ISSN: 2278-4632
Vol-14, Issue-10, October: 2024

A Comparative Study of Deep Learning Approaches for anomaly detection in surveillance videos

MS. R. MARISWARI[1], Research Scholar, Reg No: 21211282282011, Department of Computer Science, St. Xavier's College (Autonomous), Palayamkottai, Affiliated to Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli-627 012, TamilNadu. India. Email: marissiram547@gmail.com

DR. V. NARAYANI[2], Assistant Professor, Department of Computer Science, St. Xavier's College (Autonomous), Palayamkottai-Tirunelveli**,** Affiliated to Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli-627 012, TamilNadu. India Email: narayaniv1979@gmail.com

Abstract

With the advent of computer vision technology, intelligent video surveillance programs have become indispensable to public safety by analyzing and comprehending lengthy video streams. One of the most important components of intelligent video surveillance is the identification and classification of anomalies. Finding anomalies automatically in a brief amount of time is the goal of anomaly detection. Video anomalies, which include anomalous activities and anomalous entities, are characterized by odd or irregular patterns in the video that deviate from the standard learned patterns. Oddities like weapons in critical areas and misplaced luggage, along with odd activities like brawling, rioting, breaking traffic laws, and stampedes, must to be promptly identified by an automated system. However, the complexity of human behavior, different ambient conditions, the ambiguity of the anomaly, and the absence of appropriate datasets make it difficult to detect video abnormalities. There are several works related to deep learning-based video anomaly detection however only few studies are there with enhance and hybrid deep learning model covering all the aspects in detecting abnormalities in surveillance video. This paper presents the eight significant deep learning model and compare with traditional CNN methods in terms of datasets, computational infrastructure, and performance metrics for both quantitative and qualitative analyses.

Keywords: *Video surveillance, video classification, deep learning, anomaly detection*

1. Introduction

The traditional manual analysis for labeling anomalies (such as traffic accidents, robberies, violent fights, etc.) in the volume of video data captured from public place monitoring is expensive and does not meet the actual requirements of video surveillance systems, given the rapid development and widespread use of these systems. As a result, there is an urgent need for an intelligent surveillance system that can identify abnormalities. In recent years, the intersection of computer vision and pattern recognition has seen a lot of activity [1], [2].

ISSN: 2278-4632
Vol-14, Issue-10, October: 2024

Among automated video surveillance's most significant, difficult, and urgent jobs is the identification of anomalous occurrences, like criminal activity, natural disasters, and traffic accidents and infractions. Thus, in recent years, video anomaly detection has emerged as a significant research topic.

The process of locating and recognizing unusual objects, actions, and occurrences in video footage is known as video anomaly detection [3]. Rather of employing particular semantic labels for fine-grained classification, it characterizes and quantifies all types of anomalous events based on a single criterion. Since the concept of abnormality is ambiguous, it is the most difficult problem in video anomaly identification. Video anomaly detection can be applied to intelligent surveillance system [3], defect detection [4], medical image processing [5], fault diagnosis [6], and so on.

Complex neural network topologies are the foundation of most recent state-of-the-art video anomaly detection techniques. Deep neural networks perform better than other neural networks on a number of machine learning and computer vision tasks, including object identification, image classification, game play, image synthesis and so on. There is also a great deal of discussion on their deficiencies with regard to the interpretability, analyzability, and dependability of their decisions when sufficiently big and comprehensive data sets are available for training [7].

Furthermore, none of the neural network-based video anomaly detection techniques have—as far as we are aware—been examined in terms of performance guarantees. Conversely, statistical and nearest neighbor approaches continue to be widely used because of their desirable features, which include resilience, computational efficiency, and amenability to performance analysis. Driven by the above-described research gaps and domain challenges. This paper analyzes the seven best performing enhanced deep learning model and compare their performance with respect to accuracy. This study focusses on detecting the crime activities in surveillance videos so the UCF Crime dataset is chosen for evaluation.

