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

Maryam Qasim Gandapur¹*, Elena Verdú²

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 | Early Access 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.

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

Keywords

Anomaly Activities, Crime Detection, ConvGRU, Convolutional Neural Network (CNN), Deep Learning, Video Surveillance.
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>
<thead>
<tr>
<th>S. No.</th>
<th>Reference</th>
<th>Model</th>
<th>Achieved Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[36]</td>
<td>Decision Tree</td>
<td>59.15%</td>
</tr>
<tr>
<td>2</td>
<td>[37]</td>
<td>KNN</td>
<td>87.03%</td>
</tr>
<tr>
<td>3</td>
<td>[38]</td>
<td>Naive Bayes</td>
<td>87.00%</td>
</tr>
<tr>
<td>4</td>
<td>[39]</td>
<td>ARIMA</td>
<td>86.00%</td>
</tr>
<tr>
<td>5</td>
<td>[40]</td>
<td>Regression Model</td>
<td>72.00%</td>
</tr>
<tr>
<td>6</td>
<td>[41]</td>
<td>SVM</td>
<td>84.30%</td>
</tr>
<tr>
<td>7</td>
<td>[42]</td>
<td>Random Forrest</td>
<td>97.00%</td>
</tr>
<tr>
<td>8</td>
<td>[43]</td>
<td>E2E-VSDL</td>
<td>98.15%</td>
</tr>
</tbody>
</table>

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...
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 \((n\times1\times1\times2048)\) 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 \((n\times3\times3\times128)\) and is fed to ConvGRU layer. Since ResNet is not using for classification, the fully connected dense layer is not utilized.
respectively, whereas xt became:

\[
\text{Multiplication with convolution operations (denoted as \(\ast\)}, \quad \text{Eq. (1) - (4)}
\]

The input to the video frames, after passing ResNet-50, feeds the ConvGRU cell which reduces the computational complexity of the model. After Hadamard product, the complexity increases due to large parameters size thereby the model is prone to overfitting. However, convolution operations, the GRU can preserve spatial information. The ConvGRU 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:

\[
\begin{align*}
\text{ConvGRU} & \text{ is a 4-d tensor, } (n \times 256 \times 3 \times 3) \text{ such that input at each time-step is } (3 \times 3) \text{ with 256 channels. The output of ConvGRU is maxpooled with } (2 \times 2) \text{ size and flattened to get a 1-d vector. The 1-d vector feed the fully-connected layers followed by batch normalization (BN) and ReLU activation. For binary classification (normal vs abnormal activity), sigmoid activation and binary cross entropy loss can be used after fully-connected layers. However, softmax with categorical cross entropy loss can be used for multi-class classification. The working flow of the proposed ConvGRU-CNN is given in Fig. 1.}
\end{align*}
\]

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 input state \([55]\). By replacing the matrix multiplication with convolution operations (denoted as \(\ast\)), Eq. (1) - (4) became:

\[
\begin{align*}
\text{ConvGRU} & \text{ is a 4-d tensor, } (n \times 256 \times 3 \times 3) \text{ such that input at each time-step is } (3 \times 3) \text{ with 256 channels. The output of ConvGRU is maxpooled with } (2 \times 2) \text{ size and flattened to get a 1-d vector. The 1-d vector feed the fully-connected layers followed by batch normalization (BN) and ReLU activation. For binary classification (normal vs abnormal activity), sigmoid activation and binary cross entropy loss can be used after fully-connected layers. However, softmax with categorical cross entropy loss can be used for multi-class classification. The working flow of the proposed ConvGRU-CNN is given in Fig. 1.}
\end{align*}
\]

<table>
<thead>
<tr>
<th>Layer Name</th>
<th>Output Size</th>
<th>50-Layers Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv1</td>
<td>(112×112)</td>
<td>(7×7), 64, Stride 2</td>
</tr>
<tr>
<td>Conv2</td>
<td>(56×56)</td>
<td>[(1×1, 64), (3×3, 64), (1×1, 256)]+3</td>
</tr>
<tr>
<td>Conv3</td>
<td>(28×28)</td>
<td>[(1×1, 128), (3×3, 128), (1×1, 512)]×4</td>
</tr>
<tr>
<td>Conv4</td>
<td>(14×14)</td>
<td>[(1×1, 256), (3×3, 256), (1×1, 1024)]×6</td>
</tr>
<tr>
<td>Conv5</td>
<td>(7×7)</td>
<td>[(1×1, 512), (3×3, 512), (1×1, 2048)]×3</td>
</tr>
<tr>
<td></td>
<td>(1×1)</td>
<td>Average Pool, 1000d Fully Conn., Softmax</td>
</tr>
<tr>
<td>FLOPS</td>
<td>3.8 × 10^9</td>
<td></td>
</tr>
</tbody>
</table>

