Journal of Theoretical and Applied Information Technology
15st September 2026. Vol.104. No.17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ENTROPY STACKED VARIATIONAL DENOISING AUTOENCODER (ESVDAE) BASED DENOISING MODEL AND ATTENTION MULTI-SCALE DILATED CONVOLUTIONAL NEURAL NETWORK (AMSDCNN) FOR SUSPICIOUS-HUMAN-ACTIVITY RECOGNITION (SHAR)

LIKHITH SR[1] , Dr. MAHALAKSHMI R[2]

[1] **PhD Scholar, Presidency School of Computer Science and Engineering, Presidency** university, Bengaluru, India.

Professor& Dean, Presidency School of Information Science, Presidency university, Bengaluru, India.

[1]likhith.20223cse0027@presidencyuniversity.in , [2]mahalakshmi@presidencyuniversity.in

ABSTRACT

SHAR plays an important role in automated surveillance systems and security monitoring, where noise, cluttered backgrounds, and variations in activity scale can reduce recognition reliability. In this paper, we propose an end-to-end deep learning framework that incorporates an Entropy Stacked Variational Denoising Autoencoder (ESVDAE) and an Attention Multi-Scale Dilated Convolutional Neural Network (AMSDCNN) for the recognition of suspicious human activity. ESVDAE combines variational and denoising autoencoders through a stacked VDAE and entropy-guided representation learning to mitigate noise interference and preserve discriminative visual information. Afterward, the denoised images are used for ResNet50-based feature extraction, whereas AMSDCNN includes attention, dilated convolution, MSFF, and depthwise separable convolution. From an information technology standpoint, the suggested framework can serve as an integrative solution for noise-robust representation learning, multi-scale context modeling, and computation-efficient video classification. Experiments are conducted using reconstruction, recognition, discrimination, and computational metrics on the UCF-Crime and CUHK Avenue benchmarks. PSNR values are 41.32 dB and 43.89 dB, while SSIM values are 0.9785 and 0.9874 for UCF-Crime and Avenue, respectively. As for the activity recognition task, the accuracies obtained by AMSDCNN are 92.75% and 94.84%, and the AUC values are 92.87% and 94.69% on the two respective datasets. Furthermore, computations show 60 FPS inference with 12.68 million parameters. Overall, the results demonstrate the viability of the suggested framework as a means of SHAR. At the same time, its computational cost, inability to handle temporal data, and cross-dataset performance remain crucial aspects to study further.

Keywords : *Variational Denoising Autoencoder (VDAE), Entropy Stacked Variational Denoising Autoencoder (ESVDAE), Suspicious Human Activity Recognition (SHAR), Denoising, ResNet50,*

1. INTRODUCTION

The increasing need for security in public spaces has caused security cameras to be used to keep track on routine activities and spot anomalous events, including airports, train stations, supermarkets, schools, and congested streets. Using video surveillance to identify suspicious human behavior, which entails categorizing behaviours as either normal or abnormal, is a significant area of research in this discipline [1]. The growing number of anomalous actions in public spaces is driving the urgent need for effective security measures [2].

This work emphasizes identifying and localizing abnormal activities in videos by analysing both temporal and spatial cues. Abnormal activities are characterized as actions that deviate from typical behavior, such as fighting, sneaking, or leaving baggage unattended in an airport. Mobile devices, radar, and cameras are examples of sensors used in human activity detection to

recognize irregular behavioural patterns. Such systems are widely applied in areas including interaction between humans and computers, monitoring, and surveillance suspicious actions for security purposes [3, 4]. Most existing solutions depend on footage captured from Closed-Circuit Television (CCTV) systems. Developing a system capable of detecting unusual or unexpected events in advance and alerting authorities would offer a significant improvement in proactive security management [5, 6].