2. Deep learning Models in Anomaly detection

2.1 Adam-Dingo deep maxout network [8]

The Deep Maxout Network (DMN) [9] provides the benefit of effective performance in a very resource-constrained environment, which is why it is used for crime detection. The multi-layer organized Deep Maxout Network, which consists of several maxout layers connected successively, receives the summary video. When the maxout function is applied, each maxout layer produces hidden activations. Furthermore,

ISSN: 2278-4632
Vol-14, Issue-10, October: 2024

the hidden units that make up these layers have non-overlapping groups. The Deep Maxout Network's activation functions can be described by the following equations, and they are trainable.

$$A_{p,q}^{1} = \frac{max}{q \in [1, q_1]} D^{x}\, U_{\dots p,q} + W_{p,q} \quad (1)$$

$$A_{p,q}^{2} = \frac{max}{q \in [1, q_2]} \left(A_{p,q}^{1}\right)^{x} U_{\dots p,q} + W_{p,q} \quad (2)$$

$$A_{p,q}^{Z} = \frac{max}{q \in [1, q_z]} \left(A_{p,q}^{z-1}\right)^{x} U_{\dots p,q} + W_{p,q} \quad (3)$$

$$A_{p,q}^{y} = \frac{max}{q \in [1, q_y]} \left(A_{p,q}^{y-1}\right)^{x} U_{\dots p,q} + W_{p,q} \quad (4)$$

$$A_{p} = \frac{max}{q \in [1, q_y]} A_{p,q}^{y} \quad (5)$$

In equation (5) y denotes the total number of layers in the DMN, while q_z, with weight $U_{\dots p,q}$ and bias W_{pq}, denotes the number of hidden units accessible in the z^{th} layer. The Deep Maxout Network can estimate any random function by changing the value of x, and when x > 2, it can estimate the non-linear activation function. The output is provided by A_p, and the Deep maxout network successfully classifies the event as either a normal or stealing event.

The goal of creating a new Adam-Dingo optimizer is to reduce losses and increase the neural network's effectiveness. The Adam-Dingo optimizer that was developed efficiently modifies the neural network's weights and learning parameters, leading to enhanced performance. According to the parameter update equation of the Adam optimizer [10], the suggested optimizer is produced by enhancing the dingoes' encircling behavior when they hunt prey in the DOX [11].

The population-based algorithm known as the DOX is driven by the dingo's social behavior. The program takes into account the methods that dingoes employ to explore, encircle, and exploit their prey. The DOX algorithm has the benefit of solving the problem in real time with little effort, but multi-objective problems were not taken into account. Adaptive moment estimation is the basis of the stochastic optimization algorithm known as the Adam optimizer, which can be used to overcome this.

The Adam optimizer just needs the first-order gradient in order to optimize the objective functions. Additionally, by estimating the first and second-order gradient moments, the learning rates are calculated.

ISSN: 2278-4632
Vol-14, Issue-10, October: 2024

Non-convex optimization problems can be solved well with the aid of the Adam optimizer. Additionally, the optimizer has a high degree of robustness, however the method has problems with weight decay. Therefore, an efficient optimization approach that addresses the shortcomings of both strategies is established by combining the two optimization techniques.

2.2 COVAD [12]

This DLM proposes a new video anomaly detection technique that is primarily based on future frame prediction by combining memory module guidance with a content-based attention mechanism. The COVAD technique maps the video's attributes to the memory storage module after first learning its temporal and spatial characteristics. It then updates the memory storage module's records. Lastly, the error is evaluated, the difference between the actual and anticipated video frames is computed, and the video features are restored using the decoder network. In contrast to earlier approaches, this study integrates an encoder/decoder network that primarily examines the video content using the features acquired by the neural network, so modifying both the encoder and decoder networks and proposing a content-oriented self-attention mechanism.

COVAD method divides the continuous video frame sequence S of length X into two parts: the input frame, $\{1, (x-1)\}$, and the label, x^{th}. The training method uses the first $x-1$ frames as the input to extract the features set; $fI_{x-1} \in D^{W,H,K}$ K is the number of channels; the similarity index matrix $V \in D^{M,W*H}$ is then obtained by reading from memory $M \in D^{M,K}$.

