ConvGRU-CNN: Spatiotemporal Deep Learning for Real-World Anomaly Detection in Video Surveillance System

Maryam Qasim Gandapur[1]*, Elena Verdú[2]

[1] Department of Law, Shaheed Benazir Bhutto University, Dir (Upper), Khyber Pakhtunkhwa (Pakistan)
[2] Universidad Internacional de La Rioja, Logroño, La Rioja (Spain)

Received 14 April 2022 | Accepted 16 August 2022 | Published 30 May 2023

Abstract

Video surveillance for real-world anomaly detection and prevention using deep learning is an important and difficult research area. It is imperative to detect and prevent anomalies to develop a nonviolent society. Real-world video surveillance cameras automate the detection of anomaly activities and enable the law enforcement systems for taking steps toward public safety. However, a human-monitored surveillance system is vulnerable to oversight anomaly activity. In this paper, an automated deep learning model is proposed in order to detect and prevent anomaly activities. The real-world video surveillance system is designed by implementing the ResNet-50, a Convolutional Neural Network (CNN) model, to extract the high-level features from input streams whereas temporal features are extracted by the Convolutional GRU (ConvGRU) from the ResNet-50 extracted features in the time-series dataset. The proposed deep learning video surveillance model (named ConvGRU-CNN) can efficiently detect anomaly activities. The UCF-Crime dataset is used to evaluate the proposed deep learning model. We classified normal and abnormal activities, thereby showing the ability of ConvGRU-CNN to find a correct category for each abnormal activity. With the UCF-Crime dataset for the video surveillance-based anomaly detection, ConvGRU-CNN achieved 82.22% accuracy. In addition, the proposed model outperformed the related deep learning models.

Keywords

Anomaly Activities, Crime Detection, ConvGRU, Convolutional Neural Network (CNN), Deep Learning, Video Surveillance.

DOI: 10.9781/ijimai.2023.05.006

I. Introduction

With the growing public safety and security challenges, demand for increasing public safety monitoring through video surveillance cameras is also growing. Human-monitored surveillance systems can mine critical and helping cue from the patterns. This can help in detecting the abnormal activities for instant reaction [1]. However, owing to the human-monitored limitations, it is difficult to mine critical and helping cues [2]. Thus, an automated method to detect abnormal activities is critical. A sub-domain to understand behaviour from the video surveillance cameras is to detect anomaly activities [3]. The anomaly detection in the video surveillance is a crucial task and can face difficulties such as actions which do not tail definite patterns are termed as anomalies. Furthermore, actions are abnormal or normal in different situations indicating that a global abnormal activity can be a usual activity in certain situations such as gun club shooting. The shooting is usually an abnormal activity, but a normal activity in shooting clubs. Alternatively, some behavior is not essentially abnormal, but might be anomalies in different situations [4]. According to some studies [5]-[6], abnormal actions ended at unusual locations and times.

Several kinds of abnormal activities are usually identified which include killing, looting, molestation, and intensive attacks. Killing is a deliberate action to kill a person. Looting is an action of stealing belongings from the people using extreme physical force and violence. Molestation is sexual exploitation of people (man, woman, and children) against their desire. This criminal activity is terrible and shows substantial consequences. Intensive attacks are illegal fights by one person against another to get something or to harm individuals [7]. Anomaly detection and prevention using deep learning is an attention-grabbing system. Many law enforcement organizations across the globe are experimenting deep learning systems to safeguard public safety. The anomaly activities are predictable and require high volume data processing, exposing the anomaly patterns which are informative for a law enforcement department. In some situations, an anomaly activity remains unreported because of external pressures from all verticals of society. For this reason, an intelligent security system is able to autonomously detect anomaly activities and supports in excluding manipulative activities by bypassing individuals and informing law enforcement departments. For example, there is a case

* Corresponding author.
E-mail address: maryam@sbbu.edu.pk

study of San-Francisco in USA and Natal in Brazil where anomaly activities have been predominant and monitored by the intelligent video surveillance systems [8].

Video camera surveillance is a key feature of monitoring systems [9]. Computer vision automates anomaly activity detection in videos by alarming a law reinforcement system when abnormal activity is observed to derive important information from recorded videos [10]-[11]. Various anomaly activities while entering or departing the public places require careful examination which might be important towards a pattern of anomaly activity [12]. In few situations, previous patterns can recognize malicious individuals in the recorded videos [13]-[14]. We can automate feedback activities when anomalies are observed to derive information from recorded videos using deep learning models [15]. According to deep learning perspective, the detection of the anomaly actions is divided into supervised, unsupervised, and semi-supervised learning models. In a single deep learning model, the model is trained on normal or abnormal activities [16]. On the other hand, both normal and abnormal activities are used to train deep learning models in multi-model learning setting [17]. Several studies took advantage of the supervised deep learning to detect anomaly activities in videos [18]-[25]. Many deep learning models including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long-Short Term Memory (LSTM), Gated Recurrent Units (GRUs), and Generative Adversarial Networks (GANs) are used for anomaly detection and prevention [26]-[28].

This study proposes anomaly activity detection in a multiple-learning perspective using a supervised deep learning model. Numerous abnormal activities in real-world are labeled as anomalies; but the focus of this study is on anomaly activities mentioned in the UCF-Crime dataset [29] which includes abnormal and violent behavior recorded by the video surveillance cameras in various public places. The proposed deep learning model for anomaly detection and prevention has implemented the ResNet-50 as a CNN model to extract high-level features from video frames. The CNN extracted features are fed to RNN model, ConvGRU, to learn temporal dependencies in the video dataset. The proposed deep learning model ConvGRU-CNN returns output indicating whether input videos include abnormal or normal behaviour. This ConvGRU-CNN model can reduce limitations of human-monitored video surveillance systems and can improve the accuracy of anomaly activity detection. In addition, ConvGRU-CNN can considerably improve the response-time. A compact neural model for anomaly detection is proposed by implementing a convolutional form of conventional GRU to learn temporal features videos. Alternative to the fully connected layer in GRU, a convolutional layer intensely reduces the parameters number. In addition, the incorporation of GRU further reduces the parameters when replaced with LSTM in ConvLSTM. There is 25% further reduction in parameters with ConvGRU. With UCF-Crime dataset, 13 classes of abnormal events are used to evaluate the proposed ConvGRU-CNN.

