DOI 10.15622/ia.23.6.12

S.R. KRISHNAN, P. AMUDHA

ENHANCING VIDEO ANOMALY DETECTION WITH IMPROVED UNET AND CASCADE SLIDING WINDOW TECHNIQUE

Krishnan S.R., Amudha P. **Enhancing Video Anomaly Detection with Improved UNET and Cascade Sliding Window Technique.**

Abstract. Computer vision video anomaly detection still needs to be improved, especially when identifying images with unusual motions or objects. Current approaches mainly concentrate on reconstruction and prediction methods, and unsupervised video anomaly detection faces difficulties because there are not enough tagged abnormalities, which reduces accuracy. This paper presents a novel framework called the Improved UNET (I-UNET), designed to counteract overfitting by addressing the need for complex models that can extract subtle information from video anomalies. Video frame noise can be eliminated by preprocessing the frames with a Weiner filter. Moreover, the system uses Convolution Long Short-Term Memory (ConvLSTM) layers to smoothly integrate temporal and spatial data into its encoder and decoder portions, improving the accuracy of anomaly identification. The Cascade Sliding Window Technique (CSWT) is used post-processing to identify anomalous frames and generate anomaly scores. Compared to baseline approaches, experimental results on the UCF, UCSDped1, and UCSDped2 datasets demonstrate notable performance gains, with 99% accuracy, 90.8% Area Under Curve (AUC), and 10.9% Equal Error Rate (EER). This study provides a robust and accurate framework for video anomaly detection with the highest accuracy rate.

Keywords: anomaly detection, I-UNET, weiner filter, ConvLSTM, cascade sliding window, anomaly score.

1. Introduction. A significant task in video anomaly identification is recognising and localising unexpected occurrences in both place and time inside a video. These anomalies depict out-of-the-ordinary behaviours or events that may indicate possible concerns or security vulnerabilities. Depending on the context, anomalies may also be referred to as abnormalities, novelties, or outliers [1]. Unattended bags at airports, persons collapsing unexpectedly, or someone lingering suspiciously outside a guarded facility are some examples of video oddities [2]. Recognising and localising unusual occurrences in both place and time inside a video is a significant task in video anomaly identification. These anomalies describe unusual actions or events that may signal potential issues or security vulnerabilities [3]. Anomalies are also known as abnormalities, novelties, or outliers, depending on the context. This detection and analysis of anomalies is critical for improving security measures and addressing possible issues in various applications [4].

Anomaly detection in the video refers to automatically recognising aberrant events or behaviour within the spatiotemporal aspects of a video. It entails detecting actions or things that do not follow expected patterns or behaviours. Notably, the detection and localisation of video anomalies are

inextricably linked [5 – 6]. Real-time detection of anomalies in video data is critical because it allows immediate action to be taken upon recognising these anomalies, thereby preventing or mitigating adverse outcomes [7]. As a result, extensive research is being conducted to automate the process of detecting unusual occurrences in video surveillance systems. However, it might be hard to spot abnormalities in video broadcasts [8].

One important machine learning application is video anomaly detection, which looks for abnormal events or patterns in video data. This technology is critical in many areas, including security, surveillance, and industrial quality control [9]. It uses advanced algorithms and deep learning approaches to detect odd behaviours or events from the norm in a video frame. These systems can detect anomalies such as intruders in a secure facility, equipment breakdowns in manufacturing, or traffic accidents on the road by training models on massive datasets of usual events [10]. Machine learning has made significant advances in video anomaly detection. However, it still needs to improve, such as the necessity for substantial labelled data, the high processing intensity of deep learning techniques, and interpretability issues. These difficulties may make machine learning less accessible to smaller organisations or applications with limited resources [11]. Real-time processing requirements put existing infrastructure under pressure, and adaptation to changing video settings and anomalies can be constrained. As a result, while machine learning offers promise for video anomaly identification, careful analysis and resolution of these difficulties are required for successful implementation in varied applications [12].

Deep learning video anomaly detection is an advanced and sophisticated way of recognising odd events or behaviours inside video frames. This field of study uses the capabilities of deep neural networks, which are artificial intelligence systems designed to imitate the complicated functions of the human brain [13]. Deep learning for video anomaly detection entails training these neural networks to recognise and interpret typical activity patterns in video data. After learning what defines regular behaviour, the model can identify deviations from these established norms as anomalies or probable outliers [14]. Deep learning approaches, including CNNs and RNNs, are especially well-suited for detecting visual anomalies. CNNs thrive at analysing spatial information inside individual video frames, but RNNs excel at capturing temporal dependencies and event sequences. Combining these two types of neural networks allows the model to understand complicated spatiotemporal correlations, making it highly effective at detecting anomalies in video data [15].

Deep CNNs for video anomaly detection represent a cutting-edge way to detect unexpected events or anomalies inside video sequences. Deep CNNs have transformed Computer Vision tasks by automatically allowing models to learn and extract complicated spatial characteristics from pictures or video frames [16]. Deep CNNs excel in capturing intricate visual patterns and deviations when applied to video anomaly detection, making them a powerful tool in this field [17]. Video anomaly identification is a complicated task involving computer vision to identify unexpected motions or objects. Existing approaches concentrate on reconstruction and prediction but need help with obstacles such as low accuracy and complexity. A unique strategy for improving efficiency and accuracy by avoiding overfitting is suggested.

The following are the research work's key contributions:

1. The Improved UNET (I-UNET) is introduced in this work to improve anomaly detection in video frames by resolving overfitting and noise concerns, improving efficiency and accuracy.
2. Using a Weiner filter to preprocess video frames effectively reduces noise, resulting in cleaner frames for analysis and increased robustness for reliable anomaly identification.
3. The model employs an encoder-decoder architecture for efficient feature extraction, improved spatial and temporal information representation, and anomaly detection accuracy.
4. The Cascade Sliding Window Technique (CSWT) is utilised for anomaly detection in the post-processing phase, giving a sophisticated examination of frames and distinguishing between normal and abnormal ones.

The remaining manuscript is arranged as follows: The research strategy was explained in detail in the third segment, which also covers existing research. The fourth section simulates the suggested method and presents the research findings. Furthermore, a summary of the study's findings is given in the conclusion.

2. Literature Review. The research of video anomaly detection utilising deep learning and computer vision has experienced spectacular advances in recent years, with the emergence of various complicated algorithms. The researcher has provided several effective methods for detecting anomalies, as listed below.

The authors in [18] proposed a 3-stage ensemble-based unsupervised deep reinforcement algorithm for automated live video frames analytics, which employs a LSTM-based RNN for generating anomaly scores. The algorithm uses the least square method for optimal score creation, and model updates are accomplished by award-based reinforcement learning.

This method is intended for GPU and TPU-supported frameworks. However, the algorithm's capacity to handle many video frames simultaneously might be difficult, especially in high-demand applications, raising scalability difficulties.

A unique convolution autoencoder architecture for visual anomaly detection is presented in [19]. The architecture distinguishes between normalcy in appearance and motion behaviour and aberrant events by separating spatial and temporal information. The temporal autoencoder simulates optical flow using RGB difference, while the spatial autoencoder models normalcy by recreating the first individual frame. The method uses a deep Kmeans cluster strategy and a variance-based attention module to boost detection performance on rapidly moving outliers. However, limitations in accurately capturing complex motion patterns increase computational complexity, especially with quickly moving outliers.

Paper [20] frequently extracted low-level spatiotemporal features while ignoring semantic data. Deep learning algorithms, specifically CNN, are capable of extracting high-level information. A new hybrid visual embedding method was introduced for anomaly identification. The technique computes feature per frame with a pre-trained deep model, learns topic distributions with multilayer nonnegative matrix factorisation, and finds typical normal clusters with K-means. Experimental data demonstrate the method's usefulness in detecting anomalies. However, complex images with clutter or overlapping components and visual embedding may need help finding anomalies, possibly emphasising unimportant parts or failing to detect minor ones.

A residual spatiotemporal autoencoder for anomaly identification in security footage is suggested by the [21]. The method takes advantage of normalcy modelling to find departures from standard patterns. The trainable end-to-end autoencoder uses reconstruction loss to detect aberrant frames. Regarding cross-dataset generalisation, residual blocks work better than deeper layers because of their incremental effectiveness. However, the proposed technique has limitations such as adaptability, complexity, and generalizability to diverse datasets compared to deeper layers.

In study [22] the authors introduced an EADN deep learning-based approach. It splits video into prominent shots, extracts spatiotemporal information with a CNN, and learns spatiotemporal features with LSTM cells. The model's utility is demonstrated by extensive testing on benchmark datasets and comparisons with the most advanced techniques. Nonetheless, there is still room in the EADN design for improving efficiency and accuracy in real-time.

In paper [23] created a video anomaly detection system that depends on unsupervised frame prediction and enhances overall performance. The method is based on a U-Net-like structure that consists of a memory module for storing standard patterns, a Time-distributed 2D CNN-based encoder and decoder, and a multi-branch structure for extracting contextual information. However, this method may result in overfitting.

In paper [24] the authors projected a reliance on the reconstruction or prediction of future frames. In most approaches, accuracy is impacted by the requirement for more excellent temporal continuity between video frames. Using a hybrid dilated convolution module and DB-ConvLSTM module, a novel technique combines these two models. Experiments show this approach detects abnormalities more correctly in diverse video settings than state-of-the-art technologies. However, the completeness of the training data for each scenario is a prerequisite for the proposed model.

