International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Hybrid ResNet-50 and LSTM Approach for Effective Video Anomaly Detection in Intelligent Surveillance Systems

Sreedevi R. Krishnan[1*], P. Amudha[2]

Submitted: 09/05/2023 **Revised**: 18/07/2023 **Accepted**: 07/08/2023

Abstract: Modern intelligent surveillance systems have given video anomaly detection much attention. Interior and outdoor monitoring devices are widespread in modern public places and smart cities. Due to the limited modelling capabilities and difficulty in capturing complicated relationships, conventional approaches have difficulty in recognizing video anomalies. This research tries to solve the problem using a hybrid strategy combining ResNet-50 (Residual Network-50) and Long Short-Term Memory (LSTM) algorithms for detecting anomaly video activity. Videos of normal and anomaly actions datasets enhanced the video clip's unique features, reducing the data required for storage and transmission. There are separate frames in each video. According to the proposed procedure, each frame is rescaled into a different pixel size before being fed into the ResNet-50 technique for feature extraction. Following feature extraction, the frames are then fed into the classification layer. The values of the new feature vectors are calculated by adding the original feature vectors acquired by ResNet-50. Finally, the LSTM model classifies the video as normal or abnormal using the information extracted from a sequence of frames. The LSTM assigns a classification to the retrieved images. The hybrid technique ResNet-50 and LSTM obtained 96.48% accuracy using the UCSD Ped 1 dataset. The proposed models outperformed the equivalent deep learning models and showed a noticeably higher performance accuracy.

Keywords: *Anomaly detection, Deep Learning, LSTM, Rescale, ResNet-50, Video surveillance.*

1. Introduction

In today's society, video surveillance data is quite relevant. Industrial facilities, educational institutions, and businesses are responsible for the extensive surveillance data availability. Similarly, cameras installed in public places like city centres, transportation hubs, and places of worship add to the public record. Several modules are used in the processing of surveillance video, including recognising objects and actions and classifying discovered behaviours into groups like abnormal or normal [1]. People can quickly lose their capacity to focus on the monitoring task since video data is growing, and only the security monitor staff can view the restricted amount of video data, which makes it difficult to spot suspicious behaviour on the screen [2]. Small spaces could only use manual surveillance systems, and security personnel were needed to watch for unusual activity. Manual surveillance, which is costly and requires extensive effort to detect suspicious activity, is not advisable [3]. An intelligent video surveillance system monitors critical locations particularly vulnerable to crime, such as ATMs, banking theft, fire detection, fight detection, etc. It can also play a critical role in security by identifying any suspicious activity [4].

Anomaly detection, essential to autonomous video surveillance, determines whether the other categories described above are successful [5]. Additionally, it is a challenging assignment because even a human observer would need help. After all, the abnormalities to be found are unknown in advance. Finding behaviours that differ from what is expected and usual behaviour is a general description of anomaly detection [6]. Three significant issues impact anomaly detection in the context of autonomous video surveillance. These issues are due to the computationally expensive and challenging nature of video data processing, space and time complexity, and the non-local temporal variations between video frames [7].

The definition of abnormal behaviour depends heavily on the standards established in the environment under consideration [8]. It is a security challenge, meaning some peculiar behaviour has caused an activity outside the parameters [9]. More and more surveillance cameras are being used for video analysis due to the growing need for security. Automatic anomaly detection may be more practical than labour-intensive anomalous event identification [10]. Anomaly identification is challenging because it is difficult to list negative samples and obtain enough negative samples due to their rarity [11].

Due to the rising demand for security, surveillance cameras are being employed for video analysis more and more. Anomaly detection may be a more practical approach than

[1] *Research Scholar, Department of Computer Science and Engineering, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India - 641043*
ORCID ID: 0000-0002-2769-7747
[2] *Professor, Department of Computer Science and Engineering, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India - 641043*
ORCID ID : 0000-0001-7763-8633
** Corresponding Author Email: sreedevirkrishnan@gmail.com*

labour-intensive abnormal event identification [12]. Popular techniques involve learning a model from recordings of normal events as training data and looking for abnormal events that deviate from the trained model. Deep learning (DL) and computer vision have been used recently to find the anomaly in the videos [13].

Deep learning has demonstrated its potential in various fields, including acoustics, pictures, and natural language processing. However, because abnormalities significantly differ from one another in various application contexts, it is challenging to create intelligent video anomaly detection systems [14]. Deep learning has shown increasingly astounding results in several previously believed to be computationally intractable tasks, such as face matching, recommendation systems, and anomaly detection. Video anomaly detection employs computer vision, which analyses footage to spot unusual activities or behaviours [15]. It employs advanced algorithms and techniques like motion analysis, object detection, recognition, and scene understanding. This approach enhances accuracy and timeliness in areas like surveillance, industrial monitoring, and critical event detection [16]. A deep learning algorithm's capacity to extract geographical and temporal elements from video footage automatically eliminates the necessity for manual feature engineering. This enables the models to detect anomalies effectively by capturing complex and dynamic video patterns. Also, the analysis of item appearances, motion patterns, and spatial connections in video data is used in computer vision-based video anomaly detection to find abnormalities that deviate from the normal [17]. Some deep learning models need to be more comprehensible and easier to understand. This can be a big problem, especially in critical applications where it is crucial to explain any anomalies found before making a decision [18].