After that, V updates the memory module by combining Mem, aggregate feature f, and V to get Agg. The model then gets the expected I nth frame by restoring the characteristics R. After that, V updates the memory module by combining Mem, aggregate feature fI_{x-1}, and V to get $A_f \in D^{2k,W*H}$.The model then gets the expected I_{xth}frame by restoring the features $A_f \in D^{2k,W*H}$. After getting the expected value from the model, the difference between the predicted and actual I nth frames is finally computed. During the training phase, a few more loss functions are implemented.

The three primary components of the algorithm suggested in this paper are the encoder, memory storage module, and decoder.

2.3 Encoders and decoders

ISSN: 2278-4632
Vol-14, Issue-10, October: 2024

U-Net has excelled in numerous international contests since it was first created as a CNN for image segmentation. Computer vision researchers have been influenced by its distinct structure and design philosophy, which includes symmetrical concepts, upsampling, and skip connections.

In order to detect video anomalies, CNNs extract video feature frames and use encoding and decoding to return the features to video frames. Because of its symmetric network topology, U-Net has a built-in advantage over alternative network designs.

It is comprised of several convolutional applications, pooling during the extract feature phase, and upsampling during the restoration phase. Maxpooling is unavoidable for upsampling, hence switch variables that store the maxpooling data, like the maximum value's location, can be included. These switches are used by upsampling in the decoder to reconstruct the current layer above into the proper places of the subsequent layer while maintaining the stimulus's structure.

3. Memory module

A sparse matrix M × C that is randomly generated makes up the module. Depending on the specific application scenario, the length and breadth of the matrix are M. It often represents the number of normal behaviors, movies, and camera positions in the training set. The breadth of the memory and the length of the features that the CNN extracted are equal to C.

3.1 IBaggedFCNet [13]

The suggested IBaggedFCNet for anomaly detection is shown in Figure 1. The two primary processes are the bagging-based deep learning classification model and the feature extraction using Inception-v3.

The Inception-v3 network [14] provides feature representations at both the frame and video levels. The openly accessible Inception-v3 tool is used to conduct unsupervised feature extraction. Principal Component Analysis (PCA) is used to produce a 1024-D vector that is then L2 normalized and quantized (1 byte per coefficient). The PCA mean vector and covariance matrix for the Train dataset were computed.

In this IBaggedFCNet, the basic model for our parallel bagging ensemble is a fully connected 3-Layer Neural Network. Averaging over the probabilistic output from multiple models is the next step in the process.

