International Journal of Safety and Security Engineering
Vol. 14, No. 3, June, 2024, pp. 953-966
Journal homepage: http://iieta.org/journals/ijsse

VidAnomalyNet: An Efficient Anomaly Detection in Public Surveillance Videos Through Deep Learning Architectures

K. Chidananda[1*], A.P. Siva Kumar[2]

[1] Department of Computer Science and Engineering, Jawaharlal Nehru Technological University Anantapur (JNTUA), Ananthapuramu 515002, India
[2] Department of Computer Science and Engineering, JNTUA College of Engineering Anantapur (JNTUACEA), Ananthapuramu 515002, India

Corresponding Author Email: chida.koudike@gmail.com

https://doi.org/10.18280/ijsse.140326

Received: 24 November 2023
Revised: 2 February 2024
Accepted: 13 March 2024
Available online: 24 June 2024

Keywords:
convolutional neural network, video anomaly detection, MobileNet, transfer learning, video surveillance

ABSTRACT

In the contemporary era, computer vision applications assume significance due to their role in the real world. Video surveillance is one such application that has become indispensable with plenty of unprecedented applications. Detection of abnormal events from surveillance videos in real time has its importance in applications like traffic monitoring, crime investigation, public safety, healthcare and operations management to mention few. With the emergence of Artificial Intelligence (AI) automatic video surveillance is taken to the next level with sophistication in learning detection of anomalies. Particularly deep learning model like Convolutional Neural Network (CNN) is found more appropriate for image processing. However, as one size does not fit all, CNN does not provide acceptable accuracy unless it is enhanced with suitable number of layers and configurations. Towards this end, in this paper, we proposed a novel deep learning architecture known as VidAnomalyNet which is based on CNN model. It is designed to have more appropriate learning process and detection of anomalies from surveillance videos. We proposed a framework to exploit our VidAnomalyNet architecture for leveraging detection performance. We also proposed an algorithm known as VidAnomalyNet for Automatic Anomaly Detection (VAAD). Automatic anomaly detection in the context of video anomaly networks refers to the use of computational methods to automatically identify unusual or abnormal patterns within a sequence of video frames. The goal is to develop models that can distinguish between normal activities and unexpected events or anomalies. Video anomaly detection is crucial in various applications, including surveillance, industrial monitoring, and public safety. At present, this algorithm detects three classes of anomalies like fire, accident and robbery. It can be easily extended to identify more number of anomalies. We also explored MobileNetV1 with transfer learning by adding new layers to the base model for video anomaly detection. Our empirical study has revealed that VidAnomalyNet outperforms MobileNetV1. Highest accuracy achieved by the proposed model is 96.35%.

1. INTRODUCTION

Video surveillance and automatic detection of anomalies has become an important research area. It is indispensable in the modern applications in urban and industrial environments involving in development, day to day operations and sustainability. This kind of research contributes towards safety of citizens, improved security, efficiency and real time approach in monitoring and making well-informed decisions. Not only in industrial environments and cyber-physical systems, video surveillance plays crucial role in areas of high human population density. As urban areas are rapidly increasing in population diversity and density living in multi storeyed buildings with increased pedestrian, crowed and vehicular movements, video surveillance has its pivotal role in facilitating human safety, security, law and order besides bestowing evidence towards speedy investigations made by law-enforcing agencies [1].

The motivation for anomaly detection in videos stems from the need to enhance security, safety, and monitoring processes in various domains. The problem statement revolves around developing models that can automatically identify deviations from normal behavior in video streams, with the ultimate aim of preventing or responding to unexpected events effectively.

There is continuous horizontal and vertical expansion of industrial and urban areas leading to exponential usage of CCTV cameras. When there are several thousands of such cameras in operation, it is not desirable to have manual observation of such video streams. It is also not ideal to monitor video footage only when certain untoward incidents occur. There should be technology-driven approach that takes this into an autonomous video surveillance process which monitors and analyses videos in real time. Towards this end, Artificial Intelligence (AI) technology has wherewithal to

support automatic analysis of surveillance videos in real time and provide its findings as the incidents occur. As explored in the previous studies [2-4], to mention few, deep learning is an AI based learning approach that has capacity to serve the purpose of autonomous and comprehensive video surveillance. Especially detection of anomalous behaviours or incidents has to be given paramount importance. Towards this end there are many existing contributions found in the literature.

Video Anomaly Networks (VANs) are designed to effectively capture spatiotemporal patterns in video sequences, making them more suitable for anomaly detection in dynamic environments compared to traditional methods or architectures. Specific features that contribute to the efficiency of VANs for learning and detection include.

Spatiotemporal Convolutional Layers: VANs often incorporate spatiotemporal convolutional layers, allowing them to simultaneously process spatial and temporal information in video frames. This is crucial for capturing the dynamic nature of video sequences.

Temporal Modeling: Incorporation of recurrent layers, such as Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) units, enables VANs to model temporal dependencies and long-term patterns in video data. This is essential for understanding the context and continuity of events over time.

Unsupervised Learning Techniques: VANs often operate in an unsupervised learning setting, where they are trained on normal video data without explicit labels for anomalies. Unsupervised learning allows the model to learn the inherent patterns of normal behavior and detect anomalies based on deviations from these patterns.

Autoencoder Architectures: Autoencoders are commonly used in VANs for unsupervised learning. These architectures learn to reconstruct normal video frames and identify anomalies by measuring the reconstruction error. Autoencoders are effective in capturing and representing relevant features of the input data.

Adaptability to Different Anomalies: VANs are designed to be adaptable to different types of anomalies. The learned representations are expected to be generic enough to capture a wide range of abnormal behaviors, making the model versatile in various applications.

Attention Mechanisms: Some VANs may incorporate attention mechanisms to focus on specific regions or frames of interest within the video sequence. This helps the model prioritize relevant information for anomaly detection, improving efficiency and accuracy.

Real-Time Processing: Efficient VAN architectures are designed to handle real-time processing of video streams, making them suitable for applications where timely anomaly detection is crucial, such as surveillance and security.

Evaluation Metrics: VANs are typically evaluated using metrics such as precision, recall, F1 score, and area under the receiver operating characteristic (ROC) curve. These metrics provide a quantitative measure of the model's efficiency in correctly identifying anomalies while minimizing false positives.

The previous studies [5-10] explored convolutional autoencoder along with optimal flow method to detect anomalies from videos. Autoencoders are also used for this kind of research [11-15], different autoencoder models are investigated to ascertain their merits in detection of video abnormalities. In there is spatio-temporal approach in video analytics by exploiting autoencoders considering both spatial and temporal dimensions. In anomaly detection is explored using variational autoencoder and Gaussian mixture on top of convolutional approach. Their method also focused on localization of anomalies. In autoencoding and attention mechanisms are combined to have spatio-temporal autoencoding process. In autoencoders are used in critical infrastructure monitoring and analysis. Research found that autoencoders are good for anomaly detection. However, they focus on learning much information and sometimes relevant information learning becomes an issue. There is need for training autoencoders with lot of data with hyperparameter tuning. There are many CNN variants [16-19], to mention few, used for anomaly detection from videos. It is found that CNN models are suitable for image data analysis. Moreover, their learning capability makes then preferred advanced neural network architectures to solve real world problems. However, "one size does not fit all" as the CNN models cannot directly provide optimal performance in every case considered. Our contributions in this paper are as follows.

(1) We proposed a novel deep learning architecture known as VidAnomalyNet. This model is based CNN model as it is found to be highly successful in processing image data. Moreover, CNN is found efficient in feature map generation and optimization. With our architecture, there is more efficient learning process and detection of anomalies from surveillance videos.

(2) We proposed a framework that makes use of our VidAnomalyNet architecture to enhance performance in anomaly detection from surveillance videos. The framework provides a set of reusable components for facilitating intended functionality.

(3) We also proposed an algorithm known as VidAnomalyNet for Automatic Anomaly Detection (VAAD). This algorithm is designed on top of the proposed deep learning architecture. It has provision for multi-class classification with ability to detect four classes such as normal, fire, accident and robbery. It can be easily extended to identify more number of anomalies.

(4) We also explored MobileNetV1 with transfer learning by adding new layers to the base model for video anomaly detection. The rationale behind this is that MobileNet is good for processing imagery data. We compared the performance dynamics of MobileNet based enhanced architecture and our VidAnomalyNet. Our empirical study has revealed that VidAnomalyNet outperforms MobileNetV1 with highest accuracy 96.35%.

The remainder of the paper is organized into several sections. The related work in Section 2 provides valuable literature insights that helps us to ascertain research gaps and the need for building a more appropriate deep learning architecture for automatic detection of anomalies from surveillance videos. Literature also throws light on merits and demerits of existing detection methods. Our proposed architecture, algorithm and underlying mechanisms including dataset details are provided in Section 3. It throws light on our VidAnomalyNet architecture which has layers configured based on the performance dynamics in the empirical study. It also provides the details of MobileNetV1 and how it is subjected to transfer learning process to improve the baseline model.