The hybrid ResNet-50 and Long Short-Term Memory (LSTM) technique improves the detection of video anomalies. The combination of ResNet-50 and LSTM creates a resilient framework for video anomaly detection by combining the respective capabilities in spatial feature extraction and temporal modelling. The hybrid model delivers improved accuracy, robustness, and generalization by utilizing optical and temporal cues, making it suitable for various applications, including surveillance, industrial monitoring, and critical event detection.

The following are the primary contributions of the video anomaly detection method:

The area of video anomaly detection has made numerous vital strides because of deep learning and computer vision techniques.

- DL technique improves the accuracy of video anomaly detection by knowledge properties and representations from raw data. In addition to recording intricate patterns for thorough analysis and improved discrimination, Convolutional Neural Networks (CNNs) are excellent at extracting spatial data from video frames. Real-time video anomaly detection is optimized by deep learning models using effective architectures, compression, and hardware acceleration.

- Achieve accurate accuracy in video anomaly detection by combining the ResNet-50 and LSTM methods. Deep architecture ResNet-50 strikes a compromise between performance and computational effectiveness. This feature extraction technique is appropriate for detecting real-time or almost real-time video anomalies.

- LSTM enables more precise anomaly identification and differentiation between normal and abnormal patterns in complex and dynamic settings. LSTM can be combined with feature extraction techniques like deep CNN to gather spatial and temporal information for more precise and reliable anomaly identification in videos.

In the remaining sessions of this document, the following is how they are organized: A description of previous research on the subject is provided in the second section. The third section provides a complete clarification of the research methodology. The fourth section investigates the simulation of the suggested approach and provides examples of the study's findings. The conclusion summarizes the findings.

2. Literature Review

The recent developments in video anomaly detection from 2019 to the present were reviewed and discussed in this session. The goal of video anomaly detection, which seeks to identify unusual portions in a video sequence, is to identify anomalous segments in a video sequence. A 3D CNN-based encoder and a 2D CNN-based decoder are used in a Hao et al. [19] suggested spatiotemporal consistency-enhanced network. The latent space vector is resampled when training with normal data, but this method may not perform as well when abnormal data is included. Combining a discriminator, a transformed video clip, and the model allows us to assess the input clip's consistency. The results are more spatiotemporally consistent when adversarial training is used. However, applying these networks to applications where spatiotemporal consistency is unimportant could add extra complexity and overhead.

Urbanisation and autonomous industrial environments have led to an increase in the demand for intelligent real-time video monitoring. Using known normalcy training datasets and tackling developing anomalous behaviours are two issues Artificial Intelligence (AI) faces for anomaly identification. Nawaratne et al. [20] proposed that real-time video surveillance can be improved using the Incremental Spatio-Temporal Learner (ISTL), which addresses video

anomaly detection. Experiments demonstrate that ISTL is suited for use in industrial and urban contexts due to its accuracy, resilience, low processing cost, and contextual indicators. Only human intervention is required for the learning model's verification and improvement.

Several challenges hinder the advancement of AI in video surveillance anomaly detection. These challenges stem from the details associated with adapting to evolving anomalous behaviours, the reliance on datasets predominantly consisting of normal instances, and the scarcity of comprehensive evaluation methods. As video monitoring expands, the significance of automated anomaly detection techniques becomes paramount in recognising unusual events. While achieving high accuracy through deep learning methods is a prominent goal, it is worth noting that many studies predominantly focus on accuracy alone. Kim et al. [21] developed a Cross U-Net paradigm that examines accuracy and speed. The structure employs a DL model and a cascade sliding window technique to determine the anomaly score of a frame.

The available computer vision and machine learning algorithms for processing images and videos vary, rely on libraries, and demand powerful hardware. Arunnehru et al. [22] aim to create an AI-based system that processes video using a live CCTV camera feed while identifying and analysing occurrences. The system addresses important aspects like object detection, counting estimate, and anomaly detection using a convolutional neural network model while being simple to integrate and utilise with API calls. However, the Smart-Surveillance System is quite a powerful range of updates.

Studies train video anomaly detectors to identify underperforming frames during reconstruction and prediction tasks. Shao et al. [23] proposed a novel method for video anomaly identification utilising a generative network topology. The method learns spatial and temporal contexts through the order links between appearances and frames. According to experiments on four datasets, the technique enhanced detection performance on anomalies with various looks and motions. The approach's precision and recall were increased when it was combined with a prediction technique.

An Alex Net-based model for detecting crowd anomalies in video frames for images was suggested by Khan et al. [24]. The method consists of four convolutional layers, three fully connected layers, and a Rectified Linear Unit (ReLU) that acts as the activation function. The Convolution Layers (CLs) generate characteristics for anomaly detection. The model's Area Under the Curve (AUC) and overall accuracy were assessed on three reference datasets. The model achieved a 98% AUC utilising the Receiver Operating Characteristic Curve (ROC), outperforming baseline trials, according to the results. AlexNet is prone to overfitting like many deep learning models, especially when trained on limited datasets.

The increased usage of cameras for surveillance and tracking abnormal human behaviour in public and private settings has made it more challenging to identify and decipher abnormal conduct in real-world contexts. Khan et al. [25] put forward the idea of employing CNNs to detect anomalies in traffic videos, leading to the development of a robust continuing prediction system with remarkable accuracy. The study's primary emphasis lies in the realm of traffic accidents. Through the analysis of traffic surveillance footage, the trained CNN model exhibited an impressive 82% accuracy in identifying instances of traffic accidents. This approach holds the potential to contribute to the reduction of accident rates and the improvement of safety protocols in urban environments.