TABLE II. ResNet-50 Architecture

Fig. 3. ResNet-50 Structure.

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 step 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,
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

<table>
<thead>
<tr>
<th>Hyper Parameters</th>
<th>Tuning</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weights Initialization</td>
<td>Glorot-Uniform (Xavier)</td>
<td>81.9%</td>
</tr>
<tr>
<td>Weights Initialization</td>
<td>Random-Uniform</td>
<td>80.2%</td>
</tr>
<tr>
<td>Weights Initialization</td>
<td>He-Uniform</td>
<td>80.2%</td>
</tr>
<tr>
<td>Optimizer</td>
<td>Adam</td>
<td>80.9%</td>
</tr>
<tr>
<td>Optimizer</td>
<td>RMSprop</td>
<td>81.3%</td>
</tr>
</tbody>
</table>

### TABLE V. Video Frames Analysis (in Example Anomalies)

<table>
<thead>
<tr>
<th>Anomaly</th>
<th>Total Video Frames</th>
<th>Focused Video Frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assault</td>
<td>130395</td>
<td>55023</td>
</tr>
<tr>
<td>Fighting</td>
<td>269255</td>
<td>132517</td>
</tr>
<tr>
<td>Shooting</td>
<td>157735</td>
<td>55427</td>
</tr>
<tr>
<td>Vandalism</td>
<td>157511</td>
<td>82163</td>
</tr>
<tr>
<td>Total</td>
<td>714896</td>
<td>156610</td>
</tr>
</tbody>
</table>

### TABLE VI. CNN+ConvGRU Comparison

<table>
<thead>
<tr>
<th>Hyper Parameters</th>
<th>Weights Initialization</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-50+ConvGRU</td>
<td>Glorot-Uniform (Xavier)</td>
<td>82.6%</td>
</tr>
<tr>
<td>ResNet-101+ConvGRU</td>
<td>RMSprop</td>
<td>86.3%</td>
</tr>
<tr>
<td>ResNet-152+ConvGRU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>InceptionV3+ConvGRU</td>
<td>Adam</td>
<td>82.5%</td>
</tr>
<tr>
<td>VGG19+ConvGRU</td>
<td>RMSprop</td>
<td>89.3%</td>
</tr>
</tbody>
</table>

The evaluation measures are given by equations as:

\[
\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}
\]  
(9)

\[
\text{Sensitivity (Recall)} = \frac{\text{TP}}{\text{TP} + \text{FN}}
\]  
(10)

\[
\text{F1 Score} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}}
\]  
(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

<table>
<thead>
<tr>
<th>Database</th>
<th>Accuracy (%)</th>
<th>Precision (%)</th>
<th>F1 (%)</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCFcrimes</td>
<td>82.22%</td>
<td>83.13%</td>
<td>82.22%</td>
<td>82.65%</td>
</tr>
<tr>
<td>Binary</td>
<td>83.54%</td>
<td>85.65%</td>
<td>83.55%</td>
<td>82.77%</td>
</tr>
<tr>
<td>Average</td>
<td>82.88%</td>
<td>82.89%</td>
<td>82.88%</td>
<td>82.71%</td>
</tr>
</tbody>
</table>

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.
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>
<thead>
<tr>
<th>Database</th>
<th>Accuracy</th>
<th>AUC</th>
<th>Computational Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConvLSTM</td>
<td>81.93%</td>
<td>81.98%</td>
<td>25% Less computational complexity with ConvGRU</td>
</tr>
<tr>
<td>ConvGRU</td>
<td>82.88%</td>
<td>82.71%</td>
<td></td>
</tr>
</tbody>
</table>

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 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 outscore 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