Paper [25] suggested a deep CNN encoder & multi-stage channel attention decoder for autonomous anomaly detection in video surveillance systems. Temporal shift methods and channel attention modules were used for contextual dependency extraction. However, the proposed method is computationally expensive, especially for high-resolution videos.

In paper [26] a novel anomaly detection approach for surveillance operations, focused on mobile cameras. Three techniques were employed to extract robust features from Unmanned Aerial Vehicle (UAV) footage: One-Class Support Vector Machine (OCSVM), two manually constructed approaches called Histogram Oriented Gradient (HOG) and Histogram Oriented Gradient 3 Dimensional (HOG3D), and a pre-trained CNN.

The model by the author in [27] for detecting anomalies was ineffective because of significant differences within and across classes. Two novel multi-view representation learning approaches were proposed: a hybrid multi-view representation learning that combined robust handcrafted features with deep features from 3D-STAE and a deep multi-view representation learning that combined features from two-frames SpatioTemporal AutoEncoder and deep features from 3D-STAE. The video anomaly detection with several existing works is tabulated in Table 1.

As a result, managing several video frames at once may present difficulties for the anomaly detection method in complicated image processing, particularly in high-demand applications. Its ability to precisely capture intricate motion patterns and growing computational complexity is likewise limited. Notwithstanding these drawbacks, the EADN design can be strengthened to increase real-time accuracy and efficiency.

Table 1. Video Anomaly Detection with state-of-art-of techniques

No.	Technique	Objectives	Advantages	Limitation/ future scope	Result
[18]	Deep Kmeans cluster strategy	Convolutional autoencoder for anomaly detection in videos, separating spatial and temporal information to capture appearance and motion behaviour separately	Enhances anomaly detection accuracy by effectively capturing appearance and motion behaviour independently by dissociating spatial and temporal representations	It may introduce additional computational complexity, potentially requiring more resources for training and inference	9% reduction in EER of UCSD Ped1, a 13% reduction in ERR of UCSD Ped2 and a 4% improvement in accuracy in both datasets
[19]	CNN	Aims to develop a CNN-based methodology for automatically detecting traffic accidents in surveillance videos from video traffic surveillance systems	Automated detection of traffic accidents in surveillance videos, reducing reliance on manual monitoring and enabling prompt emergency response	It potentially limits its performance in detecting accidents that must be adequately represented in the training data	AUC (UCSD Ped2 dataset): 96.7% AUC (Avenue dataset): 87.1% AUC (ShanghaiTech): 73.7%
[20]	CNN	Develop a residual spatiotemporal autoencoder for anomaly detection in surveillance videos, leveraging normality modelling to identify irregularities as deviations from standard patterns	The method uses normality modelling and reconstruction loss to identify abnormal spatiotemporal events in surveillance videos accurately	It can be challenging in complex and dynamic surveillance environments, potentially leading to false positives or negatives	UCSD Ped1 dataset: EER=8.1 AUC=93.9 Accuracy=90.3 UCSD Ped2 dataset: EER=6.1 AUC=97.3 Accuracy=95.4
[21]	Residual spatiotemporal autoencoder	To enhance anomaly detection by integrating reconstruction and future frame prediction models, addressing limitations in existing methods	Captures spatial features at various scales, enabling the model to effectively detect anomalies by considering objects of different sizes and complexities in surveillance videos	The increased model complexity could lead to overfitting	AUC (Avenue dataset): 0.82 AUC (LV dataset): 0.63

Continuation of the Table 1

No.	Technique	Objectives	Advantages	Limitation/ future scope	Result
[22]	Deep convolutional neural network-based encoder and a multi-stage channel attention-based decoder	Aims to develop an advanced anomaly detection system for video surveillance by effectively integrating spatial and temporal information	The proposed method effectively captures spatial and temporal features, enhancing anomaly detection accuracy in surveillance videos	It may lead to increased computational complexity, requiring significant resources for training and inference	UCSDped1 Accuracy:93 False alarm rate: 0.08 UCSDped2 Accuracy:97.0 False alarm rate:0.06 CUHK Avenue dataset Accuracy:97.0 False alarm rate: 0.04 UCF-Crime dataset Accuracy:98.0 False alarm rate:0.03
[23]	Convolutional Neural Network (CNN) and two popular handcrafted methods (Histogram of Oriented Gradient (HOG) and HOG3D). One Class Support Vector Machine (OCSVM)	To address the limitations of existing stationary camera surveillance systems in anomaly detection by proposing new techniques suitable for Unmanned Aerial Vehicle (UAV)-based surveillance missions	Enhancing anomaly detection by capturing comprehensive surveillance footage from various angles and viewpoints	Utilising multiple feature extraction methods increases computational complexity and resource requirements	UCSDped1 AUC:83.8 EER:22.2 UCSDped2 AUC:97.6 EER:6.6 Avenue dataset AUC:89.0 EER:18.1
[24]	3D spatiotemporal autoencoder	To enhance automatic surveillance of human activities by addressing the challenges posed by complex real-time scenarios, such as camera movements, cluttered backgrounds, and occlusion	The method captures high-level semantic information and fine-grained details, providing a more comprehensive representation of surveillance video data	increases the computational complexity and resource requirements of the proposed methods	AUC raises 2.0%, 1.2%, and 1.6% for UCSD Ped1, UCSD Ped2, and CUHK Avenue datasets compared with it
[25]	attention-based residual autoencoder	It efficiently utilises spatial and temporal information by adopting both spatial and temporal branches in a single network	The model exceeds the state-of-the-art results on three standard benchmark datasets, even without an optical flow detector	It may be generalised to 3D data for real-world engineering applications	This model achieved 97.4% for UCSD Ped2, 86.7% for CUHK Avenue, and 73.6% for the ShanghaiTech dataset in terms of AUC

Continuation of the Table 1

No.	Technique	Objectives	Advantages	Limitation/ future scope	Result
[26]	OCSVM, HOG, HOG3D, CNN	The goal is when a mobile camera records videos for a surveillance mission with the assistance of a UAV	The potential of UAVs to provide an original aerial perspective is one of its primary advantages	The upper layers of the pre-trained CNN can also be adjusted using transfer learning to better match the target problem in future	HOG model Recall:1 Precision:0.7070 F1 score:0.8284 Accuracy:78.97 PCA-HOG model Recall:1 Precision:0.7073 F1 score:0.8286 Accuracy:79.00 HOG3D model Recall:1 Precision:0.8421 F1 score:0.9143 Accuracy:90.13 GoogleNet model Recall:1 Precision:0.8837 F1 score:0.9383 Accuracy:93.57
[27]	Hybrid multi-view representati on learning	This model uses handcrafted spatiotemporal autocorrelation of gradient, and raw video segments are taken as input for learning the regular patterns in surveillance videos	This model can be accomplished by extracting features from various representations (views) of the raw input data	In the future, anomalies' dependency will be captured in context, leading to better discrimination and overall performance	Avenue dataset Accuracy:82.4 AUC:0.83 LV dataset: Accuracy:64.9 AUC:0.60 BEHAVE dataset Accuracy:80.05 AUC:0.81

Nonetheless, overfitting could happen, and the model depends heavily on the completeness of the training set. The suggested approach is computationally costly, mainly when used in high-resolution videos.

3. Proposed work. Video anomaly detection and segmentation in smart cities is a crucial computer vision problem for smart surveillance and public safety. However, existing research faces challenges such as limited scalability, difficulty detecting complex anomalies, and potential overfitting. The model is computationally demanding and relies on training data accuracy, leading to low efficiency, accuracy, and overfitting in identifying and segmenting video anomalies

The proposed approach overcomes the problem of overfitting in anomaly detection within video frames by introducing a novel I-UNET. Overfitting happens when a model becomes too specialised to the training data, resulting in decreased efficiency and accuracy when applied to fresh,

previously unseen data. The I-UNET is meant to detect irregularities in video frames, which improves detection accuracy. Video frames frequently contain a significant amount of noise, which can impede the accurate detection of anomalies. To overcome this issue, a Weiner filter is applied to the frames during preprocessing to reduce noise. This step seeks to improve the input data quality and the overall presentation of the anomaly detection technique. During anomaly detection, the proposed approach considers spatial and temporal information. To accomplish this, a Convolutional Long Short-Term Memory (ConvLSTM) is added to the model. The ConvLSTM allows the model to consider both the spatial properties of the video frames and the temporal dependencies between consecutive frames, resulting in a more complete comprehension of the information. The suggested model uses the cascade sliding window technique (CSWT) to produce an anomaly score during the post-processing stage. The CSWT analyses the video frames and assigns a score reflecting the chance of an abnormality. This anomaly score was utilised to determine whether a specific frame contains an anomaly or is within the normal range. The suggested diagram's overall architecture is depicted in Figure 1 below.