The use of stationary cameras is frequently the focus of visual surveillance systems. Urban missions increasingly use Unmanned Aerial Vehicles (UAVs), which provide a fresh aerial perspective. The absence of a study on anomaly identification in drone videos presents difficulties for automatic anomaly detection by UAVs. One-Class Support Vector Machine (OCSVM), Histogram of Oriented Gradient- 3 Dimensional (HOG3D), and CNN are three feature extraction algorithms are used. Chriki et al. [26] novel approach to anomaly detection for UAV-based surveillance missions are suggested with an AUC of 0.78 at the worst and 0.93 at the best; experiments on a dataset substantiate the efficacy of the offered tactics. The proposed method’s shortcoming is the failure to recognise some aberrant frames because of how similar they are to normal ones.

Massive amounts of video data are produced by surveillance systems, making it challenging for security experts to detect odd behaviour, Ul Amin et al. [27] suggested a DL model for anomaly detection in surveillance systems called EADN, which is just moderately complex. The technique combines a shot boundary detection technique to separate video into salient shots, a convolutional neural network to extract spatiotemporal information, and LSTM cells to detect anomalies. Studies using benchmark datasets and comparisons with cutting-edge techniques reveal a considerable improvement in the model's performance.

The amount of video data produced by surveillance systems makes analysis difficult for computer vision specialists. Anomaly detection with deep learning decreases the need for human labour while ensuring public safety. A practical methodology for deep features-based intelligent anomaly detection in surveillance networks is presented by Ullah et al. [28]. The framework captures spatiotemporal information from frames to reliably categorise anomalous/normal occurrences in intricate surveillance scenes. Subsequently, it uses a multilayer Bi-Directional

Long Short-Term Memory (BD-LSTM) model after sending them to a pre-trained CNN model. Studies reveal an improvement in accuracy of between 3.41% and 8.09% when compared to cutting-edge techniques.

Most of the currently available literature review focuses on the numerical difficulties of handling HD video data and finding contextual anomalies in surveillance video streams. To surmount the obstacles, present in video anomaly detection, the proposed approach presents a hybrid solution combining the ResNet-50 and LSTM techniques.

3. Methodology

Conventional techniques have issues in detecting video anomalies because of the limited modelling capacity and effort to extract nonlinear, complex correlations from video data. They have limited modelling capacity and find capturing complex and nonlinear relationships in video data difficult. As a result, they may fail to achieve the high-level semantics required for accurate anomaly detection. A hybrid approach combining ResNet-50 and LSTM methods as an effective solution for leveraging DL in video anomaly detection is proposed to address the abovementioned challenges. Here, characteristics from the video frame are extracted using a hybrid approach that combines ResNet-50 and LSTM approaches for video anomaly identification, and it demonstrated performance in image-based tasks, spatial and temporal context awareness, transfer learning abilities, robustness to changes, scalability, and computational efficiency. ResNet-50 is a viable option for extracting functional characteristics from video frames and enhancing the precision and efficacy of anomaly detection systems because of these benefits. Moreover, when used for classification in video anomaly detection, LSTM offers benefits in temporal dependency capture, sequential anomaly detection, memory retention, robustness to variations, contextual understanding, adaptability to variable-length sequences, and the capacity to use transfer learning. The overall performance of video anomaly detection systems is enhanced with the proposed model, which helps to accurately and effectively classify anomalies in video data.

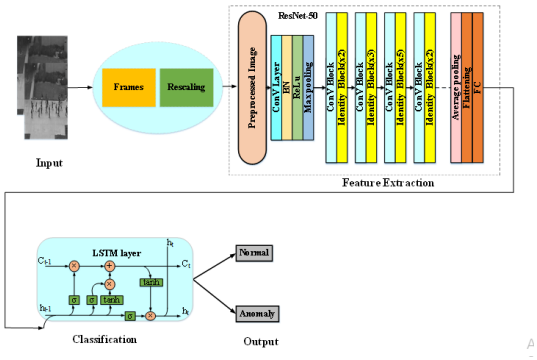

Fig. 1. Architecture diagram for the proposed model

The architecture of the suggested method, which finds video anomalies using LSTM classification, ResNet-50 feature extraction, and rescaling for preprocessing, is shown in Fig. 1. It comprises splitting the video into frames, resizing each frame, normalising pixel values, and utilising a pre-trained ResNet-50 technique. A sequence of feature vectors is created using the ResNet-50 model by extracting features from each frame. Sequential data's temporal dependencies are captured by an LSTM network, and an activation function creates a probability score for categorising video as normal or abnormal. The training and validation sets are built from the labelled video dataset, and the test set is used to assess the model. Following are the various stages of the suggested model.

Video input: The University of California, San Diego (UCSD) Ped1 dataset consists of video clips from UCSD security cameras.

Frame Extraction: After collecting video input, every video frame is resized to 256×256 pixels. The video is composed of multiple frames. The proposed approach is based on extending well-known greyscale textural features to sturdy, intense optical flow fields.

Preprocessing: Preprocessing is done using the rescaling method after the input is received, and it involves separating individual frames from the input video and resizing them to meet the input dimensions of the feature extraction method.