ISSN: 2278-4632
Vol-14, Issue-10, October: 2024

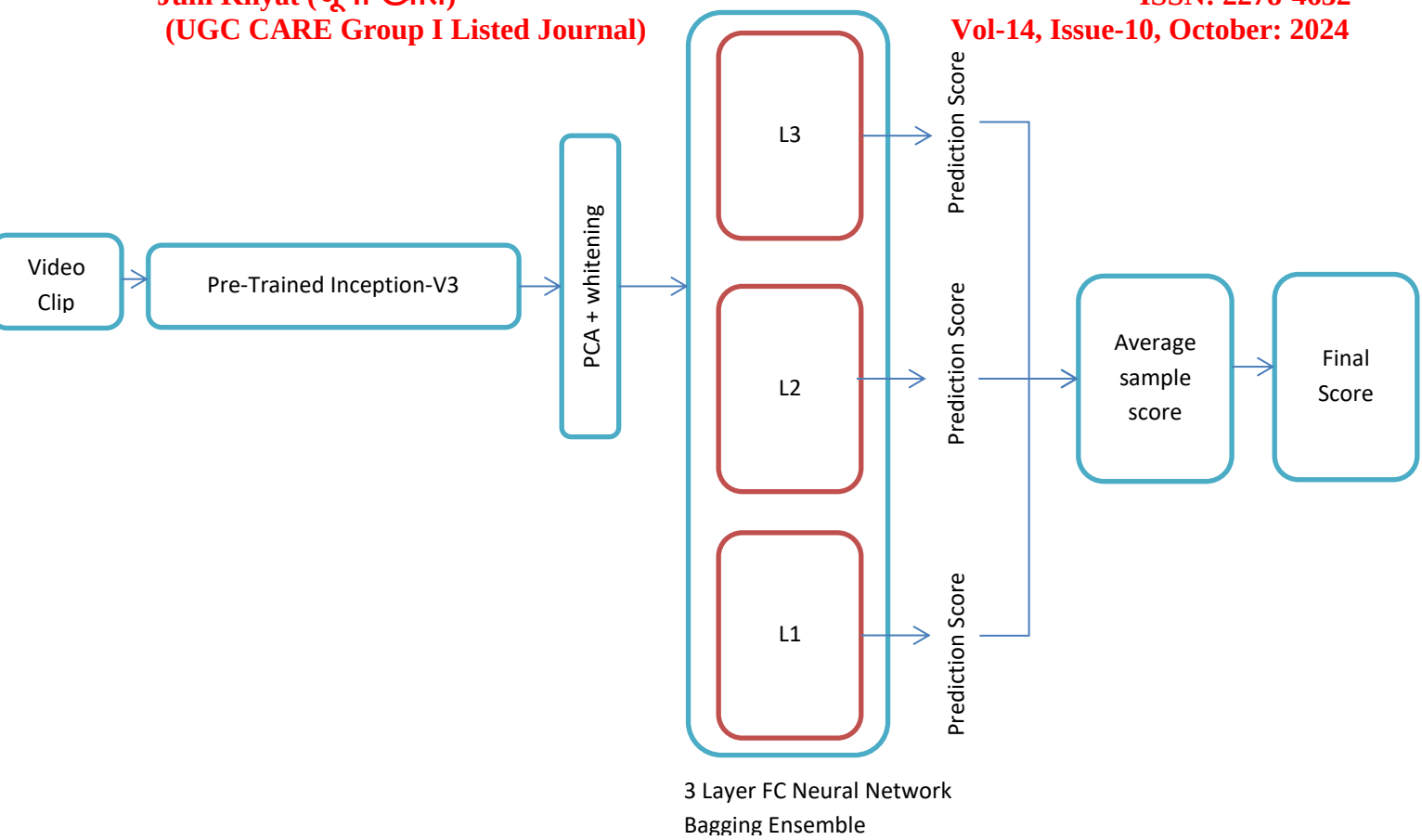

Figure: 1. IBaggedFCNet model architecture in anomaly detection

3.2 KFCRNet [15]

Combining the CNN and RNN models, the KFCRNet model has a feature selection method. After the CNN has extracted pertinent information from the video frames, the RNN uses these features to classify the frames. Because of this combination, the model is able to use both the geographical and temporal information when determining whether or not there is violence in the video. Compared to other violence detection models, the KFCRNet model has several advantages due to its utilization of CNNs and RNNs.

The CNN offers a strong approach to feature extraction, and the RNN enables the model to comprehend the temporal information found in the video. By combining the two models, the KFCRNet can manage video frame fluctuation more efficiently, which improves accuracy and performance. The KFCRNet model, which considers both the spatial and temporal information included in the video frames, is a potent and efficient method for detecting violence.

3.3 A3DConvNet [15]

ISSN: 2278-4632
Vol-14, Issue-10, October: 2024

A3DConvNet performs a more thorough and in-depth evaluation of fine-grained features than previous networks. With 19 3D convolution operations and 5 3D max-pooling operations, the network has 15 layers deep. There are two concatenation layers, three batch normalization layers, two completely connected layers, a flatten layer, and an output layer in this network. For every layer, there are different numbers and sizes of filters.

The architecture employs several kernel sizes, including$(3 \times 3 \times 3)$, $(5 \times 5 \times 5)$, and $(7 \times 7 \times 7)$, to carry out convolution operations in order to obtain precise information. The majority of convolution procedures employ the size $(3 \times 3 \times 3)$ filter because it can catch subtle, gradual information at a low level. It also reduces weight-sharing and calculation costs, which eventually results in smaller weights for back propagation. Kernels with dimensions of $(5 \times 5 \times 5)$ and $(7 \times 7 \times 7)$ can, nevertheless, extract general information from the data representation. This method is more comprehensible than others because of the features that are obtained via convolutional operation through kernels of both small and large sizes.