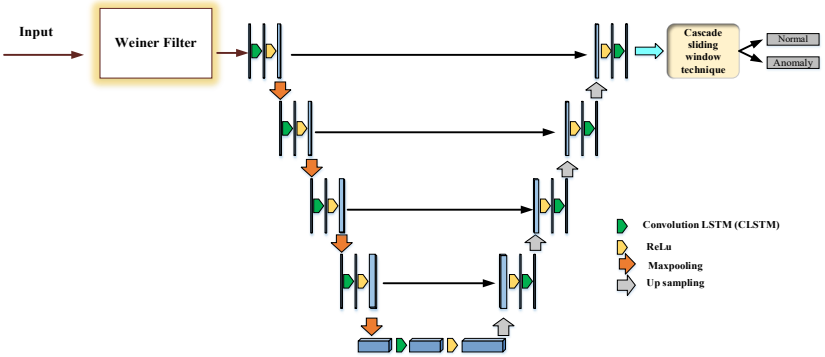

Fig. 1. Architecture diagram of the proposed model

3.1. Weiner Filter for pre-processing. The Weiner filter is used to reduce noise from images [28]. Weiner filters use Linear Time-Invariant (LTI) filtering of an observed noisy process with known stationary signal, noise spectra, and additive noise to estimate a desired or target random process. The Weiner filter decreases the mean square error between the estimated and calculated random processes. Using a related signal as an input and filtering it to get the approximation as an output, the Weiner filter computes a statistical estimate of an unknown signal. Stated differently, the Weiner filter is an adaptive filter that determines the neighbourhood's mean

and variance before applying a lower level of smoothing when the variation is significant and a higher level of smoothing when it is negligible.

The filter reduces the error between the expected and original signals. The error measure, given an original image and a processed image in Figure 7, represents the original and pre-processed images with a wiener filter.

The filter lessens the variation between the original and approximated signals. The error measure for an original image f and a processed image $\hat{f}$ is as follows in equation (1):

$$e^2 = E\{(f - \hat{f})^2\}, \quad (1)$$

where E{.} is the argument's predictable value, generating an approximated image boils down to locating the quadratic error function's minimum. The frequency domain is used to accomplish this, and the following presumptions are made: the image and noise have a zero mean, the noise and image are uncorrelated, and a linear function reduces the intensity levels in the expected picture. Depending on these circumstances, the error function's minimum is provided in equation (2):

$$\hat{F}(u,v) = \left[\frac{H^*(u,v)S_f(u,v)}{S_f(u,v)|H(u,v)|^2 + S_f(u,v)}\right] G(u,v), \quad (2)$$

where $\hat{F}(u,v)$ represent the predictable image in the frequency domain, $H(u,v)$ denote the transform of the degradation function, $G(u,v)$ denote the transform of the degraded image, $H^*(u,v)$ denote the complex conjugate of $H(\mathrm{u},\mathrm{v})$ and $S_f(u,v)$=$|\mathrm{F}(\mathrm{u},\mathrm{v})|^2$ is the power spectrum of the non-degraded image. The magnitude of the complex value squared represents the result of multiplying a complex value by its conjugate, according to the filter's general principle. Consequently, in equation (3):

$$\hat{F}(u,v) = \left[\frac{1}{H(u,v)} \frac{|H(u,v)|^2}{|H(u,v)|^2 + S_\eta(u,v)/S_f(u,v)}\right] \quad G(u,v), \quad (3)$$

where $S_\eta(\mathrm{u},\mathrm{v}) = |N(u,v)|^2$ represent the power spectrum of noise. The term $S_\eta(u,v)/S_f(u,v)$ is substituted by a constant K due to the rarity of knowing the non-degraded image's power spectrum.

The Weiner filter can correct digital image processing noise caused by continuous power additive noise. Thus, the neighbourhood size and noise power are the parameters of the Weiner filter.

Figure 2 depicts both the original and pre-processed images. The Weiner filter method is used in pre-processing to eliminate noise from the frame.

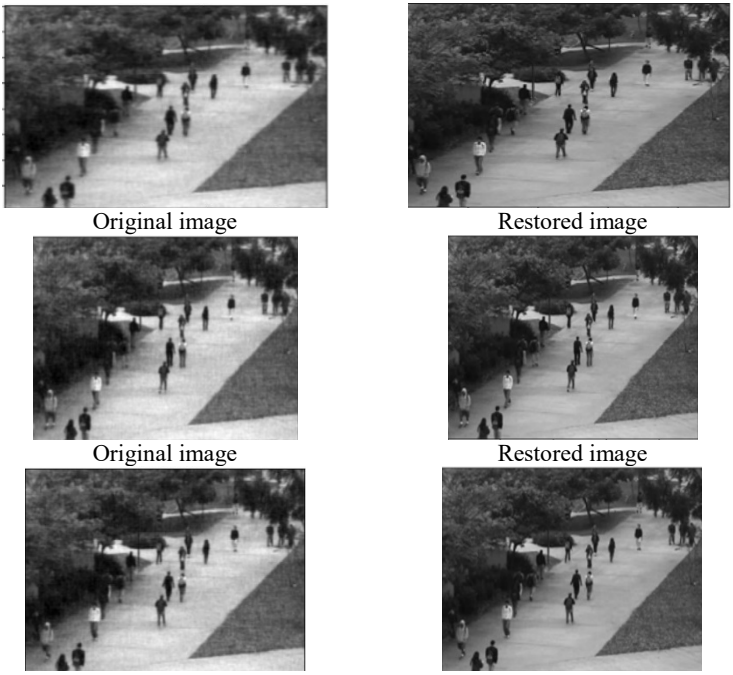

Fig. 2. Original and Pre-processed images with Weiner filter

3.2. Improved UNET for feature segmentation

3.2.1. U-NET. The U-Net architecture is a popular and useful paradigm for segmenting video images. It is U-shaped, symmetrically decoder-path and encoder-path [29]. Because of its U-shaped design, the model can record local information and information about the larger surroundings. In the encoder approach, convolutional and pooling layers are utilised to gradually downscale the input image, which aids in extracting high-level features and collecting contextual data. Each down-sampling step reduces the spatial size of the feature maps while improving their depth. Information is transferred from the encoder to the decoder using skip connections. These links connect the essential layers between the encoder and decoder routes. The skip connections allow the decoder to access high-resolution information from the encoder while acting as a gradient flow

shortcut during training. The decoder path performs feature map up-sampling using deconvolutional layers or up-sampling followed by convolutional layers. This method retains the contextual information the encoder learned as the feature maps' spatial resolution is progressively restored. Skip connections enhance segmentation results by merging feature maps from the encoder and decoder.

3.2.2. I-UNET. When renovating from standard convolution layers to ConvLSTM [30] layers inside the U-NET structure, it is vital to account for the inherent spatiotemporal dependencies in video data. In this improved approach, the Encoder with ConvLSTM is used strategically to capture spatial dependencies over the entire video frame. Concurrently, the decoder is upgraded with ConvLSTM layers to manage temporal dependencies during decoding properly. The introduction of ConvLSTM layers allows the model to include spatial properties within individual frames and temporal correlations between subsequent frames. As a result, the last layer of the decoder is modified to provide an anomaly prediction map for each frame. This holistic approach enables the U-Net to more comprehensively understand and leverage the spatiotemporal intricacies inherent in video data, making it well-suited for tasks such as video anomaly detection. The architecture of the Improved U-NET is displayed in Figure 3.

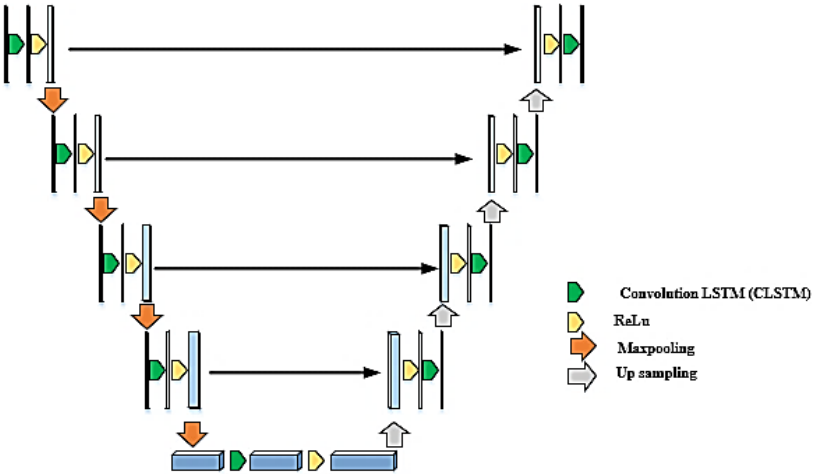

Fig. 3. Improved U-NET Architecture

3.2.2.1. ConvLSTM. Convolutional Long Short-Term Memory, or "ConvLSTM" architecture, combines the best features of Long Short-Term Memory (LSTM) networks with the concepts of Convolutional Neural

Networks (CNNs) to improve processing speed for sequential data, such as video frames. The forget gate is an essential component of classic LSTM networks that helps choose, keep, or discard data from earlier steps. By addressing the vanishing gradient issue, this method helps the network to handle longer sequences and aggregate higher-level information efficiently. Enhancing this concept, the ConvLSTM model substitutes convolutions for conventional matrix operations. The creation of spatial feature maps is enhanced by this adjustment, which lowers the model's parameter count (weights). The ConvLSTM model is especially well-suited for tasks like video frame prediction because it works with convolutions to capture spatial dependencies within the input data effectively. Sequence modelling for video data is greatly advanced by the convolutional procedures combined with LSTM units. The model's capacity to recognise intricate temporal patterns and spatial correlations in video data is improved by this method, which produces a more efficient and spatially aware representation of the input sequences. Convolutions are used instead of matrix operations in ConvLSTM instead of the typical fully connected LSTM (FC-LSTM). ConvLSTM uses convolution for hidden-to-hidden and input-to-hidden connections, which results in better spatial feature maps with less weight requirements. The following equations (4-9), which explain how data moves through the network and how the forget gate, input gate, and output gate are calculated, can be used to sum up the working principles of the ConvLSTM unit:

$$f_t = \sigma(W_f * [h_{t-1}, x_t, C_{t-1}] + b_f, \quad (4)$$

where f_t indicates the forget gate activation vector at time step t, σ is the sigmoid function, the input vector is represented by the variable x_t, the hidden state is represented by h_t, b_f represents the bias vector for the forget gate, W_f represents the weight matrix for the forget gate and the cell state is represented by C_t at times *t*.