Ultimately, the choice of deep learning approach depends on the nature of the task, the characteristics of the data, and computational resources available. Hybrid models and ensembles are also common, combining the strengths of

different architectures for improved performance. It's important to consider the specific requirements and challenges of each application when selecting a deep learning approach.

2. RELATED WORKS

The Convolutional Long Short-Term Memory (Conv-LSTM) network, which explains real-time crowd AD. To anticipate violent acts and assist stakeholders in exhibiting such activities in real time, a Deep Learning (DL)-centric strategy was adopted. Conv-LSTM was used to both detect violent actions and capture the frame. The suggested system produced better accuracy at a quicker rate. However, due to the difficulty of classifying individual or group activities, accuracy was still inadequate [20].

They demonstrated an AD module and a human detection module that together made up a supervised Local Distinguish Ability improving Network (LDA-Net). An inhibitory loss function and embedding were devised to reduce misclassification in highly imbalanced datasets. The results of the simulation demonstrated how the constructed supervised LDA-Net produced the state-of-the-art results. However, the created model's considerable computational complexity resulted from the addition of a new axis [21].

The proposed an online AD technique for surveillance videos utilizing transfer learning as well as continual learning. The developed algorithm utilized the Feature Extraction (FE) power of neural network-centered methodologies and statistical detection methods. Simulation outcomes considerably gave pre-eminent accuracy for the built system. Nevertheless, it was still challenging to learn to detect abnormalities promptly. The established an approach centered on bidirectional prediction, and subsequently built the loss function utilizing the real target frame along with its bidirectional prediction frame. Moreover, with a focus on the prediction error map's foregrounds, an anomaly score estimation approach centered on the SW scheme was built. Better AD was delivered by the experimental outcomes with higher scores. However, it highly depends on assumption-centric data generation. Thus, the false alarm rate was high [22].

The outlined a convolutional neural network (CNN)-centric, lightweight solution to AD. In sequence learning, the generated residual attention-centric LSTM was trained with the retrieved spatial CNN features, and it demonstrated both recognition and AD efficiency. Extensive experiments were conducted to validate the efficiency of the developed paradigm. Modelling based on normal activity was impractical because normal was such a general term, and it would be challenging to classify everything that fell under it [23].

A data-driven adaptive AD method for human activities was presented by the. Behaviour modelling was obtained by using the Consensus Novelty Detection Ensemble, which is an ensemble for novelty detection systems and includes a One-Class SVM. The simulation results demonstrated the good performance of the designed system. The deep systems have a problem in that they require more data and a lot of computing power [24].

3. METHODS

This section presents details of our methods, dataset used for empirical study, our deep learning architecture, mechanisms, algorithm defined and evaluation procedure used.

3.1 Dataset

Public surveillance videos are collected from [25] for the empirical study in this paper. These videos do have many realistic anomalies and also normal instances. This dataset is widely used by many researchers, such as the study [26], in computer vision applications. It is known as UCF-Crime dataset which covers 13 classes of real world anomalies. Out of 13 classes we have extracted four classes such as fire, accident, robbery and normal. The fire class videos show the fire intentionally set to properties. The accident class consists of videos with traffic accidents where cyclists, pedestrians and vehicles are involved. The robbery class consists of videos reflecting thieves taking money unlawfully by threatening or force. But this class does not include shooting kind of threats. The normal class of videos contain indoor and outdoor scenes but do not reflect any occurrence of crime.

Figure 1 shows the data distribution dynamics for 4 classes. From the collected dataset 4 classes of data are extracted for the research carried out in this paper. Out of total number of surveillance videos (16853 n), we have taken 3123 normal instances, 1563 fire instances, 2654 accident instances and 2276 robbery instances. Figure 2 shows an excerpt from each class of the collected dataset.

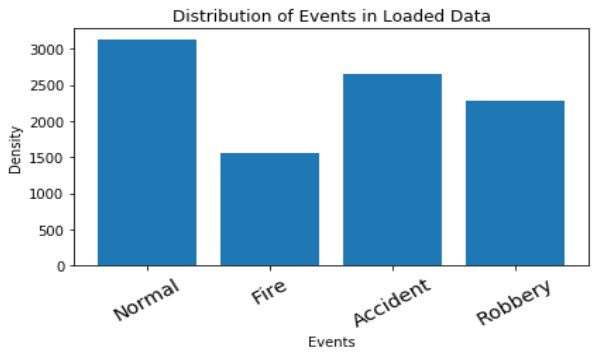
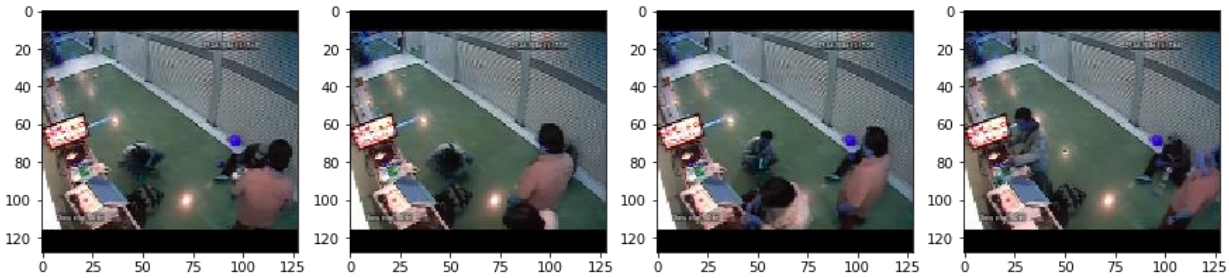
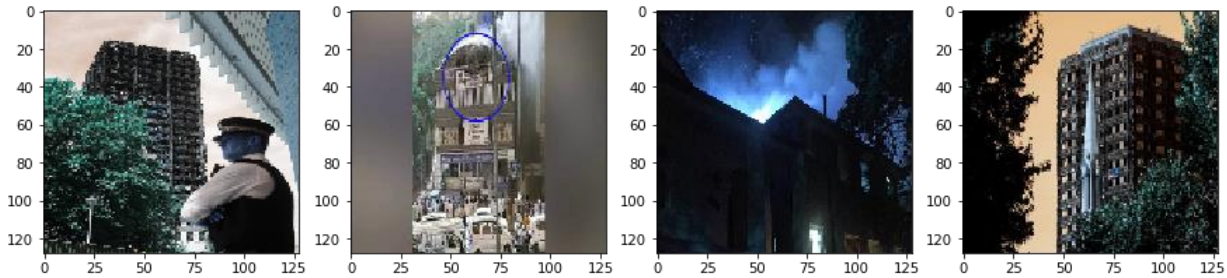


Figure 1. Data distribution dynamics reflecting 4 classes

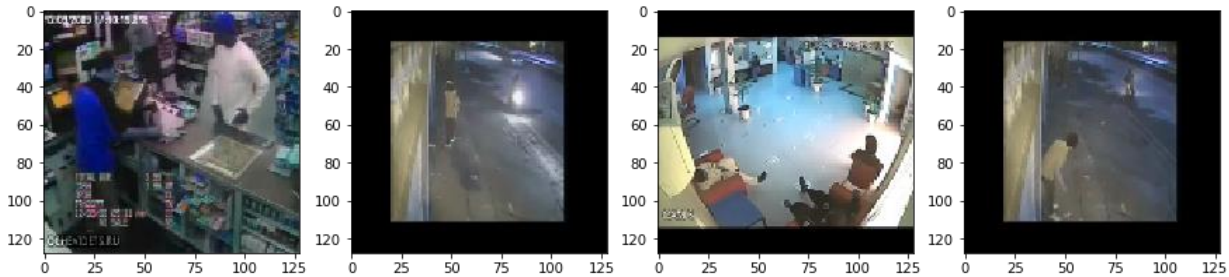


(a) Normal data



(b) Fire data

(c) Accident data



(d) Robbery data

Figure 2. An excerpt from dataset reflecting all four classes

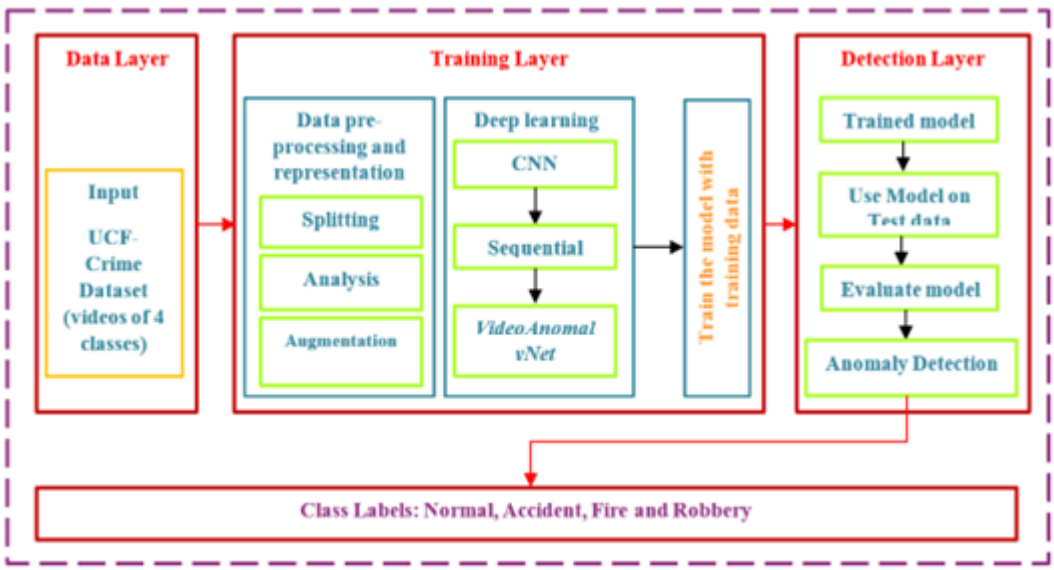


Figure 3. Overview of the proposed methodology for anomaly detection from surveillance videos