II. Related Studies on Video Surveillance Systems

The goal of anomaly detection system is to predict and prevent the abnormal (criminal) activities. Though, the conventional non-deep learning approaches are beneficial but they operate independently. Hence, a machine which is able to integrate the important aspects of conventional approaches would extremely be advantageous. A study [28] has compared the violent criminal patterns between the community's dataset and the real criminal statistical data by using Waikato Environment for Knowledge Analysis (WEKA) platform. Three models including the linear regression, additive regression, and decision stump are implemented. The linear regression on selecting the random samples in testing was able to handle randomness showing a better detection among models and proved the success of deep learning in detecting the violent patterns and criminal trends.

A study [30] examined the anomaly detection in urban areas where anomaly has combined to grid size 200×250m and examined retrospectively. An ensemble model of logistic regression and neural network is proposed to detect anomaly. The results indicate that fortnightly predictions are improved remarkably as compared to monthly predictions. Anomaly activities are detected and examined in another work [31] using anomaly data of Vancouver for the last 15 years. A boosted decision tree and K-nearest neighbor (KNN) detected anomaly activities. A total of 560,000 records are examined and the anomaly activities are predicted with accuracy between 39% and 44%.

Another study [32] predicted anomaly statistics in Philadelphia to determine the trends of anomaly. Ordinal regression, KNN, logistic regression, and decision tree are trained with the datasets to get anomaly predictions. The models were able to determine the trend of anomaly activity with an accuracy of 69%. Data science models are implemented to detect the anomaly activities from the Chicago criminal dataset. Logistic regression, SVM/KNN classification, decision trees, random forest, and Bayesian models were examined and the most accurate model was selected for training. The KNN classification obtained the best accuracy of 78.7%.

A GUI-based deep learning model to predict the anomalies is presented in another study [33]. The results of supervised models are compared to predict anomalies. A feature-level data-fusion-based deep neural network (DNN) is proposed to predict anomaly with high accuracy by combining multi-model data from different domains with environmental context knowledge [34]. The data to train models (SVM, regression analysis, Kernel density estimation) was taken from online crime statistic database. SVM and KDE obtained 67.01% and 66.33% accuracies, whereas the proposed model obtained 84.25% accuracy. Another work [5] used previous crime locations to predict anomaly likely to happen in old locations. Bayesian neural networks, Levenberg Marquardt algorithm, and a scaled algorithm are implemented to examine and understand the data. The scaled algorithm showed the best results. The ANOVA verified that the scaled algorithm reduced crime rate by 78%, with 0.78 accuracy.

A framework to predict anomaly has been proposed [35] examining a dataset of formerly committed anomalies with their patterns. KNN and decision tree with adaptive boosting and random forest are implemented to boost the prediction accuracy. The records are divided into rare and frequent classes. The deep learning framework was trained with criminal activities recorded in a period of 12-years in San Francisco, USA. By applying oversampling and undersampling with random forest, 99.2% accuracy was achieved. Other studies have also achieved state of the art results for crime detection [36]-[44]. Table I summaries the previous work with achieved accuracies.

TABLE I. Summary of Previous Work on Crime Detection

S. No.	Reference	Model	Achieved Accuracy
1	[36]	Decision Tree	59.15%
2	[37]	KNN	87.03%
3	[38]	Naive Bayes	87.00%
4	[39]	ARIMA	86.00%
5	[40]	Regression Model	72.00%
6	[41]	SVM	84.30%
7	[42]	Random Forrest	97.00%
8	[43]	E2E-VSDL	98.16%

Various CNN typed have been formulated such as the AlexNet, ResNets, VGG, Inceptions and their variants. Many studies combined these CNNs with a softmax layer [45], and morphological analysis [46] to detect anomaly. Besides CNN, other studies [47]-[48] proposed

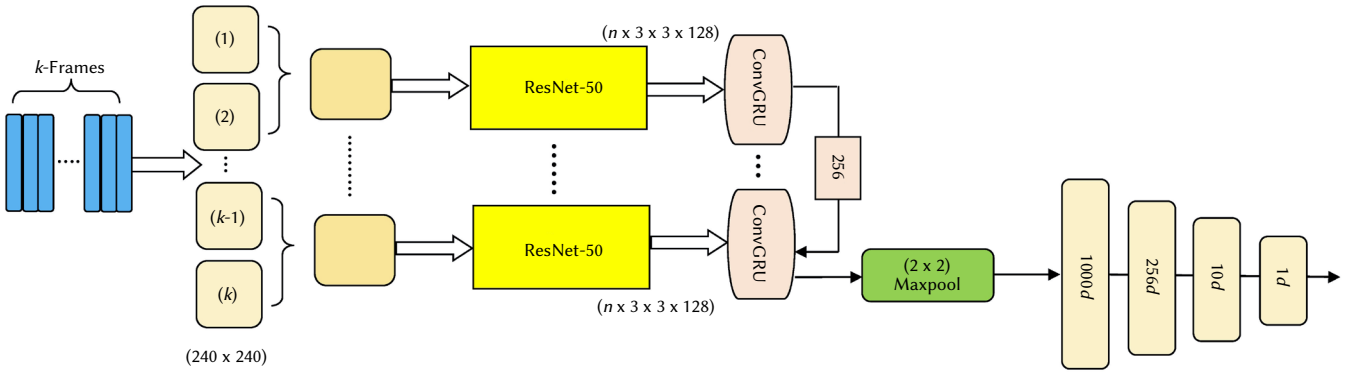


Fig. 1. The structure of the proposed ConvGRU-CNN.