Feature extraction: Residual networks (ResNets) are effective neural models for feature extraction in Deep Neural Networks (DNN). They excel in extracting spatial features from input streams and have demonstrated outstanding performance on various widely used datasets. On many standard datasets, ResNets have displayed excellent performance. There are numerous ResNet variations, including ResNet-18, ResNet-26, ResNet-50, ResNet-101, and ResNet-152. Since videos may contain numerous redundant images, providing the complete video as input for preprocessing is not practical. Second, there are no time annotations available for videos of anomalies. It is helpful in these situations to extract significant features from videos and use them to train the model. To find an abnormality, the foundation of feature extraction is ResNet-50. The anomaly detector model receives the 1538 feature vector as input. The video is categorised as either class 0 (anomaly) or class 1 (normal). The network learns complex features with the aid of the activation function. ReLU is one of the most frequently utilised activation functions. In the feature extraction function, it is employed for video anomalies.

Classification: The suggested system's main objective is to enable visual anomaly detection. An LSTM was used in this detection to categorise a video frame. CNN is involved in the initialisation of the coarse and fine labels. Recurrent

Neural Network (RNN) is used to improve productive behaviour, and in the end, LSTM collaborates with RNN by using memory cells to give instructions to the learning model. In LSTM, internal memory cells are controlled by the input and forget gate network, developed to address the problem of vanishing or exploding gradients in RNN. The forget gate and the input gate, which alter the cell state, are ranked below the cell state. Choosing how much data from the previous memory should be transferred into the next time step is the primary responsibility of the forget gate in an LSTM. The output gate uses the tan h function to construct a vector, much how the input gate chooses how much new data should be fed into the memory cell. These gates enable the LSTM to manage the sequential information's short- and long-term dependencies. The following is the expression for the LSTM:

$$C_t^s = \tanh\left(\widehat{W}\mathfrak{h}_g * \mathfrak{h}_{t-1}^s + \widehat{W_{xg}^s} * x_t^s\right) \tag{1}$$

$$\mathfrak{f}_t^s = \sigma\left(\widehat{W}\mathfrak{h}_\mathfrak{f} * \mathfrak{h}_{t-1}^s + \widehat{W}_{xf}^s * x_t^s\right) \tag{2}$$

$$\mathfrak{i}_t^s = \sigma\left(\widehat{W}\mathfrak{h}_\mathfrak{o} * \mathfrak{h}_{t-1}^s + \widehat{W}_{x\mathfrak{i}}^s * x_t^s\right) \tag{3}$$

$$\mathfrak{o}_t^s = \sigma\left(\widehat{W}\mathfrak{h}_\mathfrak{i} * \mathfrak{h}_{t-1}^s + \widehat{W}_{x\mathfrak{o}}^s * x_t^s\right. \tag{4}$$

$$C_t^s = \mathfrak{f}_t^s \odot \mathfrak{o}_{t-1}^s + \widehat{W}_{x\mathfrak{o}}^s * x_t^s \tag{5}$$

$$\mathfrak{h}_t^s = \mathfrak{o}_t^s \odot \tanh(C_t^s) \tag{6}$$

The weights $\widehat{W}_{\mathfrak{h}*}$ throughout the parallel LSTM architecture, including the $\widehat{W}\mathfrak{h}_g, \widehat{W}\mathfrak{h}_\mathfrak{f}, \widehat{W}\mathfrak{h}_\mathfrak{i}$, and $\widehat{W}\mathfrak{h}_\mathfrak{o}$ parameters, are used in the equations above to control and monitor the information about the hidden states from the previous time step in addition to $\widehat{W}_{xg}, \widehat{W}_{xf}$, and $\widehat{W}_{x\mathfrak{o}i}$, which are applied for the weight matrices for the current input time steps. Input sequence information is shown by the superscript s, time step information is shown by the subscript t, the sigmoid function is utilised σ, and elementwise multiplication is represented by the symbol Θ.

The concept of residual learning was introduced to train ultra-deep CNNs specifically for image recognition tasks. The residual idea is used to explain the sequential information of the top-level layers, and levels are reformulated by locating residual functions with the input layer. The RL function is typically expressed as follows:

$$y = f(\ddot{X}, \ddot{W}) + \ddot{X} \tag{7}$$

The layers' input and subsequent sequential information vectors are called $\ddot{X}$ and y in Equation (7). The expression f(X, W) shows the residual knowledge from the connected layers $f(\ddot{X}, \ddot{W})$. Residual learning, which builds a sequence using the input given and nonlinear residual, contains the results of these layers. This method establishes a shortcut function across multiple layers for more effective model training, which is one of its main advantages. Additionally, it aids in preventing the fundamental issue of vanishing gradients brought on by composition with the adaptive residual $f(\ddot{X}, \ddot{W})$.

In this study, the normalisation layer in a residual LSTM is used to reduce the dynamic hidden state, normalise the information of the neurons for the LSTM, and shorten the deep RNN training time.

$$\tilde{n}_t = \frac{1}{h}\sum_{i=0}^{h}(\mathfrak{h}t)i \tag{8}$$

$$\delta_t = \sqrt{\frac{1}{h}\sum_{i=0}^{h}(((\mathfrak{h}t)^i - \tilde{n}_t)^2} \tag{9}$$

$$\dot{y}_t = f\left(\frac{\dot{g}}{\delta_t} \odot (\mathfrak{h}t - \tilde{n}_t) + b\right) \tag{10}$$

Where b is the trainable weights utilised to rescale the input sequence of the activation function, f, and $(\mathfrak{h}t)i$ is the hidden state in each layer of the LSTM of the ith neuron, and the time step is denoted by the subscript. Before the forward connection of each layer of the LSTM, a dropout threshold of 0.5 was applied to lessen overfitting. The baseline study improved the model's effectiveness for captioning video using an encoder and decoder with attention techniques.