3.2 Our methodology

We proposed a methodology and a novel CNN based deep learning architecture known as VidAnomalyNet for efficient detection of anomalies from surveillance videos. Before going to technical details of the VidAnomalyNet, we delve into overall methodology which provides the modus operandi of our approach in some detail. Overview of the proposed methodology for anomaly detection from surveillance videos is illustrated in Figure 3.

The UCF-crime dataset is used to extract only 4 classes for our experiments. They are known as fire, accident, robbery

and normal. In the extracted dataset (now onwards we simply call it UFC crime dataset or dataset), there are 9616 instances covering all 4 classes. Each class related videos in the dataset is subjected to splitting into training set (75%) and testing (25%). Afterwards, the data is analysed to know whether it is balanced and need augmentation. Data augmentation is carried out in order to have more quality in the training process. Afterwards, the training data is given to our proposed novel deep learning architecture known as VidAnomalyNet (explained in Section 3.3). The deep learning model is trained with the training videos covering 4 classes. After completion of the training the model is persisted to secondary storage for further reuse instead of re-inventing the wheel every time when new surveillance video arrives for anomaly detection task. The saved knowledge model is then used to perform detection of anomalies by using test videos. Then the proposed deep learning model VidAnomalyNet is evaluated and compared against state-of-the-art model such as MobileNetV1 and MobileNetV1 with transfer learning implemented. The final outcome with regard to anomaly detection includes classification of test videos with four classes such as normal, accident, fire and robbery. Figure 4 shows the flow of the proposed methodology.

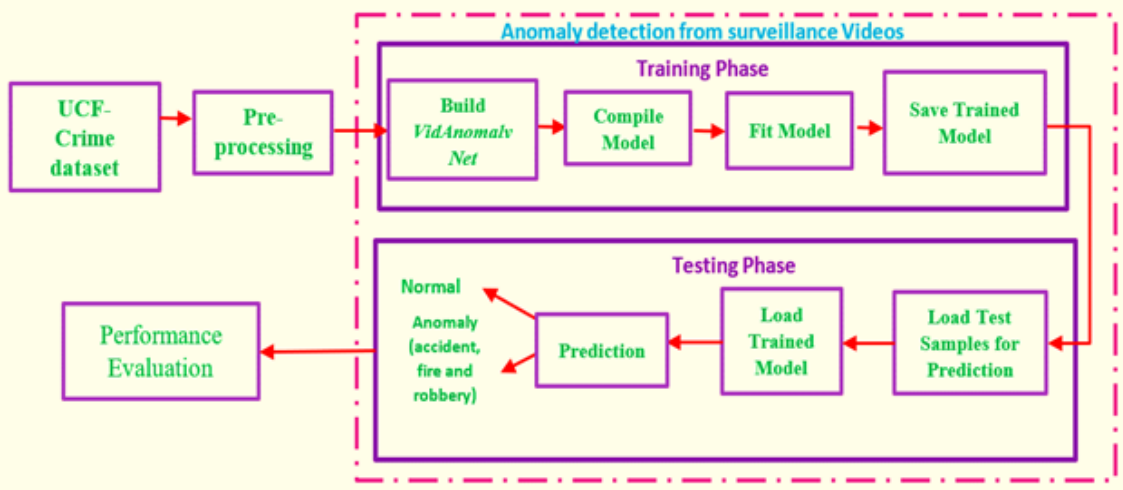


Figure 4. Flow of the proposed methodology

Layer (type)	Output Shape	Param #
separable_conv2d (SeparableC	(None, 128, 128, 16)	211
batch_normalization (BatchNo	(None, 128, 128, 16)	64
max_pooling2d (MaxPooling2D)	(None, 64, 64, 16)	0
separable_conv2d_1 (Separabl	(None, 64, 64, 32)	688
batch_normalization_1 (Batch	(None, 64, 64, 32)	128
max_pooling2d_1 (MaxPooling2	(None, 32, 32, 32)	0
separable_conv2d_2 (Separabl	(None, 32, 32, 64)	2400
batch_normalization_2 (Batch	(None, 32, 32, 64)	256
separable_conv2d_3 (Separabl	(None, 32, 32, 64)	4736
batch_normalization_3 (Batch	(None, 32, 32, 64)	256
max_pooling2d_2 (MaxPooling2	(None, 16, 16, 64)	0
flatten (Flatten)	(None, 16384)	0
dense (Dense)	(None, 128)	2097280
activation (Activation)	(None, 128)	0
batch_normalization_4 (Batch	(None, 128)	512
dropout (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 128)	16512
activation_1 (Activation)	(None, 128)	0
batch_normalization_5 (Batch	(None, 128)	512
dropout_1 (Dropout)	(None, 128)	0
dense_2 (Dense)	(None, 4)	516
activation_2 (Activation)	(None, 4)	0

Total params: 2,124,071
Trainable params: 2,123,207
Non-trainable params: 864

Figure 5. Layers and parameters involved in the proposed VidAnomalyNet architecture

Table 1. Parameters for different optimizers used for model training

References	Approach	Techniques	Dataset
[27]	Deep learning	DCNN	UCSD, CUHK, ShanghaiTech
[28]	Deep Learning	DTM technique	UCSD, Mall, UMN and MED
[29]	Deep Learning	Neural Network models	Custom dataset
[30]	Deep Learning	SFE technique	TUT 2016
[31]	Deep Learning and Bio-Inspired	CRN along with AntHocNet	Custom dataset
[32]	Deep Learning	IIN	UCF-Crime, UCSD
[33]	AI	DCNN	Custom dataset
[34]	Deep Learning	DCNN methods	Seven benchmark datasets
[35]	Deep Learning	DNN	UCF-Crime
[36]	Deep Learning	LMNN	Custom dataset

The dataset is subjected to pre-processing which results in 75% training set and 25% test set for each class. Then there are two phases involved in the system. Since it is supervised learning phenomenon, it has training and testing phases. In the former, our novel deep learning model named VidAnomalyNet is built with the architecture show in Figure 5. The model is compiled and trained with different optimizers and parameter as presented in Table 1. Afterwards the trained model is saved for reuse. The saved model is used in the testing phase of the system. The trained model is used to use test data to perform prediction of anomalies. The prediction results are four classes including normal, accident, fire and robbery.

3.3 Proposed VidAnomalyNet architecture

It is based on CNN model as CNN is found suitable for image analysis. VidAnomalyNet is made up of many kinds of layers. They include Convolutional 2D layers, batch normalization layers, max pooling 2D layers, flatten layer, dense layers, dropout and activation layers. Figure 5 shows the layers, their output shape and number of parameters.

VidAnomalyNet architecture is designed with our empirical study to maximize performance in anomaly detection from surveillance videos. Our model includes multiple layers. In the context of video anomaly detection, separable convolution 2D can be employed to efficiently process spatiotemporal information in video data. Traditional 2D convolutions involve applying a filter/kernel to the input data in both the spatial dimensions (width and height) simultaneously. Separable convolution, on the other hand, decomposes the standard 2D convolution into two consecutive operations: a spatial convolution along each spatial dimension separately (width and height), followed by a 1D temporal convolution along the time dimension and to a standard 2D convolution, leading to computational efficiency and results available in Table 2.

Table 2. Parameters for different optimizers used for model training

Optimizer	Learning Rate	Batch Size	Number of Epochs
SGD	le-2	64	100
Adam	le-0.001	64	100
RMSProp	le-0.001	64	100

It is followed by batch normalization and max pooling 2D with pool size (2,2). The softmax layer is made up of a dense layer followed by activation layer. The VidAnomalyNet is finally built with width 128, highest 128 and depth 3 besides number of classes 4.

The model is trained with three kinds of optimizers namely SGD, Adam and RMSProp. The loss function used is known as sparse categorical cross entropy.