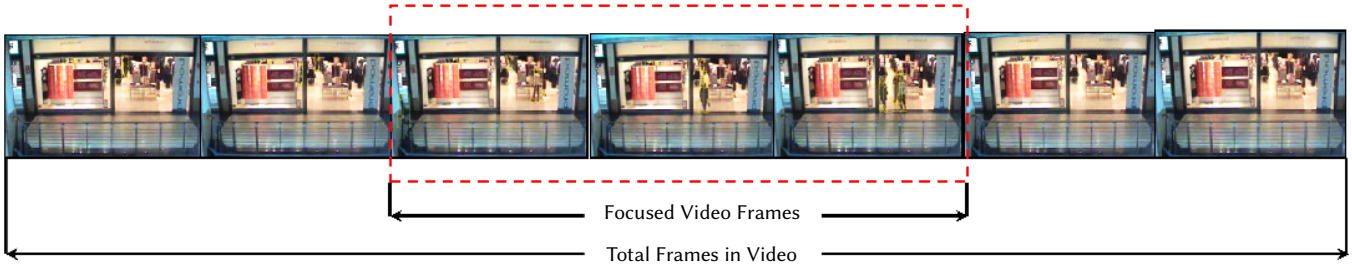


Fig. 2. Focused Bag frames extraction.

using autoencoders. Bayesian nonparametric [49] is also proposed to detect abnormal events in videos. Since surveillance camera feeds are sequential data, the LSTMs have gained attraction for anomaly detection. Encoder-decoder LSTM [50] is proposed in an unsupervised learning fashion. Spatiotemporal networks (STNs) are gaining popularity to learn spatial and temporal features [51] where RNNs and CNNs jointly extract spatiotemporal features for anomaly detection. ConvLSTM [52] is another model where a convolutional layer filters the output of CNNs before feeding a LSTM. Alternative to the fully connected layer in LSTM, a convolutional layer intensely reduces the parameters number. The GRUs further reduces the parameters if replaced with LSTM in ConvLSTM, obtaining a 25% reduction in the parameters. There are very limited studies implementing ConvGRU to detect anomaly in video streams.

III. Proposed Anomaly Detection Model

Residual Networks (ResNets) are effective neural models to extract features in DNNs [53]. First, ResNet-50 is implemented to extract spatial features from the input video streams. In the next stage, the ConvGRU as RNN is used to extract the temporal dependencies in videos. The video streams are divided into sequences of k frames and fed to ResNet-50 as inputs. The outputs are further fed to ConvGRU. The spatiotemporal features are passed through maxpooling and fully connected layers to detect the anomaly.

A. Video Pre-Processing

Fig. 1 shows the proposed ConvGRU-CNN where input video is preprocessed and divided into fixed frames k with 30 frames/second. Therefore, for 60 sec video, the total number of frames is 1800. To consider the spatial movements for all input frames after selection, the difference between every frame and adjacent frames is calculated. Three categories from UCF-Crime dataset are selected. We split the exact time of the abnormal activity for each video and labelled them as Anomaly, such that, the remaining video is labelled as Normal. After that, the videos are divided into the same length. As a result, n frames are selected from k frames. Thus, only abnormal activities are focused. Further, the normal activities are also selected from the same videos which include the anomaly activities. Except for actions, all other setting remains the same as in UCF-Crime dataset. Such arrangements help system in better detecting the anomaly activities. Full-length training videos result in a massive computational cost. Therefore, to understand the motion information in the recorded videos during training, we have considered a training framework over a defined set of frames, that is, focused bag which contains major information needed to understand the motions in the videos followed by block formation and selection. A set of frames composed of the activities in full length recorded video has been named as the focused bag and its extraction is shown in Fig. 2 where only a small part of the full-length recorded video is labelled as the suspicious/ criminal activity. Hence, L-frames out of M-frames are considered as a focused bag. This entire procedure is adopted and repeated for all recorded videos in the database thereby significantly minimized the training data by removing the redundant information.

B. ResNet-50 CNN Model

ResNets have shown excellent performance on many standard datasets such as ImageNet [16]. ResNets have many variants, such as, ResNet-18, ResNet-26, ResNet-50, ResNet-101, and ResNet-152. However, because of better performance and excellent architecture, ResNet-50 is instigated in ConvGRU-CNN. To avoid difficulties in labelling anomalies, Transfer Learning [54] is used in the model. As a result, ConvGRU-CNN is pre-trained on the ImageNet dataset, which includes 1000 sets of images. By executing ResNet-50 on ImageNet, the model parameters are initialized and updated thereby ready to execute on the preferred datasets. The input frame size is (240×240) allowing the ResNet-50 to process (240×240×3) dimension data. After passing through the convolutional and pooling layers, a 4-d tensor (n×1×1×2048) output is obtained from the Deep Residual Features (DRF), which is reshaped before fed to the ConvGRU filters. ResNet-50 structure is shown in Fig. 3 whereas the architecture is given in Table II. The ResNet50 output is reshaped into (n×3×3×128) and is fed to ConvGRU layer. Since ResNet is not using for classification, the fully connected dense layer is not utilized.

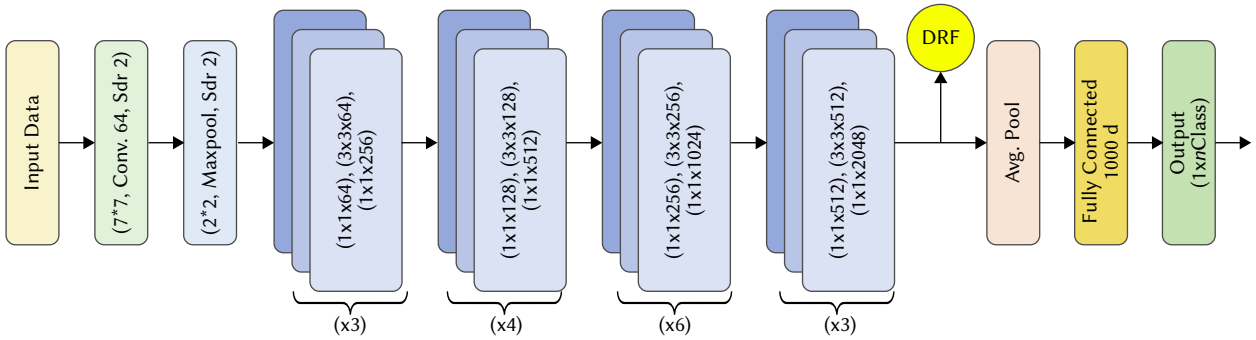


Fig. 3. ResNet-50 Structure.