A decoder that inputs a video feature matching to the next word based on the words the model had before created was employed by them for word generation. That uses latent correlation between characteristics in different places to account for short-term and long-term dependence. Contrary to the video captioning model, which inputs a feature vector from the video frame sequence to the LSTM and employs a single block of features for sequence learning.

$$k_t = \frac{1}{\mathbb{h}}\sum_{i=1}^{\mathbb{h}} \widehat{W}_t^i \mathfrak{f}_i \tag{11}$$

$$\dot{S}_t = \widehat{W}^T \tan h(\widehat{W}_\mathfrak{h}\,\mathfrak{h}t + M_\mathfrak{h}R_\mathfrak{f} + b_\mathfrak{h}) \tag{12}$$

$$\mathrm{A}_t = softmax\,(S_t) \tag{13}$$

The parameters learned for the frame features $\mathfrak{f}_i$ according to the attention weight $\widehat{W}_t^i$ to return the score S_t. are $\widehat{W}^T, \widehat{W}_\mathfrak{h}, M_\mathfrak{h}, b_\mathfrak{h}$ in the equations. The output possibilities obtained from the Softmax activation layer are shown in A_t at the end. The Softmax layer is used to make the final predictions, and the deep characteristics collected are then used to identify whether the sequence consists of typical events or aberrant activity.

4. Results and Discussion

Thousands of video frames from video datasets were used to train the model, and comprehensive tests were used to assess the effectiveness of the new approach.

4.1. Datasets

The UCSD dataset comprises many video clips previously labelled by the creator as normal or with anomalies. The UCSD Pedestrian 1 (Ped 1) dataset contains 1307 training image frames and 231 testing ones. A fixed camera observing pedestrian walkways was used to gather the UCSD Anomaly Detection Dataset. Non-pedestrian creatures or strange pedestrian motion patterns caused abnormal dealings. Bicyclists, skateboarders, tiny carts, pedestrians crossing walkways, and wheelchair users were frequent outliers. Binary flags per frame and manually created pixel-level binary masks for Peds2 were included in the ground truth annotations, which made it possible to assess how well an algorithm can locate abnormalities.

4.2. Performance Evaluation

This section shows how effectively video anomaly image frames divided into normal or anomalous categories may be classified using the proposed hybrid ResNet-50 and LSTM methods. The training process about the 50 epochs displays the accuracy and loss analyses of the suggested model. The suggested model gains significantly more accuracy and loses utility. It shows an appreciably fast convergence of the suggested model. Fig. 2. (a) and (b) show the evolution of training in accuracy and loss analysis.

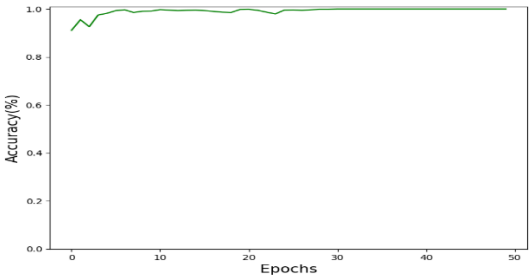

Fig. 2 (a). Training progress of accuracy analysis

The accuracy level of this training progression is 99.67%, which is very high and comparable to the 50th epoch.

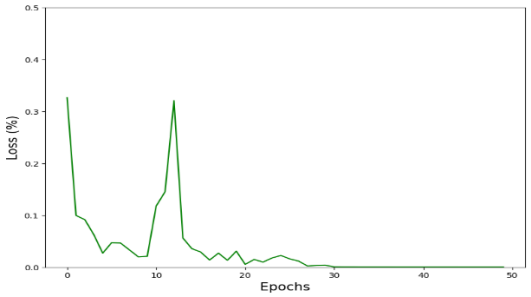

Fig. 2 (b). Training progress of loss analysis

In this training progress, the loss level is very high near the 50th epoch; the loss level of this process is 0.011%.

4.3. Performance Metrics

The following are the performance metrics: Accuracy, Precision, Specificity, Recall, F1-score, and AUC.

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+TN+FP} \tag{14}$$

The "True Positive Rate," called the Sensitivity or Recall criteria, is the percentage of positive events correctly classified as positive.

The Precision criteria display the percentage of examples from a specific class that were correctly identified compared to all cases from that class, regardless of whether they were correctly or wrongly classified.

$$Precision = \frac{TP}{TP+FP} \tag{15}$$

The Specificity criteria, called the "True Negative Rate," is used when accurate negative detections are essential. It is the antithesis of the Sensitivity parameter. The percentage of incidents that were classified as negative is shown by this metric.

$$Specificity = \frac{TN}{TN+FP} \tag{16}$$

A recall is a performance metric widely used in classification tasks to measure how well a model can pick out positive examples from a dataset.

True Positives (TP) refer to the count of events that the technique correctly classified as positive.