$$i_t = \sigma(W_i * [h_{t-1}, x_t, C_{t-1}] + b_i, \quad (5)$$

where i_t indicates the Input gate activation vector at time step t, W_i represents the weight matrix for the input gate, b_i denotes the bias vector for the input gate.

$$\hat{C}_t = tanh(W_C * [h_{t-1}, x_t] + b_C, \quad (6)$$

where $\hat{C}_t$ represents the candidate cell state at time step t, the hyperbolic tangent function is represented by $\tanh$, W_C denotes the weight matrix for the candidate cell state, the bias vector for the candidate cell state is denoted by b_C.

$$C_t = f_t \otimes C_{t-1} + i_t \otimes \hat{C}_t, \tag{7}$$

where C_t is the new cell state at time step t, f_t represents the forget gate activation vector at time step t, the previous cell state at time step $t-1$ is indicated by C_{t-1}, i_t denotes the input gate activation vector at time step t. The sign $\otimes$ indicates the Hadamard product.

$$o_t = \sigma(W_o * [h_{t-1}, x_t, C_{t-1}] + b_0), \tag{8}$$

where o_t is the output gate activation vector at time step t, the weight matrix for the output gate is represented by W_o, b_0 is the bias vector for the output gate.

$$h_t = o_t \otimes tanh(C_t), \tag{9}$$

where h_t is the hidden state at time step t, the output gate activation vector at time step t is indicated by o_t, $tanh$ represents the hyperbolic tangent function, C_t is the cell state at time step t.

Convolutional filters replace the set of weights for each link in the input (the symbol * denotes a convolution operation). Because of its capacity to convey spatial attributes temporally through each ConvLSTM state, ConvLSTM works better with images than fully connected LSTM.

3.3. Cascade Sliding Window Technique for classification. Cascade sliding window is a technique used in object detection that can also be applied to video anomaly detection. Finding strange occurrences or behaviours in video frames is the goal of video anomaly detection. The cascade sliding window technique employs a multi-stage process to look for objects or abnormalities inside each frame at various scales and places.

The cascade sliding window method is a methodical way to get an anomaly score in the context of video analysis. This method is intended to determine whether or not a specific frame in a video sequence contains an anomaly. The process entails scanning each frame at different positions and scales using a sliding window. A classifier is used at each stage of this procedure to evaluate the information within the sliding window, discriminating between typical and abnormal patterns. The cascade

structure incorporates numerous layers of classifiers, each contributing to determining the anomaly score. The first stage usually uses simpler and faster classifiers to quickly exclude non-anomalous regions, decreasing the computational cost for the following stages. More advanced classifiers are used to refine the anomaly detection as the process progresses along the cascade. This methodology's anomaly score is a critical output as a quantitative measure to infer the likelihood of an anomaly's presence in a given frame. By defining appropriate criteria for anomaly scores, it is feasible to make educated decisions about whether a frame exhibits anomalous behaviour. This allows for effective identification and analysis of odd events in the video frame. The cascade sliding window framework contributes to accuracy and processing efficiency in the anomaly identification process, making this method an organised and efficient way of distinguishing anomalies in video data. In Algorithm 1, the Cascade Sliding Window technique is illustrated.

The frame size R represents a frame's height and width in Algorithm 1, and the window size is indicated by $\hat{R}$. The technique generates an image I by squaring the distinction between an actual frame and one that has been predicted. This method is chosen because it allocates higher values to pixels in the abnormal region of I rather than making use of the fundamental distinction between an expected and actual frame. Compared to the distinction between a genuine frame and a prediction frame, it is more successful at identifying aberrant frames. The average of an anomalous frame with a small abnormal section may resemble or be less than that of a regular frame if the anomalous frame employs the difference between an actual frame and a forecasted frame. The window on $\boldsymbol{I}$ starts to move to the right as much as $\widehat{\boldsymbol{R}}$ at positions x=0 and y=0. The window travels to the left side of I and up to $\widehat{\boldsymbol{R}}$ if it reaches the right side of $\boldsymbol{I}$. The window will then start to slide to the right by the same amount as $\widehat{\boldsymbol{R}}$ as you lower $\widehat{\boldsymbol{R}}$ as the window size v decreases. Continue in the same manner until the window reaches $\boldsymbol{I}$'s upper right corner. In case the window is unable to get either the top or right sides of $\boldsymbol{I}$ due to the remaining space in $\boldsymbol{I}$ being less than the window, it will go to $\boldsymbol{y = R - \widehat{R}}$ for the top side of $\boldsymbol{I}$ and to $\boldsymbol{x = R - \widehat{R}}$ for the right side $\boldsymbol{I}$. It is provided in lines 9 and 18 of Algorithm 1. Determine the average for a frame $\boldsymbol{P_k}$ that corresponds to the moving window. It is comparable to $\boldsymbol{P_k}$, the mean squared error (MSE) between a real frame and a prediction frame. When the window reaches the upper right corner of $\boldsymbol{I}$, take n frames from the front of $\boldsymbol{P_k}$ and sort $\boldsymbol{P_k}$ in increasing order. The anomalous score S is then calculated by averaging the n patches.

Algorithm 1. Cascade Sliding Window

Input: actual present frame $F_{i,j,}$ anticipated present frame $\hat{F}_{i,j,}$ frame size R, window size $\hat{R}$, window decrease size v

Results: score S for anomalies

1 Set up the coordinates x=0 and y=0 and the P_k mean for each frame.

2 $I_{i,j} = (F_{i,j} - \hat{F}_{i,j})^2$ /* image on the square of the difference between $F_{i,j}$ and $\hat{F}_{i,j}$ */

3 While y < R do

4 if $y + \hat{R} \leq R$ then

5 while $x < R$ do

6 if $x + \hat{R} \leq R$ then

7 $P_k = \frac{1}{\hat{R}^2} \sum_{i=x}^{x+\hat{R}} \sum_{j=y}^{y+\hat{R}} I_{i,j}$

8 else

9 $P_k = \frac{1}{\hat{R}^2} \sum_{i=R-\hat{R}}^{R} \sum_{j=y}^{y+\hat{R}} I_{i,j}$

10 end if

11 $x = x + \hat{R}$

12 end while

13 else

14 while x < R do

15 if $x + \hat{R} \leq R$ then

16 $P_k = \frac{1}{\hat{R}^2} \sum_{i=x}^{x+\hat{R}} \sum_{j=R-\hat{R}}^{s} I_{i,j}$

17 else

18 $P_k = \frac{1}{\hat{R}^2} \sum_{i=R-\hat{R}}^{R} \sum_{j=R-\hat{R}}^{R} I_{i,j}$

19 end if

20 $x = x + \hat{R}$

21 end while

22 end if

23 $x = 0$

24 $y = y + \hat{R}$

25 $\hat{R} = \hat{R} - v$

26 end while

27 arrange (P_k) in ascending order

28 $S = \frac{1}{n} \sum_{i=1}^{n} P_i$

29 Return S

The cascade sliding window's decreasing window size, shown by v in line 25 of Algorithm 1, is essential. It assumes that objects get smaller and farther away from the items under video monitoring. Contrasting it with another approach that uses the MSE between an actual and a forecasted present frame shows how well the cascade sliding window technique performs.

3.3.1. Anomaly detection. Using the cascade sliding window, compute the anomaly score for every frame. The anomaly score of the I-UNET model output frame runs from 0 to $colordepth^2$, and the colour depth in the framework is 256. Consequently, the range of the anomaly score generated by the cascade sliding window is 0 to 65536. This range is too broad to establish the anomalous frame threshold. As a result, the anomaly score must be normalised. The anomaly score between 0 and 1 is normalised using the following formulas equation (10):

$$S'(t) = 1 - \frac{S(t) - min_t S(t)}{max_t S(t) - min_t S(t)}, \quad (10)$$

S'(t) is the normalised anomaly score, and S(t) is the anomaly score for frame t. Videos have two anomaly scores: $min_t S(t)$ and $max_t S(t)$ for maximum and minimum anomaly scores, respectively. However, while obtaining a new frame in the actual world, the $max_t S(t)$ and $min_t S(t)$ values could change. It results in a recalculation of the threshold and the acquired anomaly scores. Normalising the anomaly score from 0 to 1 will solve this problem using equation (11).

$$S'(t) = \frac{S(t)}{colordepth^2}, \quad (11)$$

where colour depth refers to the colour depth of the R-Net model's output frame, even if the maximum and minimum anomaly scores are altered, this normalisation does not necessitate recalculating the threshold and anomaly scores.

4. Results and Discussions. The model was trained using thousands of video frames from a video dataset, and a thorough testing procedure was conducted to ascertain the efficacy of the novel approach.