TABLE II. ResNet-50 Architecture

Layer Name	Output Size	50-Layers Model
Conv1	(112×112)	(7×7), 64, Stride 2 (3×3), Maxpool, Stride 2
Conv2	(56×56)	[(1×1, 64), (3×3, 64), (1×1, 256)]×3
Conv3	(28×28)	[(1×1, 128), (3×3, 128), (1×1, 512)]×4
Conv4	(14×14)	[(1×1, 256), (3×3, 256), (1×1, 1024)]×6
Conv5	(7×7)	[(1×1, 512), (3×3, 512), (1×1, 2048)]×3
	(1×1)	Average Pool, 1000d Fully Conn., Softmax
FLOPS		3.8 × 109

C. ConvGRU Layer

The cell inputs, outputs, and states in GRUs are 1-d vectors; therefore, GRUs is unable to hold spatial relations between video pixels. As a result, GRUs is inappropriate for spatial sequence data [55]. In ConvGRU, due to the convolutional layers, cell states/inputs/outputs, and the spatial dimensions are 3-d tensors. Since the ConvGRU structure consists of convolutional gates, it can deal with spatial and temporal sequential data. The ConvGRU is a regular GRU but replaces the matrix multiplication with convolution operations. With convolution operations, the GRU can preserve spatial information. The formulation of the ConvGRU simply takes the standard linear GRU as:

$$z_t = \sigma_g(W_z x_t + U_z h_{t-1} + b_z) \quad (1)$$

$$r_t = \sigma_g(W_r x_t + U_r h_{t-1} + b_r) \quad (2)$$

$$c_t = \rho_g(W_c x_t + r * U_h h_{t-1} + b_z) \quad (3)$$

$$h_t = (z_t \odot h_{t-1} + (1 - r_t) \odot c_t) \quad (4)$$

Where $\mathbf{W}_z$, $\mathbf{W}_r$, $\mathbf{W}_c$ are weight matrices, b_z, b_r, b_c are biased terms, respectively, whereas x_t is input state [55]. By replacing the matrix multiplication with convolution operations (denoted as *), Eq. (1) - (4) became:

$$z_t = \sigma_g(W_z * x_t + U_z * h_{t-1} + b_z) \quad (5)$$

$$r_t = \sigma_g(W_{r*} x_t + U_r * h_{t-1} + b_r) \quad (6)$$

$$c_t = \rho_g(W_c * x_t + r * U_{h*} h_{t-1} + b_z) \quad (7)$$

$$h_t = (z_t \odot h_{t-1} + (1 - r_t) \odot c_t) \quad (8)$$

From the complexity viewpoint, GRU operates at 1-d vectors and after Hadamard product, the complexity increases due to large parameters size thereby the model is prone to overfitting. However, ConvGRU has a unique internal structure and requires fewer parameters which reduce the computational complexity of model. The video frames, after passing ResNet-50, feed the ConvGRU cell composed of 256 hidden states with (3×3) kernel size. The input to ConvGRU is a 4-d tensor, (n×256×3×3) such that input at each time-step is (3×3) with 256 channels. The output of ConvGRU is maxpooled with (2×2) size and flattened to get a 1-d vector. The 1-d vector feed the fully-connected layers followed by batch normalization (BN) and ReLU activation. For binary classification (normal vs abnormal activity), sigmoid activation and binary cross entropy loss can be used after fully-connected layers. However, softmax with categorical cross entropy loss can be used for multi-class classification. The working flow of the proposed ConvGRU-CNN is given in Fig. 1.

IV. Experiments

A. Dataset

In this paper, the proposed model is implemented on the UCF-Crime dataset [29] which includes abnormal, illegal and violent behaviour recorded by surveillance video cameras located in public places such as stores and streets. The UCF-Crime dataset is prepared from everyday actual events which is the key reason to select this dataset. Many studies have used handicraft datasets or particular datasets with the same backgrounds and environments (for example fighting and movies dataset), which is not according to our daily life. The UCF-Crime dataset is including lengthy surveillance video cameras feeds covering 13 different classes of anomaly events such as the Abuse, Arrest, Arson, Assault, Road Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, and Vandalism in addition to Normal events class. Fig. 4 shows example samples of the UCF-Crime dataset. For comparison with other related studies the training and testing data is arranged as (75%-25%) in experiments. Two variants of the UCF-Crime dataset including Ucfcrimes and Binary are used in the experiments where the Ucfcrimes contains 14 classes whereas Binary has 2 classes, one compiling the 13 abnormal activities and the normal one. The quantity of videos for each class from the Ucfcrimes and Binary datasets are given in Table III.

B. Model Settings and Model Selection

In the experiments the proposed model is applied by using ResNet-50 and ConvGRU, which are available in the Keras library. To tune the model, several hyperparameters are used to attain the best performance. Table IV shows the results of experiments with different types of weight initialization and optimizers. As a result, to initialize the ConvGRU-CNN weights, glorot-uniform (Xavier) is utilized whereas to optimize the model, RMSprop optimizer is imposed. The learning rate and number of epochs are fixed to 0.0001 and 100, respectively. However, early stop is applied when loss converges. The video sequence length is fixed to 20 frames. Since, focused bag is used for video frames. Table V provides a comparison of total video frames and frames in focus bag. In our experiments, we used different kinds of evaluations. During the first step, ConvGRU is tested with several CNN models such as InceptionV3, VGG19, ResNet-50, ResNet-101,

TABLE III. Number of Videos for Ucfcrimes and Binary Datasets

Anomalies	Videos (Ucfcrimes)	Videos (Binary)	Anomalies	Videos (Ucfcrimes)	Videos (Binary)
Abuse	50	50	Road Accident	50	150
Arrest	50	50	Robbery	50	150
Arson	50	50	Shooting	50	50
Assault	50	50	Shoplifting	50	50
Burglary	50	100	Stealing	50	100
Explosion	50	50	Vandalism	50	50
Fighting	50	50	Normal	50	950
Total	350	400	Total	350	1500

and ResNet-152, available in the Keras library. Table VI shows a comparison in terms of accuracy (in %). According to accuracy with less computational complexity, ResNet-50 was selected for integration with ConvGRU.