In the proposed VidAnomalyNet architecture separable convolution 2D layer is preferred as it could reduce number of parameters without compromising performance. The rationale behind this is that it has provision factorization that results in reduction of parameters leading to reduction in model size and computation time. Separable convolutional 2D Eq. (3) exploits pointwise Eq. (1) and depth wise convolution Eq. (2) variants towards optimization.

$$\mathrm{pc} = \sum_{\mathrm{m}}^{\mathrm{M}} \mathrm{W_m}.\,\mathrm{y(i,j,m)} \tag{1}$$

$$\mathrm{dc(W,y)_{(i,j)}} = \sum_{\mathrm{k,l}}^{\mathrm{K,L}} \mathrm{W_{(k,l)}} \odot \mathrm{y(i+k, j+l)} \tag{2}$$

$$\mathrm{sc(W_p, W_d, y)_{(i,j)}} = \mathrm{pc_{(i,j)}(W_p, dc_{(i,j)}(W_d, y))} \tag{3}$$

Here pointwise approach is the regular convolutional method. Depth wise variant makes use of single kernel but results in more number of parameters. Max pooling 2D layers are used to ensure spatial invariance and optimize feature maps. Pool size is (2, 2) and as per that pooling window based processing takes place. It exploits subsampling towards optimizing feature maps as expressed in Eq. (4).

$$a_j = \tanh(\beta \sum_{N\,X\,N} a_i^{nxn} + b) \tag{4}$$

It considers averaging inputs and multiply them using a trainable scalar denoted as β. Then it adds trainable bias denoted as b and the result is passed via non-linearity. The max pooling function is as expressed in Eq. (5).

$$a_j = \max_{N\,XN}(a_i^{nxn} u(n,n)) \tag{5}$$

The map pooling layer exploits a window function denoted as u(x,y) for given input patch and it computes maximum of the neighbourhood. The outcome is in the form of optimized feature map.

Batch normalization is the technique of normalization that is taken care between layers in the proposed model VidAnomalyNet. Instead of using full data, multiple batches are used in order to make the learning process easier, adopt to learning rates and speed up the training phenomenon. The batch normalization is made as in Eq. (6).

$$z^N = \left(\frac{z - m_z}{S_z}\right) \quad (6)$$

The mean of output of the neurons is denoted as m_z while standard deviation of the output of neurons is denoted as S_z.

In the proposed VidAnomalyNet model dense layers are also used. Each neuron in the dense layer receives input from previous layer's all neurons. Thus it is called as dense layer. It has capability to classify given image based on the received output from convolutional layers.

3.4 Architecture of MobileNet

In our empirical study MobileNet and MobileNet with transfer learning are used as existing models used for comparing with the proposed VidAnomalyNet model. The original model of MobileNet V1 [37, 38] is used in our empirical study. Then it is further improved with transfer learning as discussed in Section 3.5. It is made up of lightweight CNN layers for computer vision applications. It makes use of depth wise separable convolutional filters where single convolution is performed on every input channel.

Then there is pointwise convolution filter that combines the result of depth wise convolution in a linear fashion considering 1x1 convolutions as illustrated in Figure 6, Figure 7 shows the architectural layers of MobileNet V1.

As presented in Figure 7, the MobileNet architecture includes depth wise convolutional layers, batch normalization and ReLU activation provide in number of layers. We trained this model with the proposed methodology and results are obtained for detection of abnormal activities from surveillance videos. Then we also experimented with MobileNetV1 with transfer learning to improve the training process as discussed in Section 3.5.

3.5 MobileNet based transfer learning architecture

Transfer learning (TL) is one of the techniques of ML which is used to preserve knowledge gained while solving a problem and reuse it layer for som other related problem. In other words, TL is used to reuse knowledge and speed up the process of training and detection. We proposed a transfer learning architecture by reusing MobileNetV1. As presented in Figure 7, some additional layers are added to exploit the existing architecture and improve its performance in video anomaly detection.

Figure 8, the additional layers added are global average pooling layers, two dropout layers and two dense layers.

By combining mobile networks with transfer learning, you can benefit from the efficiency of lightweight architectures while still achieving good performance on video anomaly detection tasks, even with limited labeled data for the specific application domain. This makes the approach particularly appealing for deploying anomaly detection models on resource-constrained devices and in scenarios where obtaining large amounts of labeled data is challenging.

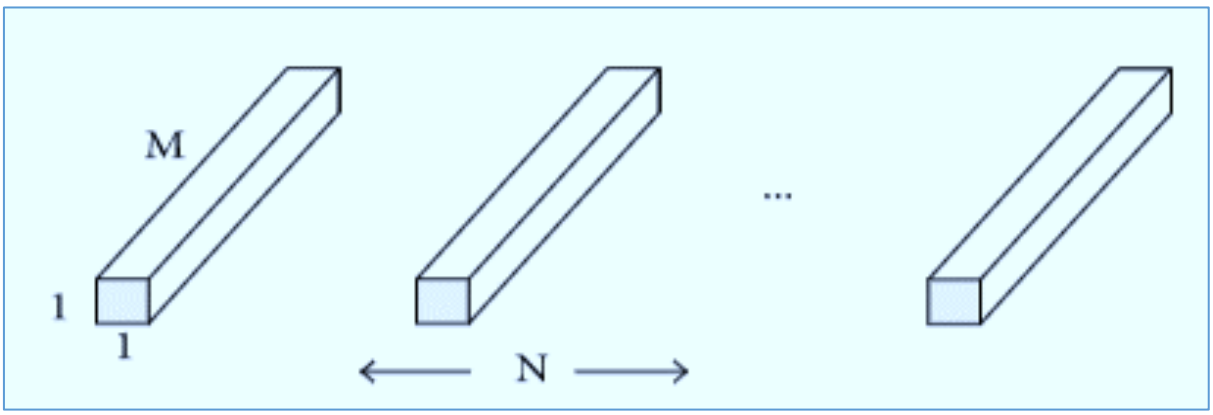


Figure 6. Depth wise convolutional filters used in MobileNet

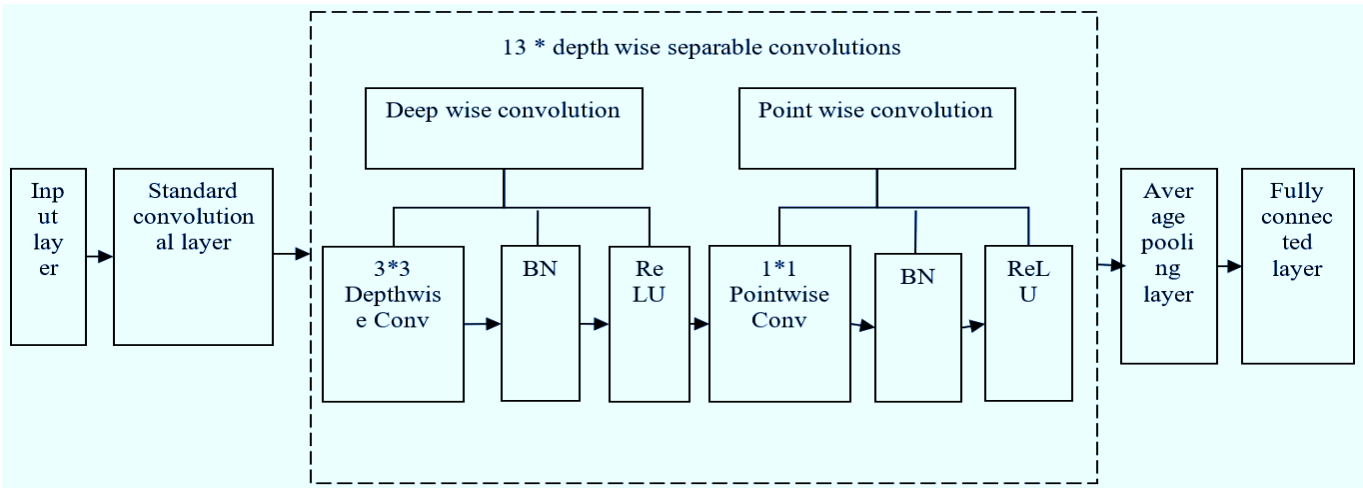


Figure 7. Architectural overview of MobileNet V1

Layer (type)	Output Shape	Param #
mobilenet_1.00_128 (Model)	(None, 4, 4, 1024)	3228864
global_average_pooling2d (Gl	(None, 1024)	0
dropout_6 (Dropout)	(None, 1024)	0
dense_9 (Dense)	(None, 128)	131200
dropout_7 (Dropout)	(None, 128)	0
dense_10 (Dense)	(None, 4)	516