4.1. Datasets. UCSD Ped2. A stationary camera positioned at a height that provided a view of pedestrian routes was utilised to collect the UCSD anomaly detection dataset. There was a range in the walkways' population density from sparse to congested. In its original state, the video only features pedestrians. Either non-pedestrian objects moving through the walkways or abnormal pedestrian movement patterns cause abnormal events. Individuals, skateboarders, bicyclists, and small carts are frequently observed strolling down a path or in the adjacent grass. There were also a few reported instances of wheelchair-using individuals. Since none of the anomalies were created to compile the dataset, they are all-natural. Two subgroups were created from the data, each representing a distinct scene. Each sequence's video clip was segmented into segments of

about 200 frames. Scenes with pedestrian movement parallel to the camera plane are categorised as Peds2, which includes twelve testing and sixteen training video examples. The ground truth annotation for every clip contains a binary flag for every frame that indicates whether an abnormality is present in that particular frame. Furthermore, pixel-level binary masks, including anomaly zones, are manually created for a subset of 12 clips for Peds2. This is meant to make it possible to assess how well algorithms perform in terms of their capacity to localise anomalies. Here, the data is split into 60% for training and 40 % for testing. (https://www.kaggle.com/datasets/karthiknm1/ucsd-anomaly-detection-dataset).

4.2. Performance Evaluation. This segment demonstrates how the suggested I-UNET approach may successfully classify video anomaly image frames designated as normal or anomalous. The suggested model's accuracy and loss analyses are displayed during the training process across the 8th epochs. The proposed model improves accuracy while losing utility. It demonstrates that the proposed model converges very quickly.

The I-UNET method trains a model across eight epochs with the Adam optimizer, consistently obtaining 99% accuracy, as shown in Figure 4. This high degree of accuracy shows how reliable and efficient the I-UNET technique is in correctly identifying and analysing data patterns. The model's high accuracy indicates its potential for various applications where precision is crucial.

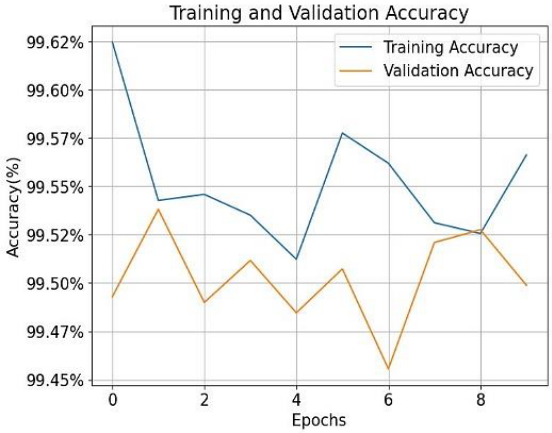

Fig. 4. Training and validation Accuracy

Employing the Adam optimizer, the training and validation AUC of the I-UNET technique was monitored throughout eight epochs. With an astounding 90.8% AUC, the model successfully identified normal and anomalous instances in the video data. This high AUC score demonstrated the strength and adaptability of the technique over all epochs, as shown in Figure 5. The Adam optimiser's practical training probably aided the model's convergence towards an optimal solution, highlighting the dependability and capacity to generalise the model and increasing the likelihood of precise anomaly discovery.

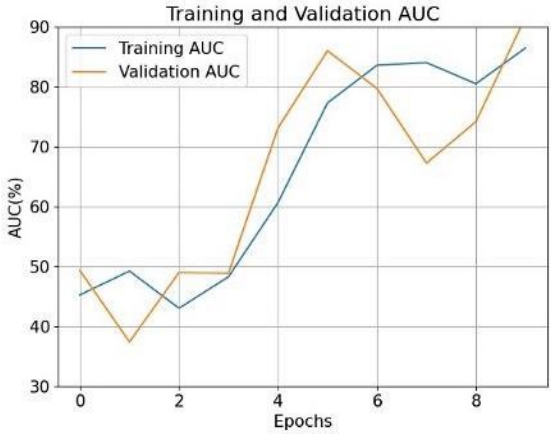

Fig. 5. Training and validation AUC

The Adam optimiser's deep learning optimisation method was used to train the I-UNET technique, and eight epochs were used to track its validation and training loss metrics, as shown in Figure 6. From the first to the eighth epoch, the model continuously reduced loss levels, demonstrating an improved capacity to minimise discrepancies between expected outputs and actual targets. With an average loss level of about 1.53%, the predictions were accurate. This effective loss reduction was probably made possible by the Adam optimiser's adaptive learning rates and parameter updates, which allowed the model to converge to the best possible solution.

Fig. 6. Training and validation loss

4.3. Performance Metrics. The performance metrics include a variety of critical indications for assessing a model's success. Accuracy, AUC, and EER are among these measurements in equation (12-14).

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}. \tag{12}$$

This metric provides an overall measure of the model's prediction accuracy, representing the ratio of successfully detected instances to total occurrences.

$$AUC = \sum \frac{(TPR[i] + TPR[i+1])}{2} * FPR[i+1] - FPR[i]). \tag{13}$$

AUC and the Receiver Operating Characteristics (ROC) curve are often associated. It shows the area under the curve, showing how the true positive and false positive rates are traded off at various classification levels. A higher AUC suggests better model performance overall. The EER metric is considered as follows:

$$EER = \frac{FP+FN}{TC}, \tag{14}$$

where TP is True Positive, TN is True Negative, FP denoted as False Positive, FN indicated as False Negative, TPR is True Positive rate, FPR represented as False positive rate.

TC displays the overall frames from the test dataset. As a result, a model that performs better has a greater AUC and a lower EER since it can discriminate more effectively. The following graph depicts the overall efficacy of video anomaly detection:

The existing methods such as the Two-stream fusion algorithm [31], AlexNet-based model [32], and Convolutional autoencoder [33] are compared with the proposed technique for analyzing the ROC curve. The Two-stream Fusion Algorithm (0.979) achieves good performance by combining temporal and spatial information. The ROC of AlexNet-based Model Convolutional Autoencoder are 0.970 and 0.9382. The proposed I-UNET is a significantly better U-Net model than the one that has been suggested, achieving flawless performance of ROC (1.00). Table 2 represents the performance of the ROC curve of the proposed model with existing works.

Table 2. Performance of ROC curve

Model	ROC curve
Two-stream fusion algorithm [31]	0.979
AlexNet-based model [32]	0.970
Convolutional autoencoder [33]	0.9382
I-UNET (Proposed)	1.00

Multiple performance measures are used to assess a model's effectiveness. Predicting accuracy requires understanding accuracy, which is defined as the proportion of successfully classified occurrences to all occurrences. AUC-ROC is a crucial metric for binary classification issues since it demonstrates the trade-off between true and false positive rates at various thresholds. The Equal Error Rate (EER) makes it easier to assess the model's performance objectively. This describes the point on the ROC curve when the rates of false rejection and mistaken acceptance are equal. These metrics evaluate a technique’s capacity to generate precise classifications over an extensive array of performance standards. Figure 7 represents the ROC curve with an actual positive rate and a false positive rate.

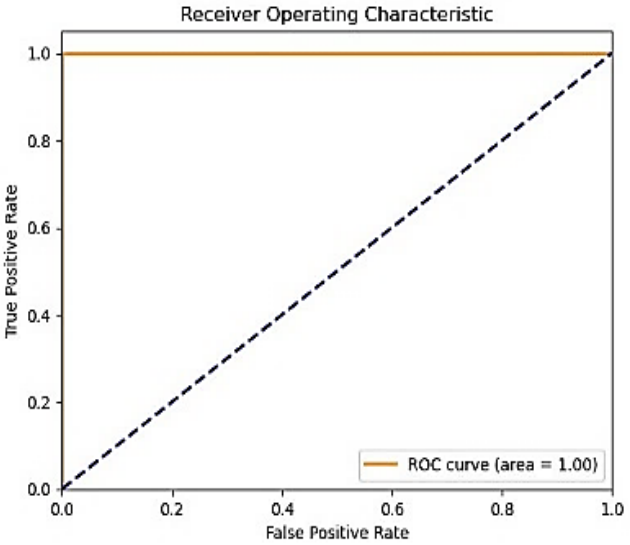

Fig. 7. ROC curve

4.4. Results Obtained. Figure 8 represents the final classification of video anomaly detection.

By adding anomaly scores, the I-UNET displays effective anomaly prediction. This model strategically integrates spatial and temporal information by employing a unique design that includes ConvLSTM layers. This new method improves the model's accuracy using encoder and decoder components that extract spatial and temporal data using ConvLSTM. Incorporating ConvLSTM enables the model to record detailed patterns across time, allowing it to recognise anomalies in video sequences more accurately. The cascade sliding window technique (CSWT) detects anomalies by calculating an anomaly score. This technique is critical in determining the existence or absence of abnormalities in each frame. It successfully analyses the successive frames, using a cascading sliding window technique to compute anomaly scores, thereby providing a dependable mechanism for identifying anomalies in video data.