TABLE IV. Hyperparameters Settings

Hyper Parameters	Tuning	Accuracy (%)
Weights Initialization	Glorot-Uniform (Xavier)	81.9%
Weights Initialization	Random-Uniform	80.2%
Weights Initialization	He-Uniform	80.2%
Optimizer	Adam	80.9%
Optimizer	RMSprop	81.3%

TABLE V. Video Frames Analysis (in Example Anomalies)

Anomaly	Total Video Frames	Focused Video Frames
Assault	130395	55023
Fighting	269255	132517
Shooting	157735	55427
Vandalism	157511	82163
Total	714896	156610

TABLE VI. CNN+ConvGRU Comparison

Hyper Parameters	Weights Initialization	Accuracy (%)
ResNet-50+ConvGRU	Glorot-Uniform (Xavier)	82.6%
ResNet-101+ConvGRU		RMSprop
ResNet-152+ConvGRU		86.3%
InceptionV3+ConvGRU	Adam	82.5%
VGG19+ConvGRU	RMSprop	89.3%

The evaluation measures are given by equations as:

$$\text{Accuracy: } \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$\text{Sensivity (Recall): } \frac{TP}{TP + FN} \tag{10}$$

$$\text{F1} - \text{Score: } \frac{2 * TP}{2 * TP + FP + FN} \tag{11}$$

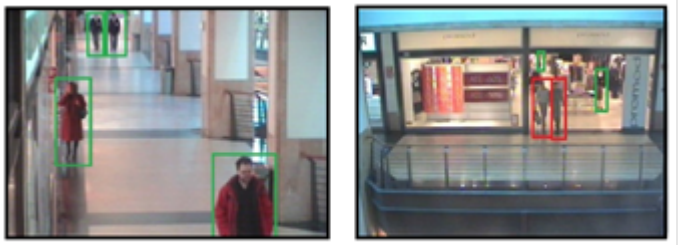
Where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative.

V. Results and Discussions

First, the proposed ConvGRU-CNN model is examined by measuring the accuracy (Acc), precision (Prc), and F1-scores. Table VII provides the Acc, Prc, and F1 scores. It is clear from the Table VII that the proposed ConvGRU-CNN achieved significant metric scores for the 14 categories of the UCF-Crime dataset. The measuring results are averaged over the 14 types of activities, and the best accuracy obtained is 82.22%. In addition, good precision and F1 are achieved with this considerable number of anomaly categories. The proposed model attained an encouraging average accuracy, precision, and F1-score of 82.88%, 82.89%, and 82.88%, respectively, at reducing the computational complexity. Therefore, an efficient model is proposed to analyze spatiotemporal features extracted from videos. Fig. 4 shows the detection of suspicious activity.

TABLE VII. Model Evaluation in Terms of Acc, Prc, and F1 Scores

Database	Accuracy	Precision	F1	AUC
Ucfcrimes	82.22%	83.13%	82.22%	82.65%
Binary	83.54%	85.65%	83.55%	82.77%
Average	82.88%	82.89%	82.88%	82.71%



Normal Activity Suspecious (Criminal) Activity

Fig. 4. Detection of Normal and Suspicious Activities.

A. Comparison With Other Models

Limited literature on anomaly detection by using the UCF-Crime dataset is available. In the experiments, the proposed ConvGRU-CNN model is compared with other CNN models by measuring the Accuracy (Acc) and Area Under the Curve (AUC). Table VIII provides the AUC scores for the binary classification on the UCF-Crime dataset for the proposed model and other models for anomaly detection. The related models include support vector machine (SVM) [56], MIL [29], 3D-CNN [11], TSN [51], AutoEncd [48], SCL [57], CNN-RNN [58], and UGD-KM [59]. The categories for all the above mentioned abnormal events are considered as the Anomaly category whereas data with no abnormal events is considered as Normal. The testing classifier indicates the probabilities of correctly classified anomaly events. Table VIII shows that the proposed ConvGRU-CNN model outperformed the related benchmark models in anomaly detection. For example, AUC score is improved from 50.10% with SVM to 82.65% with ConvGRU-CNN and achieved 32.65% AUC gain. Similarly, AUC with AutoEncd is improved from 50.6% to 82.65% with large performance gain of 32.05%. Fig. 5 the shows performance improvement over competing models. In comparison to 3D-CNN, the proposed ConvGRU-CNN improved the AUC from 81.05% to 82.65% whereas with MIL, the AUC is improved by 8.21%. The t-SNE results for normal and criminal activities are illustrated in Fig. 6.

TABLE VIII. Model Comparison in Terms of AUC Scores

Reference	Models	AUC (%)
(Erfani et al., 2016) [56]	SVM	50.10%
(Hasan et al., 2016) [48]	AutoEncd	50.60%
(Sultani et al., 2018) [29]	MIL (Loss with No Constraints)	74.44%
(Sultani et al., 2018) [29]	MIL (Loss with Constraints)	75.41%
(Zhong et al., 2019) [51]	TSN	78.08%
(Tran et al., 2015) [11]	3D-CNN	81.01%
(Vosta and Yow, 2022) [58]	CNN-RNN	81.77%
(Khan et al., 2018) [59]	UGD-KM	64.30%
Proposed	ConvGRU-CNN	82.65%

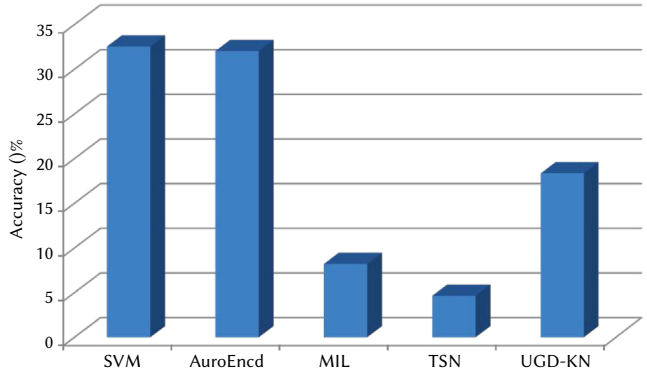


Fig. 5. percentage improvement over competing models.