Total params: 3,360,580
Trainable params: 3,338,692
Non-trainable params: 21,888

Figure 8. Architectural layers of MobileNetV1 with transfer learning

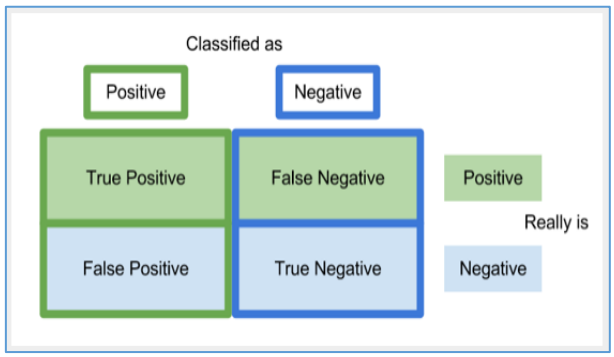


Figure 9. Confusion matrix

3.6 Performance evaluation method

Precision is one such metric expressed in Eq. (7). It is the ratio between correctly classified anomalous instances and both correctly and incorrectly classified anomalous instances.

$$\text{Precision (p)}=\frac{True\ Positve}{True\ Positive+False\ Positive} \quad (7)$$

Recall is another metric which is the ratio between correctly classified anomalous instances and both correctly classified anomalous instances and incorrectly classified anomalous instances. This measure is expressed as in Eq. (8).

$$\text{Recall (r)}=\frac{True\ Positive}{True\ Positive\ +False\ Negative} \quad (8)$$

F1-score is the measure which is the harmonic mean of both precision and recall. This measure is expressed as in Eq. (9).

$$\text{F1-score}=2*\frac{(precision*recall)}{(precision+recall)} \quad (9)$$

Accuracy is yet another widely used metric for performance evaluation. This metric is as expressed in Eq. (10).

$$\text{Accuracy}=\frac{True\ Positive\ +True\ Negative}{True\ Positive+True\ Negative+False\ Positive\ +False\ Negative} \quad (10)$$

All these metrics result in a value between 0.0 and 1.0 reflecting least and highest performance respectively.

4. RESULTS AND DISCUSSIONS

This section presents experimental results of the proposed model known as VidAnomalyNet along with existing model known as MobileNetV1 and transfer learning variant of MobileNetV1. In all experiments the surveillance videos dataset, for each class, is divided into 75% for training and 25% for testing. Each model is evaluated with three different optimizers such as Adam, SGD and RMSProp. Batch size used is 64 and number of epochs is 100. Learning rate with SGD optimizer is le-2 and for Adam and RMSProp optimizers it is set to le-0.001. Implementation of the models is made using Python 3.9 and Jupyter IDE.

4.1 Data analysis and augmentation

All the samples used for empirical study are clubbed and they are analysed to know whether there is need for data augmentation.

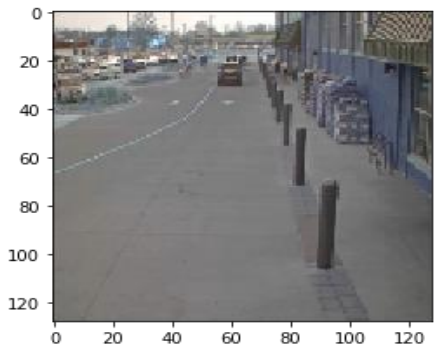


Figure 10. A sample picked for analysis

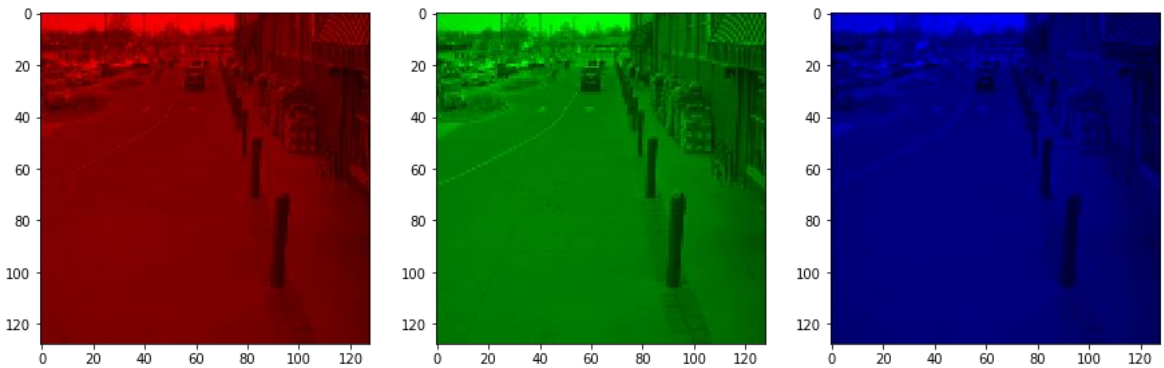


Figure 11. Different channels extracted from Figure 10

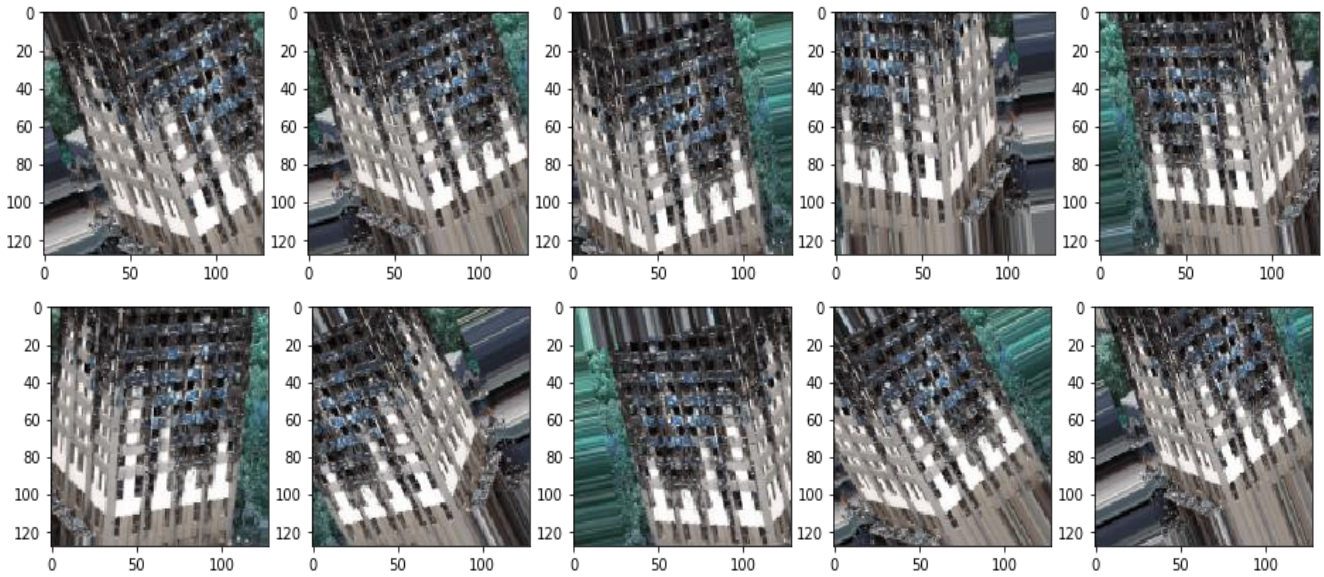


Figure 12. Visualization of data augmentation

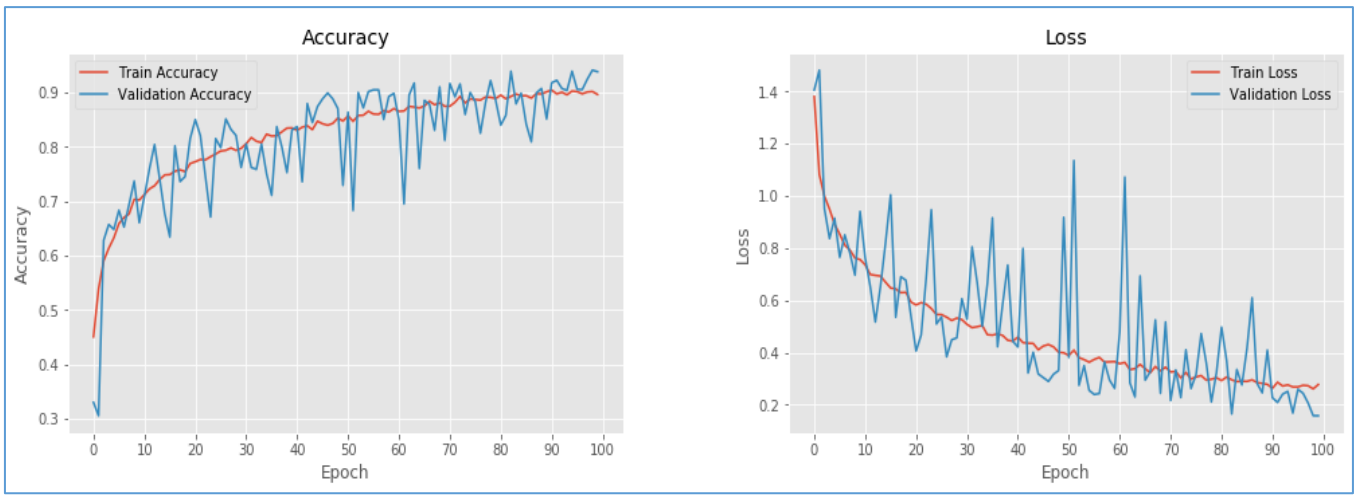


Figure 13. Training and validation accuracy and loss of VidAnomalyNet model with SGD optimizer

As presented in Figure 10, Figure 11 and Figure 12, there is data analysis and data augmentation used in order to improve training quality prior to exploiting deep learning models for anomaly detection from surveillance videos.

