime activity, however, there is still much that can be done to improve their performance. In this paper, the proposed system implemented a developed deep-learning approach for crime activity detection. With the UCF Crime dataset, the proposed system attained the lowest loss and the highest accuracy on the training and validation corpus. The proposed system especially had difficulty in classifying the abnormal class within the DCSASS dataset, and the main reason for this may be that the DCSASS dataset partitions the abnormal UCF crime videos into abnormal and normal. However, concerning the proposed CNN Approach, the structure implemented effectively detects and differentiates crime

activities. From this, we can infer that the structure modeling parameters provided an outstanding distinction in classifying various criminal classes. In the experiments, the attained results of the performance measures conclude that the proposed approach outperformed the other implemented approaches (EfficientNet-B7, and MobileNet-V2) and other relevant works for detecting criminal activities.

Future initiatives revolve around collecting new annotated datasets that effectively depict various crime activities, and developing approaches (like 3D-CNNs) capable of processing and capturing temporal information video feeds in real time without compromising accuracy. Additionally, it is preferable to exploit developed image processing and incorporate Large Language Models (LLMs) into crime activity detection systems. Together, these endeavors aim to create responsive, accurate, and efficient systems that excel at detecting crimes across various and challenging environments. Furthermore, we will concentrate on combining multi-modal data (like sensor data and audio) with video frames.

References

[1] K. Hu, L. Li, X. Tao, J. D. Velásquez, P. Delaney, "Information fusion in crime event analysis: A decade survey on data, features and models", Information Fusion, vol. 100, 2023.

[2] F. J. Rendón-Segador, J. A. Álvarez-García, J. L. Salazar-González, T. Tommasi, "CrimeNet: Neural Structured Learning using Vision Transformer for violence detection", Neural Networks, vol. 161, pp. 318-329, 2023.

[3] K. B. Sahay, B. Balachander, B. Jagadeesh, G. A. Kumar, R. Kumar, L. R. Parvathy, "A real time crime scene intelligent video surveillance systems in violence detection framework using deep learning techniques", Computers and Electrical Engineering, vol. 103, 2022.

[4] F. L. Sánchez, I. Hupont, S. Tabik, F. Herrera, "Revisiting crowd behaviour analysis through deep learning: Taxonomy, anomaly detection, crowd emotions, datasets, opportunities and prospects", Information Fusion, vol. 64, pp. 318-335, 2020.

[5] J. Waleed, A. T. Azar, S. Albawi, W. K. Al-Azzawi, I. K. Ibraheem, A. Alkhayyat, I. A. Hameed, N. A. Kamal, "An Effective Deep Learning Model to Discriminate Coronavirus Disease From Typical Pneumonia," International Journal of Service Science, Management, Engineering, and Technology (IJSSMET), vol. 13, no.1, pp. 1-16, 2022..

[6] A. J. Yousif and M. H. Al-Jammas, “Real-time Arabic Video Captioning Using CNN and Transformer Networks Based on Parallel Implementation ”, Diyala Journal of Engineering Sciences, vol. 17, no. 1, pp. 84–93, Mar. 2024.

[7] S. Albawi, M. H. Arif, J. Waleed, "Skin cancer classification dermatologist-level based on deep learning model", Acta Scientiarum. Technology, vol. 45, no. 1, e61531, 2022.

[8] S. Vosta, K.-C. Yow, "A CNN-RNN Combined Structure for Real-World Violence Detection in Surveillance Cameras", Applied Sciences, vol.12, no. 3, 2022.

[9] B. Asal, A. B. Can, "Ensemble-Based Knowledge Distillation for Video Anomaly Detection", Applied Sciences, vol. 14, no. 3, 2024.

[10] W. Ullah, A. Ullah, T. Hussain, Z. A. Khan, S. W. Baik, "An Efficient Anomaly Recognition Framework Using an Attention Residual LSTM in Surveillance Videos", Sensors, vol. 21, no. 8, 2021.

[11] A. Gulati, S. Pawar, P. Sheoran, C. Sali, "SUSPICIOUS BEHAVIOUR DETECTION USING CNN AND ROLLING AVERAGE", International Research Journal of Modernization in Engineering Technology and Science, vol.3, no.5, 2021.

[12] K. V. Thakare, N. Sharma, D. P. Dogra, H. Choi, I.-J. Kim, "A multi-stream deep neural network with late fuzzy fusion for real-world anomaly detection", Expert Systems with Applications, vol. 201, 2022.

[13] P. Khaire, P. Kumar, "A semi-supervised deep learning based video anomaly detection framework using RGB-D for surveillance of real-world critical environments", Forensic Science International: Digital Investigation, vol. 40, 2022.

[14] M. Qasim, E. Verdu, "Video anomaly detection system using deep convolutional and recurrent models", Results in Engineering, vol. 18, 2023.

[15] J. Park, J. Kim, B. Han, "End-to-end learning for weakly supervised video anomaly detection using Absorbing Markov Chain", Computer Vision and Image Understanding, vol. 236, 2023.

[16] A. Patwal, M. Diwakar, V. Tripathi, P. Singh, "An investigation of videos for abnormal behavior detection", Procedia Computer Science, vol. 218, pp. 2264-2272, 2023.

[17] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. -C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 4510-4520, 2018.

[18] M. Tan, Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks", International Conference on Machine Learning, 2019.

[19] W. Sultani, C. Chen, M. Shah, "Real-World Anomaly Detection in Surveillance Videos," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA , pp. 6479-6488, 2018.

[20] https://www.kaggle.com/datasets/mateohervas/dcsass-dataset.

[21] J. Waleed, S. Albawi, H. Q. Flayyih and A. Alkhayyat, "An Effective and Accurate CNN Model for Detecting Tomato Leaves Diseases," 2021 4th International Iraqi Conference on Engineering Technology and Their Applications (IICETA), Najaf, Iraq, pp. 33-37, 2021

[22] S. Albawi, J. Waleed and A. J. Abboud, "Deep CNN-Based-Flower Species Recognition System," 2023 3rd International Scientific Conference of Engineering Sciences (ISCES), Diyala, Iraq, pp. 54-58, 2023.