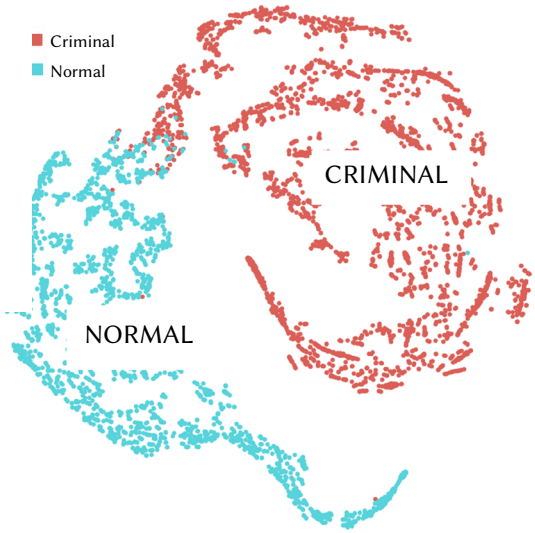


Fig. 6. t-SNE plots for normal and criminal activities.

To observe the efficiency of ConvGRU for crime detection, we have implemented ConvLSTM for the said task, and compared the accuracy, precision, F1, and AUC. This set of experiment was performed on the same experimental settings as done for ConvGRU. Table IX shows the results of the study. The results indicate that ConvLSTM underperforms in terms of accuracy, AUC, and computational cost, respectively.

TABLE IX. Comparison With LSTM in Terms of Accuracy, AUC and Computational Cost

Database	Accuracy	AUC	Computational Cost
ConvLSTM	81.93%	81.98%	25% Less computational complexity with ConvGRU
ConvGRU	82.88%	82.71%	

VI. Laws Preventing the Anomaly/Criminal Events

Anomaly events including criminal events and criminal intimidation has raised with hike in inflation in Pakistan and across the globe. It is imperative to devise new effective ways for preventing them. The conventional ways to detect the anomaly patterns are taking their part but technology has moved forward. A study [60] recommended deep learning models to the law enforcement agencies for predicting, detecting, and solving the anomaly activities at higher rates to reduce the crimes in society. So, it is recommended to use Artificial Intelligence (AI) to prevent the criminal events before their happening. The governments and other relevant institutes are responsible to prevent and reduce the criminal rate. As a result, the modern technology is one of the major solutions. To curtail this issue, law making authorities that is the legislature enacted Prevention of Electronic and other Crimes Act provided a detailed legal framework relating to different types of electronic crimes, procedures for the investigation and prosecution. Any act which is forbidden by the penal laws/ laws of the land amounts to criminal acts liable to be punished. With technological advancement, the already existing criminal patterns can be learnt to avoid future crimes and hence the punishment will become easy for law enforcement departments. To curtail heinous crimes, Serious Crimes Prevention Order is provided in Serious Crime Act 2007 in Pakistan. It is an adjudication order to safeguard public at large by preventing and restricting an individual participation in crimes.

VII. Conclusions, Limitations, and Future Work

This study proposes a novel deep learning framework by linking ResNet-50 and ConvGRU for detecting anomaly activities in the UCF-Crime dataset. Some anomalies took place in videos where persons cannot be seen such as car accidents. Besides, many anomaly events happen for few seconds and in a small length video (10 sec), most part of such videos shows a normal event. Regardless of stated limitations, the ConvGRU-CNN model outscores other models on the UCF-Crime dataset with 82.65% AUC and 82.88% accuracy. In addition to 14 classes of the UCF-Crime dataset, dividing the dataset into two major classes (Ucfcrimes and Binary) shows improved results in terms of accuracy and AUC. The focused video frames extracted from the original videos of anomaly events have greatly improved the detection accuracy. Among other CNN models implemented for anomaly event, ResNet-50 [61] provides improved results when combined with ConvGRU. An excellent features extraction with ResNet-50 and ConvGRU significantly improved the performance measures. With ResNet-50+ConvGRU, the classifier efficiently detected the anomaly classes for Ucfcrimes and Binary datasets. The experimental results demonstrate that ConvGRU-CNN performed better than other related models in terms of accuracy, precision, and AUC, yet we look to improve classification of all kinds of anomalies in the UCF-Crime dataset. One of the approaches is to add attention layers to the ConvGRU-CNN as future work. The attention layer is possible to integrate with CNN and/or ConvGRU. With this approach, the future model can be focused more precisely on the anomaly events in a video. Silent videos can be used to detect anomaly in terms of audio signals since most of the time only silent video is available. Therefore, if we incorporate audio signal synthesis, video surveillance can be made more effective [62].

References

[1] T. Hospedales, S. Gong, and T. Xiang, "Video behaviour mining using a dynamic topic model," *International journal of computer vision*, vol. 98, no. 3, pp. 303-323, 2012.

[2] M. Cristani, R. Raghavendra, A. Del Bue, and V. Murino, "Human

behavior analysis in video surveillance: A social signal processing perspective," *Neurocomputing*, vol. 100, pp. 86-97, 2013.

[3] B. Tian, B.T. Morris, M. Tang, Y. Liu, Y. Yao, C. Gou, ... and S. Tang, "Hierarchical and networked vehicle surveillance in ITS: a survey," *IEEE transactions on intelligent transportation systems*, vol. 16, no. 2, pp. 557-580, 2017.

[4] J. Yu, K.C. Yow, and M. Jeon, "Joint representation learning of appearance and motion for abnormal event detection," *Machine Vision and Applications*, vol. 29, no. 7, pp. 1157-1170, 2018.

[5] S. R. Bandekar and C. Vijayalakshmi, "Design and analysis of machine learning algorithms for the reduction of crime rates in India," *Procedia Computer Science*, vol. 172, no. 122-127, 2020.

[6] J. Varadarajan and J. M. Odobez, "Topic models for scene analysis and abnormality detection," in *2009 IEEE 12th International Conference on Computer Vision Workshops*, ICCV Workshops, IEEE, 2009, pp. 1338-1345.

[7] C. Tabedzki, A. Thirumalaiswamy, P. van Vliet, S. Agarwal and S. Sun, "Yo home to Bel-Air: predicting crime on the streets of Philadelphia," University of Pennsylvania, CIS, 520, 2018.