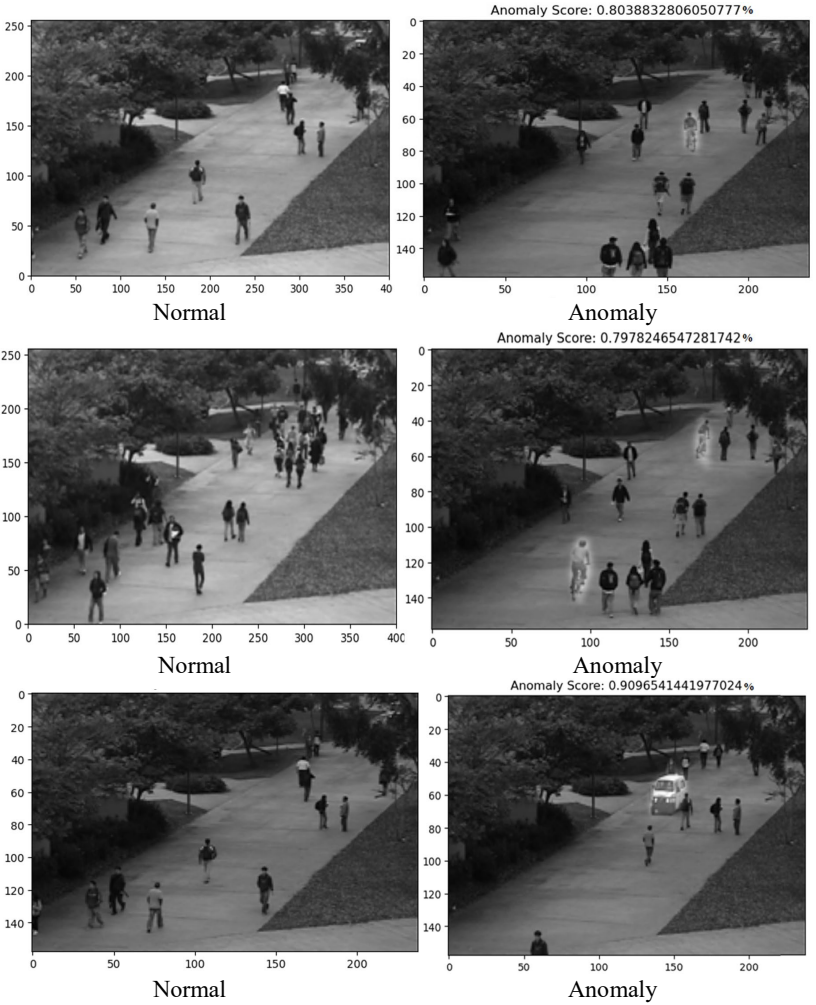

Fig. 8. Prediction result of image normal and anomaly image

4.5. Comparative Analysis. This section illustrates the recommended methodology's superior performance, with the innovative I-UNET serving as the fundamental framework. Compared to other models with fewer parameters, the unique I-UNET regularly outperforms, giving comparable or even greater performance. This highlights the efficiency of the proposed approach, establishing the I-UNET as a reliable and effective model for the given task. It combined the results of this model with those of

current research studies that specified strategies such as Approximated optical flow monitors algorithm (AOFM) [18], Mixture of Dynamic Texture (MDT) [18], Social Force SF) [18], Social Force and Mixture of Probabilistic Component Analyser (SF+MPPCA) [18], Hybrid Ensample Recurrent Reinforcement Model (HERR) [18], Mixture of probabilistic Principal component Analysis (MPPCA) [30], Convolution Auto-encoder (Conv-AE) [30], Convolution-Long short term memory-Auto-encoder (Conv-LSTM-AE) [30] unmasking [30].

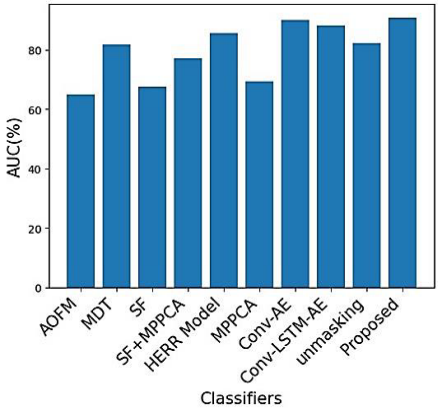

Fig. 9. Comparison of AUC

The research on the performance of different anomaly detection techniques emphasises the Area Under the Curve (AUC) measure. These AUC comparisons are shown in a figure, most likely Figure 9. The article emphasises how much better an innovative method known as I-UNET is at raising anomaly detection's AUC. The analysis shows that the accuracy of the I-UNET technique is higher than that of numerous conventional methods, such as AOFM, MDT, SF, SF+MPPCA, the HERR, MPPCA, Conv-AE, Conv-LSTM-AE, and unmasking model. According to these methodologies, the improvement percentages are as follows: 63%, 85%, 63%, 71%, 89.98%, 69.3%, 90.0%, 88.1%, and 82.2%. In terms of accuracy, conventional approaches continue to outperform the novel approach despite the remarkable performance increases attained by the I-UNET methodology. Although impressive, the AUC of 90.8% produced by the I-UNET methodology is not as high as that of existing methods.

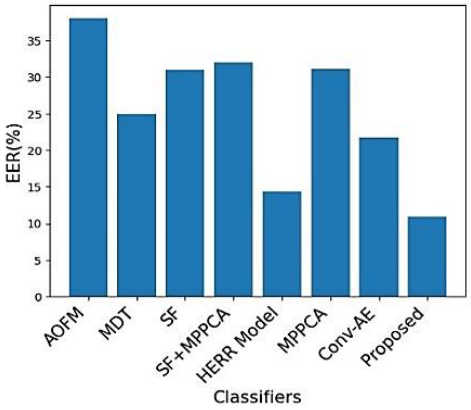

Fig. 10. Comparison of EER

The Equal Error Rate (EER) comparison of several anomaly detection techniques is shown in Figure 10. It demonstrates how well a brand-new strategy known as I-UNET works to lower anomaly detection's EER. The results show that the EER of the I-UNET approach is lower than that of numerous traditional approaches, such as AOFM, MDT, SF, SF+MPPCA, HERR, MPPCA, and Conv-AE model. The corresponding improvement percentages over these strategies are 42%, 25%, 31%, 32%, 14.33%, 31.1%, and 21.7%. Significantly, with an EER of 10.9%, the innovative I-UNET methodology outperforms conventional techniques in accuracy. This suggests a notable decrease in mistake rates in contrast to traditional methods. In addition, Table 3 offers a thorough summary of both AUC and EER values, enabling a more in-depth analysis of how well various approaches perform across these parameters.

Table 3. Comparisons of AUC and EER

Method	AUC(%)	EER(%)
AOFM [18]	63	42
MDT [18]	85	25
SF [18]	63	31
SF+MPPCA [18]	71	32
HERR model [18]	89.8	14.33
MPPCA [30]	69.3	31.1
Conv-AE [30]	90.0	21.7
Conv-LSTM-AE [30]	88.1	-
Unmasking [30]	82.2	-
I-UNET (Proposed)	90.8	10.9%

When assessing the effectiveness of a classification system, accuracy and Equal Error Rate (EER) are essential metrics, mainly when there are uneven class distributions or different sorts of errors. Although accuracy gauges forecasts' accuracy, it might give a partial picture in skewed conditions. By balancing the rates of false positives and false negatives, EER measures the model's performance. A lower EER indicates equal error costs and balanced class distribution produces better performance and positive correlations.

Compared to existing works such as AOFM, MDT, SF, SF+MPPCA, HERR, MPPCA, Conv-AE, Conv-LSTM-AE, and unmasking model, the proposed model attains high AUC and low EER. In this model, three datasets are used: UCF, UCSDped1, and UCSped2 datasets. Among these datasets, the UCSDped2 attains the highest accuracy and AUC but has a lower EER than other datasets. The performance analysis of metrics such as accuracy, AUC and EER for different datasets is represented in Table 4.

Table 4. Performance analysis of different datasets for the proposed model

Datasets	Accuracy (%)	AUC (%)	EER (%)
UCF	92	82.5	23.5
UCSDped1	96.5	87.8	18.6
UCSDped2	99	90.8	10.9

5. Conclusion. The researchers have focused on developing algorithms for reconstruction and prediction to tackle the complex problem of video anomaly identification in computer vision. Existing techniques encountered difficulties in unsupervised anomaly recognition due to resolution constraints and a lack of labelled anomalies, resulting in lesser accuracy. This paper presents a unique approach designated as Improved UNET (I-UNET) to reduce the risk of overfitting by addressing the need for sophisticated models capable of managing fine-grained data in video anomalies. A Weiner filter is used in the preprocessing stage to remove noise from video frames. The proposed architecture integrates spatial and temporal information in both the encoder and decoder sections by employing a ConvLSTM layer, ensuring anomaly detection precision. The Cascade Sliding Window Technique (CSWT) is used to calculate anomaly scores and assess the presence of anomaly frames to improve post-processing. The results show that the proposed network successfully segments anomalies, resulting in significantly better performance metrics such as Accuracy of 99%, AUC of 90.8%, and EER of 10.9%. This demonstrates the efficacy of the suggested methodology in detecting high-precision video anomalies, a significant advancement in the field. Fine-

tuning and adaptation are crucial in tailoring pre-trained models to specific anomaly detection tasks. This process involves optimising hyperparameters, implementing regularisation techniques, and devising effective adaptation strategies. The future potential of utilising transfer learning techniques with pre-trained models for anomaly detection on comparable tasks or datasets warrants examination. This approach holds promise for identifying anomalies in scenarios where labelled data is limited or unavailable.