4.2 Performance evaluation

The performance of proposed model VidAnomalyNet is provided in this section along with comparing the same with MobileNetV1 and MobileNetV1 with transfer learning.

Both accuracy and loss metrics are used for visualizing performance of the proposed model. Higher accuracy denotes better performance while lower in loss denotes better performance. As presented in Figure 13 training accuracy, validation accuracy, training loss and validation loss are provided against number of epochs.

As presented in Figure 14, confusion matrix shows the prediction details of the proposed model in terms of TP, FP, TN and FN. It shows ground truth and predicted labels for all four classes. In confusion matrix 0 indicates normal class, 1 indicates fire, 2 indicates accident and 3 denotes robbery.

As presented in Figure 15 and Figure 16, performance and related confusion matrix are provided for the proposed model with Adam optimizer.

As presented in Figure 17 and Figure 18, performance and related confusion matrix are provided for the proposed model with RMSProp optimizer.

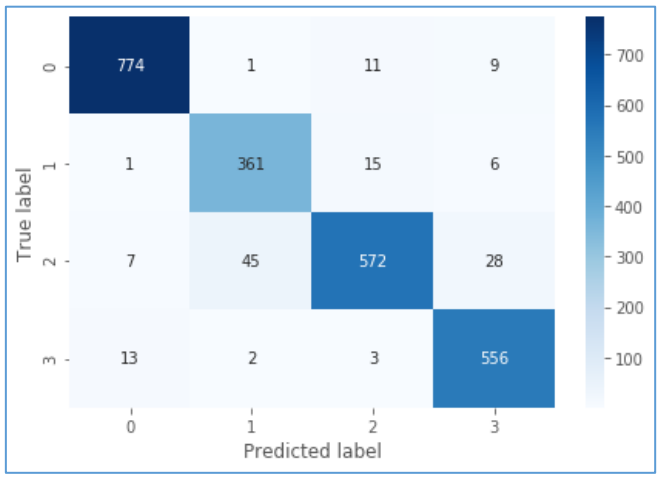


Figure 14. Confusion matrix reflecting predictions of VidAnomalyNet model with SGD optimizer

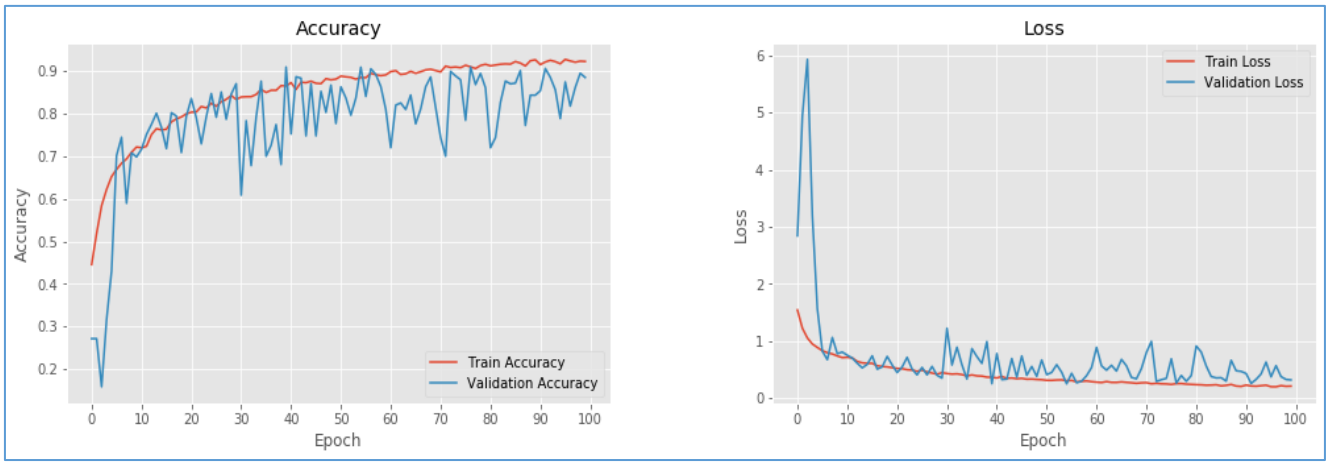


Figure 15. Training and validation accuracy and loss of VidAnomalyNet model with Adam optimizer

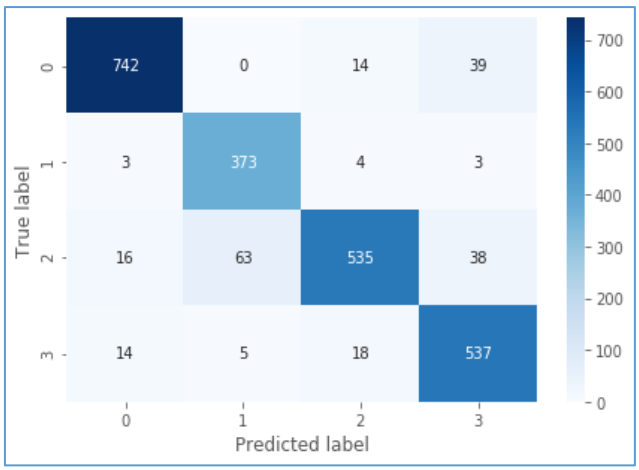


Figure 16. Confusion matrix reflecting predictions of VidAnomalyNet model with Adam optimizer

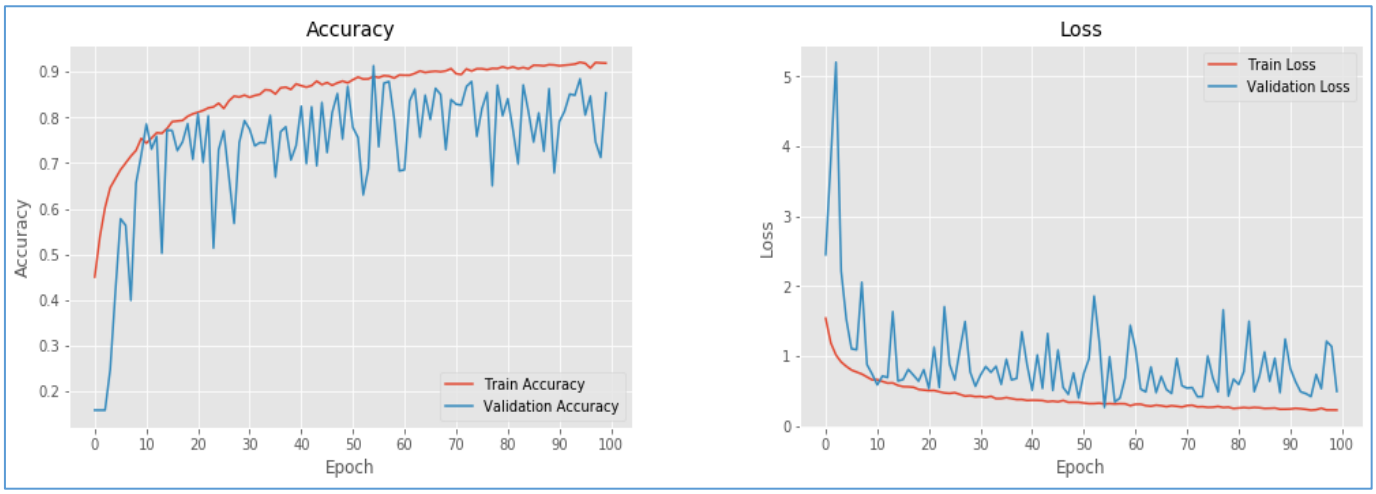


Figure 17. Training and validation accuracy and loss of VidAnomalyNet model with RMSProp optimizer

As presented in Figure 19, it is observed that deep learning based proposed model VidAnomalyNet takes more time at the time of training. However, the trained model can be used for detection of anomalies in surveillance videos in real time with negligible time delay. With SGD optimizer it took 3071710 milliseconds or 3071 seconds which amounts to around 51 minutes. The model with Adam optimizer needed 3051855 milliseconds and with RMSProp 3058404 milliseconds.

As presented in Figure 20, it is one of the test images associated to surveillance videos of Accident class. It does mean that ground truth is Accident class. The proposed model VidAnomalyNet also predicted it as Accident class. The learned model took 23 milliseconds to detect it.

Selecting the right assessment measures is crucial. Metrics like precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) can be used to compare different models. The objectives and specifications of the anomaly detection task determine which metrics are used.