The images of the size ($120width \times 120height \times 3chanels$) are the input of the suggested architecture. With a kernel size of $(3 \times 3 \times 3)$, the first four 3D convolutions contain 16 filters. Subsequently, feature evaluation is carried out in separate branches for various filters of sizes, such as $(3 \times 3 \times 3)$, $(5 \times 5 \times 5)$, and $(7 \times 7 \times 7)$, and they are merged appropriately. The 3D convolved representation is flattened to obtain the final features, which are then sent to two fully linked layers of size 1024 and 128 respectively, before being passed to an output layer. Both fully linked layers use a 40% dropout rate as a regularizer to keep the network from overfitting. Equation 2 shows how the output layer labels the input video in its appropriate class using the softmax function.

$$Pro\left(\frac{C}{a}\right) = \frac{Pro\left(\frac{a}{c}\right) \times pro(c)}{\sum_{k=1}^{c} \quad Pro\left(\frac{a}{k}\right) \times pro(k)} \qquad (6)$$

Conditional probability is shown here by $Pro(\frac{c}{a})$. Class prior probability is denoted by P(c), and the value of C represents the total number of classes.

3.4 Lightweight Neural Network (LWNN) [16]

ISSN: 2278-4632
Vol-14, Issue-10, October: 2024

This technique for identifying anomalies in videos is precise and lightweight. The suggested technique examines the complete movie, automatically extracting and identifying the key characteristics needed to classify a video as normal or abnormal.

Assume that the dataset is $D = \{(I_a, L_a)\}$, where I_a represents the ith video in the training dataset and L_a is the label associated with I_a. $L_a = \{0, 1\}$, where 0 denotes ordinary and 1 abnormality. Only the normal condition is captured in every frame of a movie that is designated as normal. Conversely, the abnormally labeled movie has a mix of normal and abnormal states in its frames. The feature extractor divides I_a into K segments, each of which is transformed into a D-dimensional feature vector R_a, b: F_a, b denotes the bth feature vector of I_a, a feature extractor trained on the Kineticts dataset.

3.5 ResNet CNN models [17]

CNN + SRU has been built on the ResNet CNN model Due to its superior architecture and performance in anomaly detection tasks. In order to prepare ResNet for use on the appropriate datasets, it is necessary to load and update the model parameters on ImageNet. Due to the (240 × 240) input frame size, the ResNet can analyze (240 × 240 × 3) dimension data. After the Deep Residual Features (DRF) travel through the convolutional and pooling layers, they are transformed into a 4-D tensor (n × 1 × 1 × 2048), which is then given to the SRUs. The SRU layers get the ResNet outputs after they have been converted into (n × 3 × 3 × 128). Since ResNet does not use a completely linked dense layer for categorization, it is not used. Better paral lelization and gradient propagation allow the components to come together to form a simple, yet effective, design that is easy to scale.

3.6 SRU layer

The lightweight recurrent unit called the Simple Recurrent Unit (SRU) [18] strikes a balance between scalability and model capacity. SRU is designed to enable highly parallelized implementation, incorporate strict initialization, and give expressive recurrence in order to facilitate the training of deep models. Combining SRU with CNNs allows it to be used for anomaly identification in video frames. Compared to the Transformer model, data translation is possible with the inclusion of SRU in the design. The SRU design is introduced and explained in this section. A single layer of SRU requires the computations shown below:

$$C_a = \sigma(PM_c i_a) + I_c \odot B_{a-1} + x_c \quad (6)$$

ISSN: 2278-4632
Vol-14, Issue-10, October: 2024

$$R_a = \sigma(PM_R i_a) + I_R \odot B_{a-1} + x_R \quad (7)$$

$$F_a = c_a \odot B_{a-1} + (1 - C_a \odot PM\ i_R) \quad (8)$$

$$K_a = R_a \odot B_a + (1 - R_a) \odot i_a \quad (9)$$

where I_c, I_R, x_c, and x_R are parameters that need to be learned during training, and PM, PM_c, and PM_R are parameter matrices. Two sections might be made out of the entire network design: a light recurrence (Eqs. (6) and (7)) and a highway network (Eqs. (8) and (9)). In order to gather sequential data, the light recurrence section reads the input vectors i_a one at a time and computes the sequence of states F_a. The highway network component facilitates gradient-based deep network training. The input i_a and the state F_a generated by the light recurrence (Eq. (4)) are adaptively combined using the reset gate R_a (Eq. (6)), where $(1 - x_c) \odot i_a$ is a skip connection that enables the gradient to immediately propagate to the preceding layer. It has been demonstrated that these connections maximize scalability.