[8] K. A. Joshi and D.G. Thakore, "A survey on moving object detection and tracking in video surveillance system," *International Journal of Soft Computing and Engineering*, vol. 2, no. 3, pp. 44-48, 2012.

[9] R. Socha and B. Kogut, "Urban video surveillance as a tool to improve security in public spaces," Sustainability, vol. 12, no. 15, 6210, 2020.

[10] A. Selvaraj, J. Selvaraj, S. Maruthaiappan, G.C. Babu and P.M. Kumar, "L1 norm based pedestrian detection using video analytics technique," *Computational Intelligence*, vol. 36, no. 4, pp. 1569-1579, 2020.

[11] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489-4497.

[12] L. Alkanhal, D. Alotaibi, N. Albrahim, S. Alrayes, G. Alshemali and O. Bchir, "Super-resolution using deep learning to support person identification in surveillance video," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 7, 2020.

[13] J. Athanesious, V. Srinivasan, V. Vijayakumar, S. Christobel and S.C. Sethuraman, "Detecting abnormal events in traffic video surveillance using superorientation optical flow feature," *IET Image processing*, vol. 14, no. 9, pp. 1881-1891, 2020.

[14] H. Zhang, P. Li, Z. Du and W. Dou, "Risk entropy modeling of surveillance camera for public security application," *IEEE Access*, vol. 8, pp. 45343-45355, 2020.

[15] B.S. Harish and S.A. Kumar, "Anomaly based Intrusion Detection using Modified Fuzzy Clustering," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 6, pp. 54-60, 2017.

[16] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, IEEE, 2009, pp. 248-255.

[17] A.A. Sodemann, M.P. Ross and B.J. Borghetti, "A review of anomaly detection in automated surveillance," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 1257-1272, 2012.

[18] I.V. Pustokhina, D.A. Pustokhin, T. Vaiyapuri, D. Gupta, S. Kumar and K. Shankar, "An automated deep learning based anomaly detection in pedestrian walkways for vulnerable road users safety," *Safety science*, vol. 142, 105356, 2021.

[19] R. Nawaratne, D. Alahakoon, D. De Silva and X. Yu, "Spatiotemporal anomaly detection using deep learning for real-time video surveillance," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 393-402, 2019.

[20] M. Á. López, J.M. Lombardo, M. López, C.M. Alba, S. Velasco, M.A. Braojos and M. Fuentes-García, "Intelligent Detection and Recovery from Cyberattacks for Small and Medium-Sized Enterprises," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 3, 2020.

[21] K. Rezaee, S.M. Rezakhani, M.R. Khosravi and M.K. Moghimi, "A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance," *Personal and Ubiquitous Computing*, pp. 1-17, 2021.

[22] F. Rezaei and M. Yazdi, "A New Semantic and Statistical Distance-Based Anomaly Detection in Crowd Video Surveillance," *Wireless Communications and Mobile Computing*, vol. 2021, 5513582, 2021.

[23] W. Ullah, A. Ullah, I.U. Haq, K. Muhammad, M. Sajjad and S.W. Baik, "CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks," *Multimedia Tools and Applications*, vol. 80, no. 11, pp. 16979-16995, 2021.

[24] W. Ullah, A. Ullah, T. Hussain, Z.A. Khan and S.W. Baik, "An Efficient Anomaly Recognition Framework Using an Attention Residual LSTM in Surveillance Videos," Sensors, vol. 21, no. 8, 2811, 2021.

[25] Y. Luo, Y. Xiao, L. Cheng, G. Peng and D. Yao, "Deep learning-based anomaly detection in cyber-physical systems: Progress and opportunities," *ACM Computing Surveys (CSUR)*, vol. 54, no. 5, pp. 1-36, 2021.

[26] K.K. Santhosh, D.P. Dogra, P.P. Roy and A. Mitra, "Vehicular Trajectory Classification and Traffic Anomaly Detection in Videos Using a Hybrid CNN-VAE Architecture," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 11891-11902, 2021.

[27] H. Fanta, Z. Shao and L. Ma, "SiTGRU: single-tunnelled gated recurrent unit for abnormality detection," *Information Sciences*, vol. 524, pp. 15-32, 2020.

[28] W. Shin, S.J. Bu and S.B. Cho, "3D-convolutional neural network with generative adversarial network and autoencoder for robust anomaly detection in video surveillance," *International Journal of Neural Systems*, vol. 30, no. 6, 2050034, 2020.

[29] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6479-6488.

[30] A. Rummens, W. Hardyns and L. Pauwels, "The use of predictive analysis in spatiotemporal crime forecasting: Building and testing a model in an urban context," *Applied geography*, vol. 86, pp. 255-261, 2017.

[31] S. Kim, P. Joshi, P.S. Kalsi and P. Taheri, "Crime analysis through machine learning," in *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, IEEE, 2018, pp. 415-420.

[32] V. Tsakanikas and T. Dagiuklas, "Video surveillance systems-current status and future trends," *Computers & Electrical Engineering*, vol. 70, pp. 736-753, 2018.

[33] S. Prithi, S. Aravindan, E. Anusuya and A.M. Kumar, "GUI based prediction of crime rate using machine learning approach," *International Journal of Computer Science and Mobile Computing*, vol. 9, no. 3, pp. 221-229, 2020.

[34] H.W. Kang and H.B. Kang, "Prediction of crime occurrence from multi-modal data using deep learning", *PLOS ONE*, vol. 12, no. 4, e0176244, 2017.

[35] S. Hossain, A. Abtahee, I. Kashem, M.M. Hoque and I.H. Sarker, "Crime prediction using spatio-temporal data," in *International Conference on Computing Science, Communication and Security*, Springer, Singapore, 2020, pp. 277-289.

[36] G.N. Obuandike, I. Audu and A. John, "Analytical study of some selected classification algorithms in WEKA using real crime data," *International Journal of Advanced Research in Artificial Intelligence*, vol. 4, no. 12, 2015.