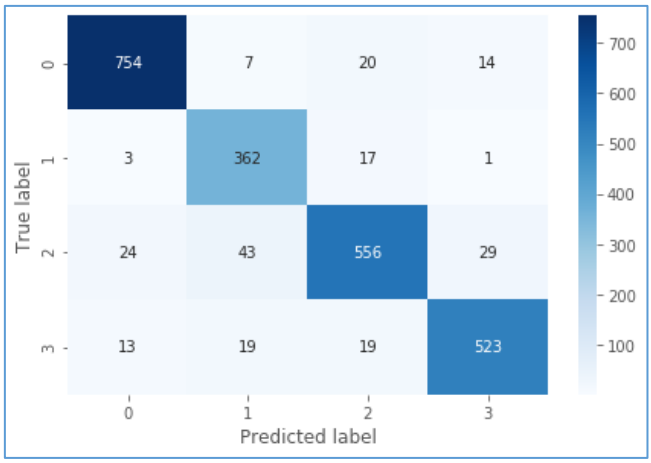


Figure 18. Confusion matrix reflecting predictions of VidAnomalyNet model with RMSProp optimizer

Figure 19. Time taken to train VidAnomalyNet with different optimizers

Figure 20. One of the correctly predicted sample as accident

4.3 Comparison with existing deep learning models

We compared the performance of our model VidAnomalyNet with the MobileNetV1 model and its transfer learning variant.

As presented in Table 3, the anomaly detection performance of VidAnomalyNet with public surveillance videos is provided with different optimizers.

As presented in Table 4, the anomaly detection performance of VidAnomalyNet using SGD optimizer with public surveillance videos is compared with the state of the art models.

Table 3. Performance of VidAnomalyNet with different optimizers

Anomaly Detection Model	Performance (%)			
	Precision	Recall	F1-Score	Accuracy
VidAnomalyNet with SGD Optimizer	0.93	0.94	0.94	0.965
VidAnomalyNet with Adam Optimizer	0.9	0.92	0.91	0.91
VidAnomalyNet with RMSProp Optimizer	0.91	0.91	0.91	0.91

Table 4. Performance of VidAnomalyNet with SGD optimizer is compared against existing models

Anomaly Detection Model	Performance (%)			
	Precision	Recall	F1-Score	Accuracy
MobileNetV1	0.78	0.87	0.81	0.921
MobileNetV1 with Transfer learning	0.91	0.78	0.82	0.95
VidAnomalyNet	0.93	0.94	0.94	0.965

As presented in Figure 21, the performance of VidAnomalyNet model is evaluated with different optimizers. Each optimizer showed its influence on the model in terms of performance. With Adam optimizer it showed precision 90%, recall 92%, F1-score 91% and accuracy 91%. With RMSProp optimizer it showed precision 91%, recall 91%, F1-score 91% and accuracy 91%. With SGD optimizer it exhibited precision 93%, recall 94%, F1-score 94% and accuracy 96.5%. Highest accuracy 96.5% is achieved when VidAnomalyNet model is used with SGD optimizer.

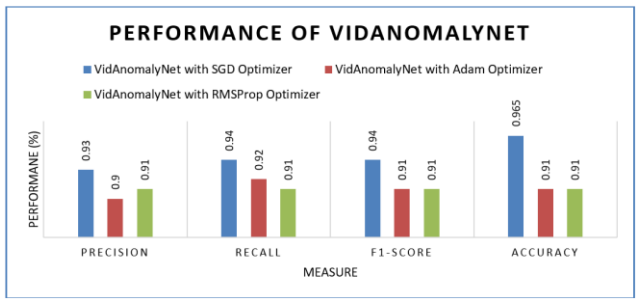


Figure 21. Performance of VidAnomalyNet model with different optimizers

As presented in Figure 22, the performance of VidAnomalyNet model with SGD optimizer is compared against state of the art models. MobileNetV1 is the existing model used in experiments. MobileNetV1 is also used with transfer learning in the empirical study. MobileNetV1 achieved precision 78%, recall 87%, F1-score 81% and accuracy 92.10%. MobileNetV1 with transfer learning showed better performance over MobileNetV1 with precision 91%, recall 78%, F1-score 82% and accuracy 95%. The proposed model VidAnomalyNet with SGD optimizer outperformed existing models with precision 93%, recall 94%, F1-score 94% and accuracy 96.50%. It is observed from the results that there is influence of transfer learning on MobileNetV1. Due to the transfer learning, MobileNetV1 with transfer learning could

improve detection performance significantly. However, both the existing models could not exceed the proposed model due to the fact that VidAnomalyNet follows novel approach in configuration of layers and their functioning. Therefore, the proposed model can be used in computer vision applications where public video surveillance is to be done in real time. Since it is a multi-class classifier with highest accuracy, its saved model can act as real time detector of anomalies from public surveillance videos.

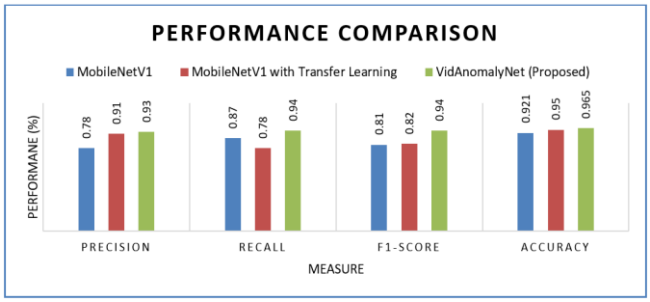


Figure 22. Performance of VidAnomalyNet model compared against state of the art

Surveillance video anomaly detection has various applications across different domains. Here is a list of specific applications where anomaly detection in surveillance videos plays a crucial role:

(1) Security and Public Safety:
Application: Identifying unusual activities or behaviors in public spaces, transportation hubs, and critical infrastructure to enhance overall security and public safety.
(2) Retail Loss Prevention:
Application: Detecting suspicious activities such as shoplifting, fraud, or other unauthorized behaviors in retail environments to minimize losses.
(3) Crowd Monitoring:
Application: Monitoring and detecting anomalous behaviors in crowded areas, such as events, stadiums, or public gatherings, to ensure crowd safety and manage potential security threats.
(4) Airport Security:
Application: Identifying unusual or threatening behaviors, abandoned objects, or unauthorized access in airport terminals to enhance aviation security.
(5) Critical Infrastructure Protection:
Application: Monitoring critical infrastructure sites, such as power plants, water facilities, or transportation networks, for any abnormal activities that could indicate security threats.
(6) Traffic Surveillance:
Application: Detecting abnormal traffic patterns, accidents, or incidents in real-time to improve traffic management and enhance road safety.
(7) Border Security:
Application: Monitoring borders and detecting unusual or unauthorized crossings, smuggling activities, or other security threats.
(8) Perimeter Security:
Application: Securing the perimeter of facilities, military bases, or sensitive areas by identifying intruders or suspicious activities.
(9) Banking and ATMs:
Application: Detecting anomalous behavior around ATMs or within bank branches, such as skimming devices or suspicious transactions, to prevent fraud.

5. CONCLUSION AND FUTURE WORK

The VidAnomalyNet deep learning architecture, which is based on the CNN model, was suggested. It is made to identify anomalies in surveillance footage and to have a more suitable learning process. The proposed a framework to exploit our VidAnomalyNet architecture for leveraging detection performance. We also proposed an algorithm known as VidAnomalyNet for Automatic Anomaly Detection (VAAD). At present, this algorithm predicts four classes. Out of them three classes pertaining to anomalies like fire, accident and robbery and the four one is NORMAL. It can be easily extended to identify more number of anomalies. We also explored MobileNetV1 with transfer learning by adding new layers to the base model for video anomaly detection. Our empirical study has revealed that VidAnomalyNet outperforms MobileNetV1. Highest accuracy achieved by the proposed model is 96.35%. In future, we intend to propose a Generative Adversarial Network (GAN) based architecture to exploit out deep learning model VidAnomalyNet along with other deep learning models for improving training quality and performance further. It is significant to remember that careful design, training, and hyperparameter tweaking are necessary for a GAN-based Video Anomaly Detection architecture to work well. Achieving effective integration also requires careful consideration of factors including training methodologies, loss functions, and GAN architecture selection. Determining the true impact on performance requires testing and validation on certain datasets.