False Negatives (FN) is the percentage of illustrations that the model misclassified as positive when they were actually negative.

$$\text{Recall} = \frac{TP}{(TP+FN)} \tag{17}$$

The F1-Score criterion compromises between the Accuracy and Precision demands when False Positive and False Negative detections have different costs. This amount where displayed as

$$F1-score = 2 \times \frac{FN}{TP+FN} \tag{18}$$

The AUC, particularly on the ROC curve, indicates the technique's capacity to distinguish between positive and negative examples; it measures the two-dimensional area beneath the curve. A higher AUC value denotes superior overall performance and discriminative capacity.

$$AUC = \sum \frac{(TPR[i]+TPR[i+1])}{2} * FPR[i+1] - FPR[i]) \tag{19}$$

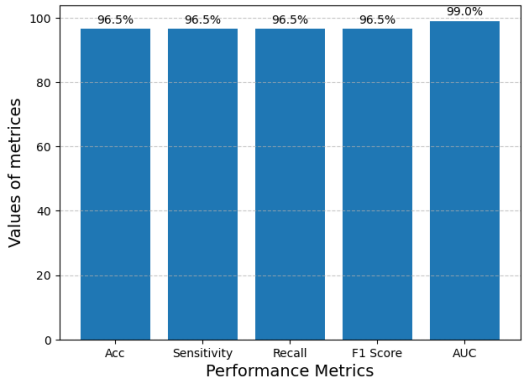

Fig. 3. Performance obtained from the proposed method

Fig. 3, which displays the performance metrics of a proposed video anomaly detection approach, a mix of ResNet-50 and LSTM methodologies, illustrates the overall effectiveness of video anomaly detection. Various performance criteria can be used to assess a hybrid model for video anomaly detection that combines ResNet-50 and LSTM. Typical measurements include Mean Average Precision (MAP), True Positive Rate (TPR), False Positive Rate (FPR), Precision, F1 Score, Accuracy, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These performance metrics aid in determining how well the hybrid ResNet-50 and LSTM model detects anomalies and prevents false alarms. Effective video anomaly detection is indicated by a high TPR, while minimising false alarms is done by a low FPR. Out of all the cases labelled as anomalies, precision determines the percentage of accurately detected anomalies. A higher F1 score indicates better overall performance. AUC-ROC assesses the model's capacity to distinguish between anomalies and regular events at various threshold settings. The average precision across various recall levels is then determined using Mean Average Precision (mAP). Combining these indicators enables a thorough assessment of the model's effectiveness in identifying abnormalities and preventing false alarms.

4.4. Results Obtained

The graph below displays typical (anomalous-free) and abnormal video frames. Fig. 4 interprets images from the video that are normal and abnormal.

Fig. 4. Video Anomaly Detection Normal image frames

Rescaling, which modifies the scale or range of pixel values in video frames, is an essential preprocessing step in detecting video anomalies. The data must be appropriate for the anomaly detection model to learn correctly, which is ensured in this stage. Techniques like normalisation or standardisation can be used to rescale. By dividing each pixel value by the highest value present in the video frames, normalisation modifies each pixel value to fall within a predetermined range, often between 0 and 1. The data is scaled to a similar range, and the mean is removed during standardisation, resulting in pixel values with a mean of 0 and a standard deviation of 1.

The properties of the video data and the precise needs of the anomaly detection task will determine whether to normalise or standardise. While standardisation is ideal for pixel values with a more extensive distribution, normalisation is appropriate for those with a specified maximum range. Rescaling preprocessing must be applied to prevent biases and provide consistency between training and test data. Based on the requirements of the anomaly detection model and the characteristics of the video data, additional preprocessing operations like scaling, cropping, or data augmentation should also be considered.

To extract valuable representations from video frames and apply them in identifying video anomalies, ResNet-50 is a deep convolutional neural network architecture. The procedure entails preprocessing the frames, loading the already trained ResNet-50 model, eliminating the classifier layers, and turning each frame into a collection of high-level characteristics known as feature vectors or embeddings. Commonly used to analyse temporal patterns, temporal aggregation produces features that can be used for anomaly detection. To aggregate temporal data, a variety of methods can be utilised, including averaging, pooling, and the usage of recurrent neural networks. By training a separate anomaly detection model or by employing unsupervised methods like clustering or density estimation, the generated feature vectors can be used for anomaly identification. This method captures rich representations of video frames using ResNet-50's potent feature extraction capabilities, assisting in detecting anomalies based on recognised visual patterns. The images of video anomaly detection after using the suggested strategies are shown in Fig. 5.

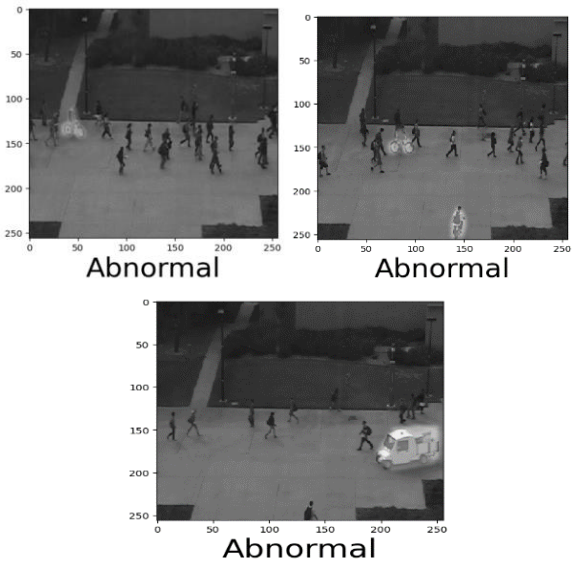

Fig. 5. Video Anomaly Detection Abnormal image frame

4.5. Comparative analysis

This section demonstrates how the proposed methods outperform other models in the hybrid ResNet-50 and LSTM method, allowing the proposed to serve as the method's primary framework. Compared to other networks with fewer parameters, a hybrid of ResNet-50 and LSTM can offer performance that is equivalent to or even better. They contrasted the outcomes of the method with those of recent research studies that have described methods like Sparse Reconstruction (SR) [29], Social Force (SF) [29], Anomaly Net [29], Abnormality Mining and Detection Network (AMDN) [29], Gaussian Mixture Model- Fully Convolutional Network (GMM-FCN) [29], Anomaly Over-exaggeration (AOE) [29], Generative Adversarial Networks (GAN) [29] Dual Spatio- Temporal Network (DSTN) [29], and Deep Regression Spatial- Temporal Network (DR-STN) [29].