The primary objective of employing the purpose of surveillance cameras is to make it possible to detect unusual human behavior early on. This task is vital in situations where timely human intervention is required, such as in crime prevention or anti-terrorism efforts. However, continuous manual monitoring is both labour-intensive and monotonous, as abnormal incidents occur in only about 0.01% of the total surveillance time meaning that approximately 99.90% of the monitoring effort is spent observing normal, uneventful activity [7].Additionally, surveillance systems generate large volumes of redundant video footage, leading to excessive storage requirements. To minimize human error and reduce storage costs, an efficient surveillance system capable of detecting unusual activities that can indicate to possible dangers must be developed. The task of recognizing abnormal activities has become increasingly challenging because real-time processing is required and there is a large volume of data. Consequently, intelligent monitoring systems are required to monitor all activities while accurately identifying the most hazardous and suspicious human behaviours [8,9].

The two types of surveillance systems are as follows. The first has a degree of autonomy, where videos are recorded and later analyzed by human experts. Such non-intelligent systems demand constant human monitoring, which is costly, inefficient, and challenging, as guards cannot continuously review all footage to detect suspicious activities. The second type is a surveillance system that is completely self-sufficient, ability to manage low-level activities including tracking, motion detection, anomalous event identification, and categorization. Particularly suitable is a system that may be effectively used both indoors and outdoors automatically identify anomalous or suspicious activity in advance notifying the appropriate authorities [10,11].

Traditional approaches depended largely on manually designed features and predefined rules, which restricted their effectiveness, especially in recognizing complex actions, as they were mostly suited for simple activities. In contrast, two deep learning models are recurrent neural networks (RNNs) and convolutional neural networks (CNNs), are models of modern machine learning approaches, have greatly enhanced the performance of SHAR systems. Unprocessed data may be used to automatically train these models features, allowing them to modify varied activity scenarios and achieve superior accuracy.

Another challenge lies in predicting human movements, which is complicated by the presence of noise in video data. Continuous monitoring of public spaces is difficult, highlighting the reason sophisticated video surveillance systems are necessary equipped with noise-removal models that can track human activities, classify them as normal or abnormal, and trigger alerts. Over the past three decades, various algorithms have been developed with differing levels of denoising performance. More recently, deep learning-based models have demonstrated significant promise by surpassing traditional methods. However, these approaches are limited by their need for large training datasets and high computational resources. Denoising Autoencoders (DAEs) have proven highly effective at recovering original signals by filtering out noise from input data. As a result, utilizing deep learning methods to denoised vibration time series data is anticipated to considerably increase maintenance prediction systems' effectiveness [12-13]. Typically, DAEs are trained by introducing random noise to enter data and gaining learning about reconstructs the clean, original signals. A Variational Denoising Autoencoder (VDAE) is an improved version of a Variational Autoencoder (VAE) that incorporates denoising techniques to enhance its performance. Then, the denoising is performed by subtracting the regenerated noise from the noisy input [14-15].

Deep learning approaches often face several challenges: 1) Excessive down sampling in the model can lead to the loss of critical feature information, is not entirely recoverable. 2) A limited receptive field prevents the model from adequately capturing local contextual information, reducing its ability to accurately detect abnormal activities. 3) Insufficient feature extraction capability within the network makes it

difficult to restore fine-grained details, causing residual noise in the frames. 4) The model struggles to detect objects or activities of varying sizes, limiting its accuracy in identifying abnormal behaviours.

The key contributions of the paper can be highlighted as follows:

Entropy-based denoising architecture: An Entropy Stacked Variational Denoising Autoencoder (ESVDAE) integrates variational and denoising learning to represent video frames in noisy environments robustly.

Representation learning in the presence of noise: An SVDAE model is designed to acquire robust and discriminative latent representations of data by minimizing reconstruction error.

Multi-Scale SHAR through attention mechanisms: An AMSDCNN model, which involves channel attention, spatial attention, dilated convolution, and Multi-Scale Feature Fusion (MSFF), has been designed to learn contextual information from different scales of receptive field.