Results and discussion

The above mentioned seven deep learning algorithms are evaluated using the python platform with keras and tensorflow library with UCF crime dataset. Table 1 shows the accuracy values of each approach in detecting the anomaly. Among the seven models COVAD achieved the highest accuracy of 96.5 and the second highest accuracy is achieved by the LWNN model with 95.72%.

Table 1: Accuracy comparison of seven DL models

S.No	Reference	Model	Accuracy
1	[8]	Adam Dingo_Deep Maxout Network (ADDMN)	94.5%
2	[12]	**COVAD**	95.72%
3	[13]	IBaggedFCNet [13]	91%
4	[15]	KFCRNet	91.24%
5	[15]	A3DConvNet	92.06
6	[16]	Lightweight Neural Network (LWNN)	96.5%
7	[17]	ResNet50 + SRU	91%

ISSN: 2278-4632
Vol-14, Issue-10, October: 2024

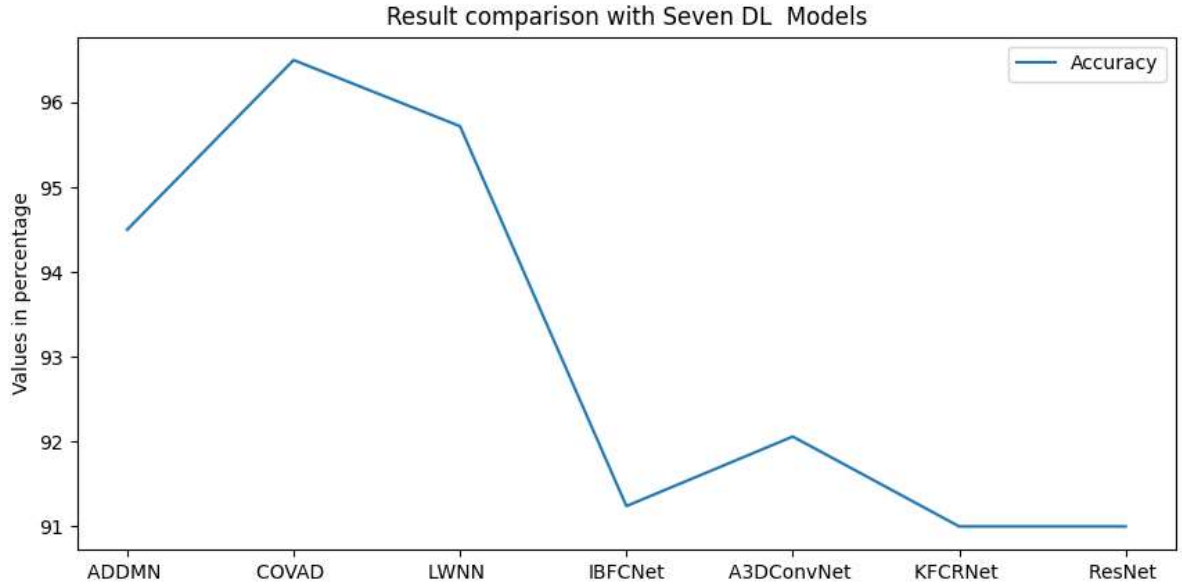

Figure: 2. Accuracy Comparison of Seven DL models

The accuracy of the discussed deep learning models is illustrated in figure 2. All the models are evaluated on the UCF crime data and performed better in detecting the anomaly in videos. Each model is executed on different parameter setting on python platform and provides the result greater than 90 %.