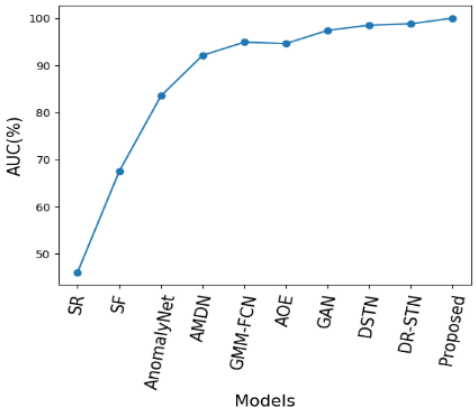

Fig. 6. Comparison of AUC

Fig. 6. displays comparisons of AUC. The AUC of anomaly detection can be increased by using a hybrid of ResNet-50 and LSTM. In terms of AUC, the proposed method performs better than SR, SF, Anomaly Net, AMDN, GMM-FCN, AOE, GAN, DSTN, and DR-STN by 67.5%, 46.1%, 92.1%, 97.4%, 83.5%, 98.5%, 94.9%, 94.6%, and 98.9%, respectively. As a result, conventional methods surpass the unique, revolutionary method, which has an AUC rate of 99.00%.

5. Conclusion

This work proposed a novel hybrid technique for finding anomalous behaviours in the UCSD dataset utilising ResNet-50 and LSTM approaches. The suggested method significantly improves the accuracy of identifying and categorising aberrant actions from the video data. Using ResNet-50, high-level features that reflect discriminative information about the activities may be effectively extracted. The LSTM is then given feature frames to differentiate between the normal and abnormal classes. The best model was proposed, with a 96.49% average classification accuracy. This model has 96.48% sensitivity, 96.48% recall, 96.48% F1-score, and 99.00% AUC values. The research has important implications for numerous practical applications using surveillance systems. ResNet-50 and LSTM-based automated abnormal activity recognition can significantly improve these domains' interactions, safety, and effectiveness. Future studies may examine how to enhance further and optimize the suggested strategy. This can entail researching other ResNet-50 and LSTM architectural variants, investigating cutting-edge data augmentation methods, or assessing the performance on more extensive and varied datasets.

References

[1] G. Sreenu, & S. Durai, "Intelligent video surveillance: a review through deep learning techniques for crowd analysis," *Journal of Big Data*, vol. 6, no. 1, pp. 1-27, 2019.

[2] A. Alshammari, & D. B. Rawat, "Intelligent multi-camera video surveillance system for smart city applications," *In 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC) IEEE*, pp. 0317-0323, 2019, January.

[3] Y. Feng, Y. Yuan, & X. Lu, "Learning deep event models for crowd anomaly detection," *Neurocomputing*, vol. 219, pp. 548-556, 2017.

[4] P. Khaire, & P. Kumar, "A semi-supervised deep learning-based video anomaly detection framework using RGB-D for surveillance of real-world critical environments," *Forensic Science International: Digital Investigation*, vol. 40, p. 301346, 2022.

[5] Y. Zhou, B. Li, J. Wang, E. Rocco, & Q. Meng, "Discovering unknowns: Context-enhanced anomaly detection for curiosity-driven autonomous

underwater exploration," *Pattern Recognition*, vol. 131, p. 108860, 2022.

[6] X. Cheng, L. Yuan, Z. Liu, & F. Guo, "Comparative analysis of video anomaly detection algorithms," *In International Conference on Computer, Artificial Intelligence, and Control Engineering (CAICE 2022) SPIE*, vol. 12288, pp. 431-436, 2022, December.

[7] R. F. Mansour, J. Escorcia-Gutierrez, M. Gamarra, J. A. Villanueva, & N. Leal, "Intelligent video anomaly detection and classification using faster RCNN with deep reinforcement learning model," *Image and Vision Computing*, vol. 112, p. 104229, 2021.

[8] L. A. Saleem, & E. V. Reddy, "A Survey on Deep Learning based Video Surveillance Framework," *In 2023 International Conference on Computer Communication and Informatics (ICCCI), IEEE*, pp. 1-6, 2023, January.

[9] K. Singh, S. Rajora, D. K. Vishwakarma, G. Tripathi, S. Kumar, & G. S. Walia, "Crowd anomaly detection using aggregation of ensembles of fine-tuned convents," *Neurocomputing*, vol. 371, pp. 188-198, 2020.

[10] N. L. Lavanya, N. Vijayananda, N. N. Sattigeri, R. K. Nisarga, & N. M. Pooja, "Survey on Abnormal Event Detection and Signalling in Multiple Video Surveillance Scenes Using CNN," *International Journal of Human Computations & Intelligence*, vol. 2, no. 3, pp. 159-168, 2023.