Efficient recognition computation: The concept of depthwise separable convolution has been used to reduce the computational burden.

Comprehensive evaluation of performance: The proposed framework has been tested on datasets such as UCF-Crime and CUHK Avenue for denoising performance, classification, discrimination, and computational cost.

Applicability assessment: The computational cost and generalization constraints of the proposed system have been discussed in detail.

2. LITERATURE REVIEW

Gao et al., [16] proposed a Stacking Denoising Autoencoder (SDAE) and LightGBM (LGB) for human activity recognition (HAR) method. Through unsupervised learning, the SDAE is used to eliminate noise from unprocessed sensor data and identify the most instructive features, while LGB accurately classifies activities by retaining the inherent connections among features. In three common application scenarios human movement patterns and users' static and dynamic behaviours four datasets with distinct sensor combinations from various devices were extensively tested. The suggested strategy has highest accuracy when compared to XGBoost, CNN, CNN with statistical features, and single SDAE, achieving superior recognition performance.

Pham et al., [17] proposed a sensCapsNet, a capsule network for spatial-temporal data from wearable devices by the usage of portable sensors for human activity recognition (HAR). CapsNet architecture is assessed by modifying capsule layer dynamic routing. To develop a non-intrusive HAR method, the 19NonSens dataset was gathered from twelve individuals who used e-shoes and a smartwatch while performing 19 activities in various contexts. The proposed method demonstrates promising results, enabling real-time computation in a life-logging application and achieving over 80% accuracy for five common upper-body activities. SensCapsNet performs superior to traditional CNN and Long Short-Term Memory (LSTM) models, according to results of the experiment.

Lv et al., [18] a unique hybrid deep learning network is suggested to determine human activity (HAR) from various types of sensor data. The network incorporates a ConvLSTM pipeline that effectively leverages temporal information extracted at each layer. To maximize information flow across layers, a Dense Connection Module (DCM) is implemented. Additionally, a Multilayer Feature Aggregation Module (MFAM) is introduced to extract spatial features and combine outputs from all convolutional layers based on the importance of features at various spatial positions. Then, using a softmax function and a completely connected layer, calculate class probabilities after the MFAM output has been processed into two LSTM layers to capture temporal constraints. University of Milano Bicocca Smartphone-based Human Activity Recognition (UniMiB-SHAR) and Opportunity are two benchmark datasets used to assess the network. Multimodal fusion and various hyperparameter settings are employed to analyze performance, while ablation and visualization to show how successful the suggested strategy is, tests are carried conducted.

Ullah et al., [19] an efficient intelligent detection of anomalies framework is developed to function in surveillance networks with less computing time, on the basis of deep features. This method captures crucial information for spotting abnormal occurrences by first employing to take a series of frames and extract statistical features using a pre-trained CNN model. In complex urban surveillance settings, these extracted deep features are subsequently input into a multi-layer Bi-Directional Long Short-Term Memory (BD-LSTM) network, which reliably identifies current occurrences as either normal or unusual. Several

benchmark anomaly detection datasets were used in extensive research are carried on to confirm the suggested framework's efficacy in challenging surveillance scenarios.

Ullah et al., [20] suggested lightweight framework for activity identification based on deep learning. First, an efficient CNN model trained on two surveillance datasets is used to identify persons in surveillance video streams. The extremely rapid MOSSE (Minimum Output Sum of Squared Error) object tracker is then used to follow each identified individual throughout the video.For each observed individual, LiteFlowNet CNN extracts the effective Pyramidal convolutional features from successive frames.A unique Deep Skip Connection Gated Recurrent Unit (DS-GRU) helps identify activity by capturing temporal changes across the frame sequence.The suggested strategy outperforms the state-of-the-art techniques in real-time surveillance, as shown by experiments on five benchmark activity recognition datasets.