REFERENCES

[1] Nawaratne, R., Alahakoon, D., De Silva, D., Yu, X. (2019). Spatiotemporal anomaly detection using deep learning for real-time video surveillance. IEEE Transactions on Industrial Informatics, 16(1): 393-402. https://doi.org/10.1109/TII.2019.2938527

[2] Kiran, B.R., Thomas, D.M., Parakkal, R. (2018). An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. Journal of Imaging, 4(2): 36. https://doi.org/10.3390/jimaging4020036

[3] Nayak, R., Pati, U.C., Das, S.K. (2021). A comprehensive review on deep learning-based methods for video anomaly detection. Image and Vision Computing, 106: 104078. https://doi.org/10.1016/j.imavis.2020.104078

[4] Revathi, A.R., Kumar, D. (2017). An efficient system for anomaly detection using deep learning classifier. Signal, Image and Video Processing, 11: 291-299. https://doi.org/10.1007/s11760-016-0935-0

[5] Duman, E., Erdem, O.A. (2019). Anomaly detection in videos using optical flow and convolutional autoencoder. IEEE Access, 7: 183914-183923. https://doi.org/10.1109/ACCESS.2019.2960654

[6] Kavikuil, K., Amudha, J. (2019). Leveraging deep learning for anomaly detection in video surveillance. In First International Conference on Artificial Intelligence and Cognitive Computing: AICC 2018. Springer

Singapore, pp. 239-247. https://doi.org/10.1007/978-981-13-1580-0_23
[7] Sultani, W., Chen, C., Shah, M. (2018). Real-world anomaly detection in surveillance videos. Computer Vision and Pattern Recognition, arXiv:1801.04264. https://doi.org/10.48550/arXiv.1801.04264
[8] Morais, R., Le, V., Tran, T., Saha, B., Mansour, M., Venkatesh, S. (2019). Learning regularity in skeleton trajectories for anomaly detection in videos. Computer Vision and Pattern Recognition, arXiv:1903.03295. https://doi.org/10.48550/arXiv.1903.03295
[9] Sánchez, F.L., Hupont, I., Tabik, S., Herrera, F. (2020). Revisiting crowd behaviour analysis through deep learning: Taxonomy, anomaly detection, crowd emotions, datasets, opportunities and prospects. Information Fusion, 64: 318-335. https://doi.org/10.1016/j.inffus.2020.07.008
[10] Chriki, A., Touati, H., Snoussi, H., Kamoun, F. (2021). Deep learning and handcrafted features for one-class anomaly detection in UAV video. Multimedia Tools and Applications, 80: 2599-2620. https://doi.org/10.1007/s11042-020-09774-w
[11] Ribeiro, M., Lazzaretti, A.E., Lopes, H.S. (2018). A study of deep convolutional auto-encoders for anomaly detection in videos. Pattern Recognition Letters, 105: 13-22. https://doi.org/10.1016/j.patrec.2017.07.016
[12] Zhao, Y.R., Deng, B., Shen, C., Liu, Y., Lu, H.T., Hua, X.S. (2017). Spatio-Temporal AutoEncoder for Video Anomaly Detection. In MM '17: Proceedings of the 25th ACM international conference on Multimedia, pp. 1933-1941. https://doi.org/10.1145/3123266.3123451
[13] Fan, Y., Wen, G., Li, D., Qiu, S., Levine, M.D., Xiao, F. (2020). Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder. Computer Vision and Image Understanding, 195: 102920. https://doi.org/10.1016/j.cviu.2020.102920
[14] Yang, B., Cao, J., Ni, R., Zou, L. (2018). Anomaly detection in moving crowds through spatiotemporal autoencoding and additional attention. Advances in Multimedia, 2018(1): 2087574. https://doi.org/10.1155/2018/2087574
[15] Pawar, K., Attar, V. (2019). Deep learning approaches for video-based anomalous activity detection. World Wide Web, 22(2): 571-601. https://doi.org/10.1007/s11280-018-0582-1
[16] Ragedhaksha, D., Shahil, N.A. (2021). Deep learning-based real-world object detection and improved anomaly detection for surveillance videos. Mater Today Proc. https://doi. org/10.1016/j. matpr, 64.
[17] Rezaee, K., Rezakhani, S.M., Khosravi, M.R., Moghimi, M.K. (2024). A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance. Personal and Ubiquitous Computing, 28(1): 135-151. https://doi.org/10.1007/s00779-021-01586-5
[18] Pang, G.S., Yan, C., Shen, C.H., Hengel, A.V.D., Bai, X. (2020). Self-trained deep ordinal regression for end-to-end video anomaly detection. Computer Vision and Pattern Recognition, arXiv: 2003.06780. https://doi.org/10.48550/arXiv.2003.06780
[19] Chackravarthy, S., Schmitt, S., Yang, L. (2018). Intelligent crime anomaly detection in smart cities using deep learning. In 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC), Philadelphia, PA, USA, pp. 399-404. https://doi.org/10.1109/CIC.2018.00060
[20] Saba, T. (2021). Real time anomalies detection in crowd using convolutional long short-term memory network. Journal of Information Science, 49(5): 1145-1152. https://doi.org/10.1177/01655515211022665
[21] Gong, M., Zeng, H., Xie, Y., Li, H., Tang, Z. (2020). Local distinguishability aggrandizing network for human anomaly detection. Neural Networks, 122: 364-373. https://doi.org/10.1016/j.neunet.2019.11.002
[22] Yang, F., Yu, Z., Chen, L., Gu, J., Li, Q., Guo, B. (2021). Human-machine cooperative video anomaly detection. Proceedings of the ACM on Human-Computer Interaction, 4: 1-18. https://doi.org/10.1145/3434183
[23] Yahaya, S.W., Lotfi, A., Mahmud, M. (2021). Towards a data-driven adaptive anomaly detection system for human activity. Pattern Recognition Letters, 145: 200-207. https://doi.org/10.1016/j.patrec.2021.02.006
[24] Shin, H., Na, K.I., Chang, J., Uhm, T. (2022). Multimodal layer surveillance map based on anomaly detection using multi-agents for smart city security. ETRI Journal, 44(2): 183-193. https://doi.org/10.4218/etrij.2021-0395
[25] UCF-Crime Dataset. (2018). Real-world anomaly detection in surveillance videos. https://www.crcv.ucf.edu/projects/real-world.
[26] Sultani, W., Chen, C., Shah, M. (2018). Real-world anomaly detection in surveillance videos. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 6479-6488. https://doi.org/10.1109/CVPR.2018.00678
[27] Le, V.T., Kim, Y.G. (2023). Attention-based residual autoencoder for video anomaly detection. Applied Intelligence, 53(3): 3240-3254. https://doi.org/10.1007/s10489-022-03613-1
[28] Abdullah, F., Jalal, A. (2023). Semantic segmentation based crowd tracking and anomaly detection via neuro-fuzzy classifier in smart surveillance system. Arabian Journal for Science and Engineering, 48(2): 2173-2190. https://doi.org/10.1007/s13369-022-07092-x
[29] Mavikumbure, H.S., Wickramasinghe, C.S., Marino, D.L., Cobilean, V., Manic, M. (2022). Anomaly detection in critical-infrastructures using autoencoders: A survey. In IECON 2022-48th Annual Conference of the IEEE Industrial Electronics Society, Brussels, Belgium, IEEE, pp. 1-7. https://doi.org/10.1109/IECON49645.2022.9968505
[30] Abbasi, A., Javed, A.R.R., Yasin, A., Jalil, Z., Kryvinska, N., Tariq, U. (2022). A large-scale benchmark dataset for anomaly detection and rare event classification for audio forensics. IEEE Access, 10: 38885-38894.
[31] Chidananda, K., Siva Kumar, A.P. (2022). Human anomaly detection in surveillance videos: A review. Information and Communication Technology for Competitive Strategies (ICTCS 2020) ICT: Applications and Social Interfaces, 791-802. https://doi.org/10.1007/978-981-16-0739-4_75
[32] Mumtaz, A., Sargano, A.B., Habib, Z. (2023). Robust learning for real-world anomalies in surveillance videos. Multimedia Tools and Applications, 82(13): 20303-20322. https://doi.org/10.1007/s11042-023-14425-x
[33] Shukla, S., Gupta, R., Garg, S., Harit, S., Khan, R. (2022). Real-Time parking space detection and management with artificial intelligence and deep learning system. Springer, pp. 127-139. https://doi.org/10.1007/978-3-

030-86749-2_7
[34] Pang, G., Aggarwal, C., Shen, C., Sebe, N. (2022). Editorial deep learning for anomaly detection. IEEE Transactions on Neural Networks and Learning Systems, 33(6): 2282-2286. https://doi.org/10.1109/TNNLS.2022.3162123
[35] Thakare, K.V., Sharma, N., Dogra, D.P., Choi, H., Kim, I.J. (2022). A multi-stream deep neural network with late fuzzy fusion for real-world anomaly detection. Expert Systems with Applications, 201: 117030. https://doi.org/10.1016/j.eswa.2022.117030
[36] Iliadis, L., Magri, L. (2022). Special issue on deep learning modeling in real life: Anomaly detection, biomedical, concept analysis, finance, image analysis, recommendation. Neural Computing and Applications, 34(22): 19397-19400. https://doi.org/10.1007/s00521-022-07832-y
[37] Ullah, W., Ullah, A., Hussain, T., Khan, Z.A., Baik, S.W. (2021). An efficient anomaly recognition framework using an attention residual LSTM in surveillance videos. Sensors, 21(8): 2811. https://doi.org/10.3390/s21082811
[38] Chidananda, K., Siva Kumar, A.P. (2023). A robust multi descriptor fusion with one-class CNN for detecting anomalies in video surveillance. International Journal of Safety & Security Engineering, 13(6): 1143. https://doi.org/10.18280/ijsse.130618