[11] J. Ren, F. Xia, Y. Liu, & I. Lee, "Deep video anomaly detection: Opportunities and challenges," *In 2021 international conference on data mining workshops (ICDMW) IEEE*, pp. 959-966, 2021, December.

[12] G. Pang, C. Shen, L. Cao, & A. V. D. Hengel, "Deep learning for anomaly detection: A review," *ACM computing surveys (CSUR)*, vol. 54, no. 2, pp. 1-38, 2021.

[13] E. Şengönül, R. Samet, Q. Abu Al-Haija, A. Alqahtani, B. Alturki, & A. A. Alsulami, "An Analysis of Artificial Intelligence Techniques in Surveillance Video Anomaly Detection: A Comprehensive Survey," *Applied Sciences*, vol. 13, no. 8, p. 4956, 2023.

[14] A. Berroukham, K. Housni, M. Lahraichi, & I. Boulfrifi, "Deep learning-based methods for anomaly detection in video surveillance: a review," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 1, pp. 314-327, 2023.

[15] M. Arunachalam, V. Sanghavi, Y. A. Yao, Y. A. Zhou, L. A. Wang, Z. Wen, & F. Mohammad, "Strategies for Optimising End-to-End Artificial Intelligence Pipelines on Intel Xeon Processors," *arXiv preprint arXiv:2211.00286*, 2022.

[16] H. T. Duong, V. T. Le, & V. T. Hoang, "Deep Learning-Based Anomaly Detection in Video Surveillance: A Survey," *Sensors*, vol. 23, no. 11, p. 5024, 2023.

[17] F. Mumcu, K. Doshi, & Y. Yilmaz, "Adversarial Machine Learning Attacks Against Video Anomaly Detection Systems," *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 206-213, 2022.

[18] Z. K. Abbas, & A. A. Al-Ani, "A Comprehensive Review for Video Anomaly Detection on Videos," *In IEEE 2022 International Conference on Computer Science and Software Engineering (CSASE)*, pp. 1-1, 2022, March.

[19] Y. Hao, J. Li, N. Wang, X. Wang, & X. Gao, "Spatiotemporal consistency-enhanced network for video anomaly detection," *Pattern Recognition*, vol. 121, p. 108232, 2022.

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

[21] Y. Kim, J. Y. Yu, E. Lee, & Y. G. Kim, "Video anomaly detection using Cross U-Net and cascade sliding window," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 6, pp. 3273-3284, 2022.

[22] J. Arunnehru, "Deep learning-based real-world object detection and improved anomaly detection for surveillance videos," *Materials Today: Proceedings*, vol. 80, pp. 2911-2916, 2023.

[23] W. Shao, R. Kawakami, & T. Naemura, "Anomaly Detection Using Spatio-Temporal Context Learned by Video Clip Sorting," *IEICE TRANSACTIONS on Information and Systems*, vol. 105, no. 5, pp. 1094-1102, 2022.

[24] A. A. Khan, M. A. Nauman, M. Shoaib, R. Jahangir, R. Alroobaea, M. Alsafyani, & C. Wechtaisong, "Crowd Anomaly Detection in Video Frames Using Fine-Tuned AlexNet Model," *Electronics*, vol. 11, no. 19, p. 3105, 2022.

[25] S. W. Khan, Q. Hafeez, M. I. Khalid, R. Alroobaea, S. Hussain, J. Iqbal, & S. S. Ullah, "Anomaly

detection in traffic surveillance videos using deep learning," *Sensors*, vol. 22, no. 17, p. 6563, 2022.

[26] A. Chriki, H. Touati, H. Snoussi, & F. Kamoun, "Deep learning and handcrafted features for one-class anomaly detection in UAV video," *Multimedia Tools and Applications*, vol. 80, pp. 2599-2620, 2021.

[27] S. Ul Amin, M. Ullah, M. Sajjad, F. A. Cheikh, M. Hijji, A. Hijji, & K. Muhammad, "EADN: An Efficient Deep Learning Model for Anomaly Detection in Videos," *Mathematics*, vol. 10, no. 9, p. 1555, 2022.

[28] W. Ullah, A. Ullah, I. U. Haq, K. Muhammad, M. Sajjad, & S. W. Baik, "CNN features bi-directional LSTM for real-time anomaly detection in surveillance networks," *Multimedia tools and applications*, vol. 80, pp. 16979-16995, 2021.

[29] T. Ganokratanaa, S. Aramvith, & N. Sebe, "Video anomaly detection using deep residual-spatiotemporal translation network," *Pattern Recognition Letters*, vol. 155, pp. 143-150, 2022.

[30] Meneses-Claudio, B. ., Perez-Siguas, R. ., Matta-Solis, H. ., Matta-Solis, E. ., Matta-Perez, H. ., Cruzata-Martinez, A. ., Saberbein-Muñoz, J. ., & Salinas-Cruz, M. . (2023). Automatic System for Detecting Pathologies in the Respiratory System for the Care of Patients with Bronchial Asthma Visualized by Computerized Radiography. International Journal on Recent and Innovation Trends in Computing and Communication, 11(2), 27–34. https://doi.org/10.17762/ijritcc.v11i2.6107

[31] Kwame Boateng, Machine Learning in Cybersecurity: Intrusion Detection and Threat Analysis , Machine Learning Applications Conference Proceedings, Vol 3 2023.