Conclusion

Advances in computer vision and the widespread usage of surveillance cameras in public spaces make research on anomalous activity identification more desirable. Applications for anomaly detection in the identification of suspicious activities have showed promise. While several works have been done on deep learning-based video anomaly detection, there aren't many that use enhanced or hybrid deep learning models that cover every facet of identifying anomalies in surveillance footage. In terms of datasets, computational infrastructure, and performance indicators for both quantitative and qualitative evaluations, this research compares seven important deep learning models with conventional CNN techniques. The seven classification model detects the anomaly effectively however the COVAD and LWNN performed well on UCF crime dataset and achieved highest accuracy. In future these two models can be enhanced with optimization technique to improve the detection accuracy.

References

ISSN: 2278-4632
Vol-14, Issue-10, October: 2024

[1] Brunetti, A., Buongiorno, D., Trotta, G.F. and Bevilacqua, V., 2018. Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. Neurocomputing, 300, pp.17-33.

[2] Szeliski, R., 2022. *Computer vision: algorithms and applications*. Springer Nature.

[3] Aziz, Z., Bhatti, N., Mahmood, H. and Zia, M., 2021. Video anomaly detection and localization based on appearance and motion models. *Multimedia Tools and Applications*, *80*(17), pp.25875-25895.

[4] Febin, I.P., Jayasree, K. and Joy, P.T., 2020. Violence detection in videos for an intelligent surveillance system using MoBSIFT and movement filtering algorithm. *Pattern Analysis and Applications*, *23*(2), pp.611-623.

[5] Bhatt, P.M., Malhan, R.K., Rajendran, P., Shah, B.C., Thakar, S., Yoon, Y.J. and Gupta, S.K., 2021. Image-based surface defect detection using deep learning: A review. *Journal of Computing and Information Science in Engineering*, *21*(4), p.040801.

[6] Saeed, H.A., Peng, M.J., Wang, H. and Zhang, B.W., 2020. Novel fault diagnosis scheme utilizing deep learning networks. *Progress in Nuclear Energy*, *118*, p.103066.

[7] Doshi, K. and Yilmaz, Y., 2020. Any-shot sequential anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 934-935).

[8] Waddenkery, N. and Soma, S., 2023. Adam-Dingo optimized deep maxout network-based video surveillance system for stealing crime detection. *Measurement: Sensors*, *29*, p.100885.

[9] W. Sun, F. Su, L. Wang Improving deep neural networks with multi-layer maxout networks and a novel initialization method Neurocomputing, 278 (2018), pp. 34-40

[10] D.P. Kingma, J. Ba Adam: A Method for Stochastic Optimization (2014) arXiv preprint arXiv:1412.6980

[11] A.K. Bairwa, S. Joshi, D. Singh Dingo Optimizer: A Nature-Inspired Metaheuristic Approach for Engineering Problems" Mathematical Problems in Engineering (2021)

[12] Shao, W., Rajapaksha, P., Wei, Y., Li, D., Crespi, N. and Luo, Z., 2023. COVAD: Content-oriented video anomaly detection using a self-attention based deep learning model. *Virtual Reality & Intelligent Hardware*, *5*(1), pp.24-41.

ISSN: 2278-4632
Vol-14, Issue-10, October: 2024

[13] Zahid, Y., Tahir, M.A., Durrani, N.M. and Bouridane, A., 2020. Ibaggedfcnet: An ensemble framework for anomaly detection in surveillance videos. *IEEE Access*, *8*, pp.220620-220630.

[14] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, ‘‘YouTube-8M: A large-scale video classification benchmark,’’ 2016

[15] Ansari, M.A., Singh, D.K. and Singh, V.P., 2023. Detecting abnormal behavior in megastore for intelligent surveillance through 3D deep convolutional model. *Journal of Electrical Engineering*, *74*(3), pp.140-153.

[16] Watanabe, Y., Okabe, M., Harada, Y. and Kashima, N., 2022. Real-World Video Anomaly Detection by Extracting Salient Features in Videos. *IEEE Access*, *10*, pp.125052-125060.

[17] Qasim, M. and Verdu, E., 2023. Video anomaly detection system using deep convolutional and recurrent models. *Results in Engineering*, *18*, p.101026.

[18] B. Riyono, R. Pulungan, A. Dharmawan, A.R. Antariksawan, A hybrid machine learning approach for improving fuel temperature prediction of research reactors under mix convection regime, Results in Engineering 15 (2022), 100612.