References

1. Ramachandra B., Jones M.J., Vatsavai R.R. A survey of single-scen4e video anomaly detection. IEEE transactions on pattern analysis and machine intelligence. 2020. vol. 44(5). pp. 2293–2312.
2. Nayak R., Pati U.C., Das S.K. A comprehensive review on deep learning-based methods for video anomaly detection. Image and Vision Computing. 2021. vol. 106(6). DOI: 10.1016/j.imavis.2020.104078.
3. Raja R., Sharma P.C., Mahmood M.R., Saini D.K. Analysis of anomaly detection in surveillance video: recent trends and future vision. Multimedia Tools and Applications. 2023. vol. 82(8). pp. 12635–12651.
4. Erhan L., Ndubuaku M., Di Mauro M., Song W., Chen M., Fortino G., Bagdasar O., Liotta A. Smart anomaly detection in sensor systems: A multi-perspective review. Information Fusion. 2021. vol. 67. pp. 64–79.
5. Pang G., Shen C., Cao L., Hengel A.V.D. Deep learning for anomaly detection: A review. ACM computing surveys (CSUR). 2021. vol. 54(2). pp. 1–38.
6. Rezaee K., Rezakhani S.M., Khosravi M.R., Moghimi M.K. A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance. Personal and Ubiquitous Computing. 2024. vol. 28(1). pp. 135–151.
7. Ackerson J.M., Dave R., Seliya N. Applications of recurrent neural network for biometric authentication & anomaly detection. Information. 2021. vol. 12(7). DOI: 10.3390/info12070272.
8. Şengönül E., Samet R., Abu Al-Haija Q., Alqahtani A., Alturki B., Alsulami A.A. An Analysis of Artificial Intelligence Techniques in Surveillance Video Anomaly Detection: A Comprehensive Survey. Applied Sciences. 2023. vol. 13(8). DOI: 10.3390/app13084956.
9. da Costa K.A., Papa J.P., Passos L.A., Colombo D., Del Ser J., Muhammad K., de Albuquerque V.H.C. A critical literature survey and prospects on tampering and anomaly detection in image data. Applied Soft Computing. 2020. vol. 97. DOI: 10.1016/j.asoc.2020.106727.
10. Jebur S.A., Hussein K.A., Hoomod H.K., Alzubaidi L., Santamaría J. Review on deep learning approaches for anomaly event detection in video surveillance. Electronics. 2022. vol. 12(1). DOI: 10.3390/electronics12010029.
11. Habeeb R.A.A., Nasaruddin F., Gani A., Hashem I.A.T., Ahmed E., Imran M. Real-time big data processing for anomaly detection: A survey. International Journal of Information Management. 2019. vol. 45. pp. 289–307.
12. Arshad K., Ali R.F., Muneer A., Aziz I.A., Naseer S., Khan N.S., Taib S.M. Deep Reinforcement Learning for Anomaly Detection: A Systematic Review. IEEE Access. 2022. vol. 10. pp. 124017–124035.
13. Berroukham A., Housni K., Lahraichi M., Boulfrifi I. Deep learning-based methods for anomaly detection in video surveillance: a review. Bulletin of Electrical Engineering and Informatics. 2023. vol. 12(1). pp. 314–327.

14. Kiran B.R., Thomas D.M., Parakkal R. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. Journal of Imaging. 2018. vol. 4(2). DOI: 10.3390/jimaging4020036.
15. Musa A.A., Hussaini A., Liao W., Liang F., Yu W. Deep Neural Networks for Spatial-Temporal Cyber-Physical Systems: A Survey. Future Internet. 2023. vol. 15(6). DOI: 10.3390/fi15060199.
16. Albuquerque Filho J.E., Brandão L.C., Fernandes B.J., Maciel A.M. A review of neural networks for anomaly detection. IEEE Access. 2022. vol. 10(5). pp. 112342–112367.
17. Borowiec M.L., Dikow R.B., Frandsen P.B., McKeeken A., Valentini G., White A.E. Deep learning as a tool for ecology and evolution. Methods in Ecology and Evolution. 2022. vol. 13(8). pp. 1640–1660.
18. Amudha L., Pushpa Lakshmi R. Performance Analysis of Hybrid RR Algorithm for Anomaly Detection in Streaming Data. Computer Systems Science & Engineering. 2023. vol. 45(3). pp. 2299–2312.
19. Chang Y., Tu Z., Xie W., Luo B., Zhang S., Sui H., Yuan J. Video anomaly detection with spatio-temporal dissociation. Pattern Recognition. 2022. vol. 122. DOI: 10.1016/j.patcog.2021.108213.
20. Rezaei F., Yazdi M. A new semantic and statistical distance-based anomaly detection in crowd video surveillance. Wireless Communications and Mobile Computing. 2021. vol. 2021. DOI: 10.1155/2021/5513582.
21. Deepak K., Chandrakala S., Mohan C.K. Residual spatiotemporal autoencoder for unsupervised video anomaly detection. Signal, Image and Video Processing. 2021. vol. 15(1). pp. 215–222.
22. Ul Amin S., Ullah M., Sajjad M., Cheikh F.A., Hijji M., Hijji A., Muhammad K. EADN: An efficient deep learning model for anomaly detection in videos. Mathematics. 2022. vol. 10(9). DOI: 10.3390/math10091555.
23. Taghinezhad N., Yazdi M. A new unsupervised video anomaly detection using multi-scale feature memorization and multipath temporal information prediction. IEEE Access. 2023. vol. 11. pp. 9295–9310.
24. Liu T., Zhang C., Niu X., Wang L. Spatio-temporal prediction and reconstruction network for video anomaly detection. Plos one. 2022. vol. 17(5). DOI: 10.1371/journal.pone.0265564.
25. Le V.T., Kim Y.G. Attention-based residual autoencoder for video anomaly detection. Applied Intelligence. 2023. vol. 53(3). pp. 3240–3254.
26. Chriki A., Touati H., Snoussi H., Kamoun F. Deep learning and handcrafted features for one-class anomaly detection in UAV video. Multimedia Tools and Applications. 2021. vol. 80. pp. 2599–2620.
27. Deepak K., Srivathsan G., Roshan S., Chandrakala S. Deep multi-view representation learning for video anomaly detection using spatiotemporal autoencoders. Circuits, Systems, and Signal Processing. 2021. vol. 40(3). pp. 1333–1349.
28. dos Santos J.C.M., Carrijo G.A., de Fátima dos Santos Cardoso C., Ferreira J.C., Sousa P.M., Patrocínio A.C. Fundus image quality enhancement for blood vessel detection via a neural network using CLAHE and Wiener filter. Research on Biomedical Engineering. 2020. vol. 36. pp. 107–119.
29. Sharma N., Gupta S., Koundal D., Alyami S., Alshahrani H., Asiri Y., Shaikh A. U-Net model with transfer learning model as a backbone for segmentation of gastrointestinal tract. Bioengineering. 2023. vol. 10(1). DOI: 10.3390/bioengineering10010119.
30. Cai Y., Liu J., Guo Y., Hu S., Lang S. Video anomaly detection with multi-scale feature and temporal information fusion. Neurocomputing. 2021. vol. 423. pp. 264–273.

31. Yang Y., Fu Z., Naqvi S.M. Abnormal event detection for video surveillance using an enhanced two-stream fusion method. Neurocomputing. 2023. vol. 553. DOI: 10.1016/j.neucom.2023.126561.
32. Khan A.A., Nauman M.A., Shoaib M., Jahangir R., Alroobaea R., Alsafyani M., Binmahfoudh A., Wechtaisong C. Crowd anomaly detection in video frames using fine-tuned AlexNet Model. Electronics. 2022. vol. 11(19). DOI: 10.3390/electronics11193105.
33. Ali M.M. Real-time video anomaly detection for smart surveillance. IET Image Processing. 2023. vol. 17(5). pp. 1375–1388.

R. Krishnan Sreedevi — Research scholar, Department of Computer Science and Engineering, Avinashilingam Institute for Home Science and Higher Education for Women. Research interests: deep learning, computer vision, machine learning, network security and anomaly detection. The number of publications — 6. 19pheop005@avinuty.ac.in; Tamil Nadu, 641043, Coimbatore, India; office phone: +91(960)535-9348.

Amudha P. — Professor, Department of Computer Science and Engineering, Avinashilingam Institute for Home Science and Higher Education for Women. Research interests: data mining, machine learning, information security. The number of publications — 56. amudha_cse@avinuty.ac.in; Tamil Nadu, 641043, Coimbatore, India; office phone: +91(902)563-6594.

УДК 004.8 **DOI** 10.15622/ia.23.6.12

Ш. Р. Кришнан, П. Амудха

УЛУЧШЕНИЕ ОБНАРУЖЕНИЯ АНОМАЛИЙ НА ВИДЕО С ПОМОЩЬЮ УСОВЕРШЕНСТВОВАННОЙ ТЕХНОЛОГИИ UNET И ТЕХНИКИ КАСКАДНОГО СКОЛЬЗЯЩЕГО ОКНА

Р. Кришнан Ш., Амудха П. **Улучшение обнаружения аномалий на видео с помощью усовершенствованной технологии UNET и техники каскадного скользящего окна.**

Аннотация. Обнаружение аномалий на видео с помощью компьютерного зрения все еще нуждается в совершенствовании, особенно при распознавании изображений с необычными движениями или объектами. Современные подходы в основном сосредоточены на методах реконструкции и прогнозирования, а обнаружение аномалий на видео без наблюдения сталкивается с трудностями из-за отсутствия достаточного количества помеченных аномалий, что снижает точность. В этой статье представлена новая структура под названием усовершенствованная UNET (I-UNET), разработанная для противодействия переобучению путем удовлетворения потребности в сложных моделях, которые могут извлекать малозаметную информацию из аномалий на видео. Видеошум можно устранить путем предварительной обработки кадров фильтром Винера. Более того, система использует сверточные слои долго-кратковременной памяти (ConvLSTM) для плавной интеграции временных и пространственных данных в свои части энкодера и декодера, улучшая точность идентификации аномалий. Последующая обработка осуществляется с использованием техники каскадного скользящего окна (CSWT) для идентификации аномальных кадров и генерации оценок аномалии. По сравнению с базовыми подходами, экспериментальные результаты на наборах данных UCF, UCSDped1 и UCSDped2 демонстрируют заметные улучшения производительности, с точностью 99%, площадью под кривой (AUC) 90,8% и равным уровнем ошибок (EER) 10,9%. Это исследование предоставляет надежную и точную структуру для обнаружения аномалий на видео с наивысшим уровнем точности.