[37] C. C. Sun, C. Yao, X. Li and K. Lee, "Detecting Crime Types Using Classification Algorithms," *Journal of Digital Information Management*, vol. 12, no. 5, pp. 321-327, 2014.

[38] M. Jangra and S. Kalsi, "Crime analysis for multistate network using naive Bayes classifier," *International Journal of Computer Science and Mobile Computing*, vol. 8, no. 6, pp. 134-143, 2019.

[39] F. Vanhoenshoven, G. Nápoles, S. Bielen and K. Vanhoof, "Fuzzy cognitive maps employing ARIMA components for time series forecasting," in *International Conference on Intelligent Decision Technologies*, Springer, Cham, 2017, pp 255-264.

[40] W. Gorr, A. Olligschlaeger and Y. Thompson, "Assessment of crime forecasting accuracy for deployment of police," *International journal of forecasting*, 743-754, 2000.

[41] C.H. Yu, M.W. Ward, M. Morabito and W. Ding, "Crime forecasting using data mining techniques," in *2011 IEEE 11th international conference on data mining workshops*, IEEE, 2011, pp. 779-786.

[42] L.G. Alves, H.V. Ribeiro and F.A. Rodrigues, "Crime prediction through urban metrics and statistical learning," *Physica A: Statistical Mechanics and its Applications*, vol. *505*, pp. 435-443, 2018.

[43] M.Q. Gandapur, "E2E-VSDL: End-to-end video surveillance-based deep learning model to detect and prevent criminal activities," *Image and Vision Computing*, vol.123, 104467, 2022.

[44] M. Adimoolam, S. Mohan, A. John and G. Srivastava, "A Novel Technique to Detect and Track Multiple Objects in Dynamic Video Surveillance Systems," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 4, 2022.

[45] P. Christiansen, L.N. Nielsen, K.A. Steen, R.N. Jørgensen and H. Karstoft, "DeepAnomaly: Combining background subtraction and deep learning for detecting obstacles and anomalies in an agricultural field," *Sensors*, vol. 16, no. 11, 1904, 2016.

[46] L. Dong, Y. Zhang, C. Wen and H. Wu, "Camera anomaly detection based on morphological analysis and deep learning," in *2016 IEEE International Conference on Digital Signal Processing (DSP)*, IEEE, 2016, pp. 266-270.

[47] D. Xu, E. Ricci, Y. Yan, J. Song and N. Sebe, "Learning deep representations of appearance and motion for anomalous event detection," 2015, arXiv preprint arXiv:1510.01553.

[48] M. Hasan, J. Choi, J. Neumann, A.K. Roy-Chowdhury and L.S. Davis, "Learning temporal regularity in video sequences," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 733-742.

[49] V. Nguyen, D. Phung, D.S. Pham and S. Venkatesh, "Bayesian nonparametric approaches to abnormality detection in video surveillance," *Annals of Data Science*, vol. 2, no. 1, pp. 21-41, 2015.

[50] T. Ergen and S.S. Kozat, "Unsupervised anomaly detection with LSTM neural networks," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 8, pp. 3127-3141, 2019.

[51] J.X. Zhong, N. Li, W. Kong, S. Liu, T.H. Li and G. Li, "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1237-1246.

[52] S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory", in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, IEEE, 2017, pp. 1-6.

[53] K. Zhang, M. Sun, T.X. Han, X. Yuan, L. Guo and T. Liu, "Residual networks of residual networks: Multilevel residual networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 6, pp. 1303-1314, 2017.

[54] Y. Wu, Q. Wu, N. Dey and S. Sherratt, "Learning models for semantic classification of insufficient plantar pressure images," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 1, pp. 51-61, 2020.

[55] M.G. Huddar, S.S. Sannakki and V.S. Rajpurohit, "Attention-based Multi-modal Sentiment Analysis and Emotion Detection in Conversation using RNN," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, pp. 112-121, 2021.

[56] S.M. Erfani, S. Rajasegarar, S. Karunasekera and C. Leckie, "High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning," *Pattern Recognition*, vol. 58, pp. 121-134, 2016.

[57] C. Lu, J. Shi and J. Jia, "Abnormal event detection at 150 fps in matlab," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2720-2727.

[58] S. Vosta and K.C. Yow, "A CNN-RNN Combined Structure for Real-World Violence Detection in Surveillance Cameras," *Applied Sciences*, vol. 12, no. 3, 1021, 2022.

[59] M.U.K. Khan, H.S. Park and C.M. Kyung, "Rejecting motion outliers for efficient crowd anomaly detection," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 2, pp. 541-556, 2018.

[60] Hosseinzadeh, M., Rahmani, A. M., Vo, B., Bidaki, M., Masdari, M., & Zangakani, M. "Improving security using SVM-based anomaly detection: issues and challenges," *Soft Computing*, vol. *25*, 3195-3223.

[61] A.A. Alvarez and F. Gómez, "Motivic Pattern Classification of Music Audio Signals Combining Residual and LSTM Networks," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, 2021.

[62] N. Saleem, J. Gao, M. Irfan, E. Verdu and J.P. Fuente, "E2E-V2SResNet: Deep residual convolutional neural networks for end-to-end video driven speech synthesis," *Image and Vision Computing*, vol. 119, 104389, 2022.

Maryam Qasim Gandapur

Maryam Qasim received her LLB and LLM degrees from Khyber Law College, University of Peshawar in 2013 and 2015, respectively. She is preparing for PhD studies. She is currently an Assistant Professor at Department of Law, Shaheed Benazir Bhutto University, Dir, Khyber Pakhtunkhwa, Pakistan. She has been attached with academia for many years. Her research has focused on Corporate Law, Comparative Human Rights Law, intelligent systems for Law enforcement, and deep learning systems for Security.

Elena Verdú

Elena Verdú received her master's and Ph.D. degrees in telecommunications engineering from the University of Valladolid, Spain, in 1999 and 2010, respectively. She is currently an Associate Professor at Universidad Internacional de La Rioja (UNIR) and member of the Research Group "Data Driven Science" of UNIR. For more than 15 years, she has worked on research projects at both national and European levels. Her research has focused on e learning technologies, intelligent tutoring systems, competitive learning systems, accessibility, data mining and expert systems.