Ключевые слова: обнаружение аномалий, I-UNET, фильтр Винера, ConvLSTM, каскадное скользящее окно, оценка аномалий.

Литература

1. Ramachandra B., Jones M.J., Vatsavai R.R. A survey of single-scen4e video anomaly detection. IEEE transactions on pattern analysis and machine intelligence. 2020. vol. 44(5). pp. 2293–2312.
2. Nayak R., Pati U.C., Das S.K. A comprehensive review on deep learning-based methods for video anomaly detection. Image and Vision Computing. 2021. vol. 106(6). DOI: 10.1016/j.imavis.2020.104078.
3. Raja R., Sharma P.C., Mahmood M.R., Saini D.K. Analysis of anomaly detection in surveillance video: recent trends and future vision. Multimedia Tools and Applications. 2023. vol. 82(8). pp. 12635–12651.
4. Erhan L., Ndubuaku M., Di Mauro M., Song W., Chen M., Fortino G., Bagdasar O., Liotta A. Smart anomaly detection in sensor systems: A multi-perspective review. Information Fusion. 2021. vol. 67. pp. 64–79.
5. Pang G., Shen C., Cao L., Hengel A.V.D. Deep learning for anomaly detection: A review. ACM computing surveys (CSUR). 2021. vol. 54(2). pp. 1–38.
6. Rezaee K., Rezakhani S.M., Khosravi M.R., Moghimi M.K. A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance. Personal and Ubiquitous Computing. 2024. vol. 28(1). pp. 135–151.

7. Ackerson J.M., Dave R., Seliya N. Applications of recurrent neural network for biometric authentication & anomaly detection. Information. 2021. vol. 12(7). DOI: 10.3390/info12070272.
8. Şengönül E., Samet R., Abu Al-Haija Q., Alqahtani A., Alturki B., Alsulami A.A. An Analysis of Artificial Intelligence Techniques in Surveillance Video Anomaly Detection: A Comprehensive Survey. Applied Sciences. 2023. vol. 13(8). DOI: 10.3390/app13084956.
9. da Costa K.A., Papa J.P., Passos L.A., Colombo D., Del Ser J., Muhammad K., de Albuquerque V.H.C. A critical literature survey and prospects on tampering and anomaly detection in image data. Applied Soft Computing. 2020. vol. 97. DOI: 10.1016/j.asoc.2020.106727.
10. Jebur S.A., Hussein K.A., Hoomod H.K., Alzubaidi L., Santamaría J. Review on deep learning approaches for anomaly event detection in video surveillance. Electronics. 2022. vol. 12(1). DOI: 10.3390/electronics12010029.
11. Habeeb R.A.A., Nasaruddin F., Gani A., Hashem I.A.T., Ahmed E., Imran M. Real-time big data processing for anomaly detection: A survey. International Journal of Information Management. 2019. vol. 45. pp. 289–307.
12. Arshad K., Ali R.F., Muneer A., Aziz I.A., Naseer S., Khan N.S., Taib S.M. Deep Reinforcement Learning for Anomaly Detection: A Systematic Review. IEEE Access. 2022. vol. 10. pp. 124017–124035.
13. Berroukham A., Housni K., Lahraichi M., Boulfrifi I. Deep learning-based methods for anomaly detection in video surveillance: a review. Bulletin of Electrical Engineering and Informatics. 2023. vol. 12(1). pp. 314–327.
14. Kiran B.R., Thomas D.M., Parakkal R. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. Journal of Imaging. 2018. vol. 4(2). DOI: 10.3390/jimaging4020036.
15. Musa A.A., Hussaini A., Liao W., Liang F., Yu W. Deep Neural Networks for Spatial-Temporal Cyber-Physical Systems: A Survey. Future Internet. 2023. vol. 15(6). DOI: 10.3390/fi15060199.
16. Albuquerque Filho J.E., Brandão L.C., Fernandes B.J., Maciel A.M. A review of neural networks for anomaly detection. IEEE Access. 2022. vol. 10(5). pp. 112342–112367.
17. Borowiec M.L., Dikow R.B., Frandsen P.B., McKeeken A., Valentini G., White A.E. Deep learning as a tool for ecology and evolution. Methods in Ecology and Evolution. 2022. vol. 13(8). pp. 1640–1660.
18. Amudha L., Pushpa Lakshmi R. Performance Analysis of Hybrid RR Algorithm for Anomaly Detection in Streaming Data. Computer Systems Science & Engineering. 2023. vol. 45(3). pp. 2299–2312.
19. Chang Y., Tu Z., Xie W., Luo B., Zhang S., Sui H., Yuan J. Video anomaly detection with spatio-temporal dissociation. Pattern Recognition. 2022. vol. 122. DOI: 10.1016/j.patcog.2021.108213.
20. Rezaei F., Yazdi M. A new semantic and statistical distance-based anomaly detection in crowd video surveillance. Wireless Communications and Mobile Computing. 2021. vol. 2021. DOI: 10.1155/2021/5513582.
21. Deepak K., Chandrakala S., Mohan C.K. Residual spatiotemporal autoencoder for unsupervised video anomaly detection. Signal, Image and Video Processing. 2021. vol. 15(1). pp. 215–222.
22. Ul Amin S., Ullah M., Sajjad M., Cheikh F.A., Hijji M., Hijji A., Muhammad K. EADN: An efficient deep learning model for anomaly detection in videos. Mathematics. 2022. vol. 10(9). DOI: 10.3390/math10091555.

23. Taghinezhad N., Yazdi M. A new unsupervised video anomaly detection using multi-scale feature memorization and multipath temporal information prediction. IEEE Access. 2023. vol. 11. pp. 9295–9310.
24. Liu T., Zhang C., Niu X., Wang L. Spatio-temporal prediction and reconstruction network for video anomaly detection. Plos one. 2022. vol. 17(5). DOI: 10.1371/journal.pone.0265564.
25. Le V.T., Kim Y.G. Attention-based residual autoencoder for video anomaly detection. Applied Intelligence. 2023. vol. 53(3). pp. 3240–3254.
26. Chriki A., Touati H., Snoussi H., Kamoun F. Deep learning and handcrafted features for one-class anomaly detection in UAV video. Multimedia Tools and Applications. 2021. vol. 80. pp. 2599–2620.
27. Deepak K., Srivathsan G., Roshan S., Chandrakala S. Deep multi-view representation learning for video anomaly detection using spatiotemporal autoencoders. Circuits, Systems, and Signal Processing. 2021. vol. 40(3). pp. 1333–1349.
28. dos Santos J.C.M., Carrijo G.A., de Fátima dos Santos Cardoso C., Ferreira J.C., Sousa P.M., Patrocínio A.C. Fundus image quality enhancement for blood vessel detection via a neural network using CLAHE and Wiener filter. Research on Biomedical Engineering. 2020. vol. 36. pp. 107–119.
29. Sharma N., Gupta S., Koundal D., Alyami S., Alshahrani H., Asiri Y., Shaikh A. U-Net model with transfer learning model as a backbone for segmentation of gastrointestinal tract. Bioengineering. 2023. vol. 10(1). DOI: 10.3390/bioengineering10010119.
30. Cai Y., Liu J., Guo Y., Hu S., Lang S. Video anomaly detection with multi-scale feature and temporal information fusion. Neurocomputing. 2021. vol. 423. pp. 264–273.
31. Yang Y., Fu Z., Naqvi S.M. Abnormal event detection for video surveillance using an enhanced two-stream fusion method. Neurocomputing. 2023. vol. 553. DOI: 10.1016/j.neucom.2023.126561.
32. Khan A.A., Nauman M.A., Shoaib M., Jahangir R., Alroobaea R., Alsafyani M., Binmahfoudh A., Wechtaisong C. Crowd anomaly detection in video frames using fine-tuned AlexNet Model. Electronics. 2022. vol. 11(19). DOI: 10.3390/electronics11193105.
33. Ali M.M. Real-time video anomaly detection for smart surveillance. IET Image Processing. 2023. vol. 17(5). pp. 1375–1388.

Р. Кришнан Шридеви — научный сотрудник, кафедра компьютерных наук и инженерии, Институт домоводства и высшего образования для женщин Авинашилингам. Область научных интересов: глубокое обучение, компьютерное зрение, машинное обучение, сетевая безопасность и обнаружение аномалий. Число научных публикаций — 6. 19pheop005@avinuty.ac.in; Тамил Наду, 641043, Коимбатур, Индия; р.т.: +91(960)535-9348.

Амудха П. — профессор, кафедра компьютерных наук и инженерии, Институт домоводства и высшего образования для женщин Авинашилингам. Область научных интересов: интеллектуальный анализ данных, машинное обучение, информационная безопасность. Число научных публикаций — 56. amudha_cse@avinuty.ac.in; Тамил Наду, 641043, Коимбатур, Индия; р.т.: +91(902)563-6594.