Ullah et al., [21] developed lightweight convolutional neural networks (LWCNNs) for surveillance situations in anomaly identification. A residual attention-based LSTM network efficiently identifies unusual activity in surveillance footage after spatial data are collected from video frames using a CNN. The model applicability for smart city surveillance is revealed by the successful anomaly identification achieved by integrating for sequence learning, the LSTM uses CNN features with residual blocks that are representative. UCF-Crime, University of Minnesota (UMN), and Avenue datasets, provide improvements over state-of-the-art techniques with accuracy increases of 1.77%, 0.76%, and 8.62%, confirming the efficiency of model in complex surveillance settings.

Khan et al., [22] proposed a hybrid model for recognizing human activities that combines CNN extracts spatial features and LSTM with temporal dynamics. Furthermore, dataset including 12 different categories of human physical activity is produced, acquired using the Kinect V2 sensor from 20 individuals. HAR is performed by thorough ablation study on a variety of both traditional machine learning algorithms and deep learning. CNN-LSTM method has highest 90.89% accuracy demonstrates how successful it is for applications involving the identification of human activities.

Qasim Gandapur and Verdú [23] suggested automated deep learning model to identify and stop unusual activity. Real-world video surveillance systems with High-level spatial features and temporal features may be retrieved using ResNet-50 from input video streams from time-series data using a Convolutional GRU (ConvGRU). ConvGRU-CNN deep learning framework is produced which can successfully detects unusual activity. This model can effectively distinguishes between normal and abnormal occurrences UCF-Crime dataset testing and demonstrating its capability to correctly categorize each anomaly. ConvGRU-CNN achieves an accuracy of 82.22% and outperforms other comparable deep learning approaches.

Hanief Wani and Faridi [24] proposed real-time video monitoring using a Long-term Recurrent Convolutional Network (LRCN) which would consequently recognize and inform the authority to suspicious activity, including fights, accidents, and robberies. The two main parts of the system are real-time alerts production and activity detection based on LRCN. State-of-the-art accuracy, precision, and recall ratings are obtained when its performance is assessed using a bespoke dataset that was assembled from two publically accessible datasets. Airports, train stations, and retail malls are just a few of the public places where this system may be installed to improve security and safety.

Dua et al., [25] proposed an Inception-inspired CNN-GRU Network (ICGNet) model by combining CNN and GRU architectures. The CNN component draws inspiration from the Inception module, employing convolutional filters of multiple sizes simultaneously on the input to capture information at different scales. For local data patches, the network may calculate more abstract features thanks on the same layer as these filters of various sizes. Furthermore, the model is able to extract useful hidden information across channels by combining information across the channel dimension using 1×1 convolutions. By combining CNN and GRU, ICGNet successfully detects both local features and long-term connections in multidimensional time series data .Health monitoring, pattern-based surveillance, interactive gaming, and robot learning are just a few of the Human-Computer Interaction (HCI) sectors in which this HAR system may be used.

Lafontaine et al., [26] proposed to Ultra-Wideband (UWB) radar data is processed using an unsupervised deep convolutional neural

network autoencoder (CNN-AE) to learn and remove background noise. The effectiveness of this filtering approach is evaluated by comparing it with using a CNN classifier on raw data under a Leave-One-Subject-Out (LOSO) scheme. The generalization capability is also tested on datasets with varying participant numbers (1, 5, 10, and all 19 participants), highlighting the challenges of generalization in HAR and the impact of dataset size and subject diversity. Results demonstrate that for HAR uses, UWB radar data may be efficiently filtered and generalized using the unstructured CNN-AE, offering simpler learning constraints and practical implementation. Overall, the performance indicates that CNN-AE-based unsupervised filtering is highly effective for human activity recognition.

Lee et al., [27] proposed missing joint reconstruction model to enhance the efficiency of skeleton-based HAR. Denoising graph autoencoder (DGAE) is introduced which regards missing joints as noise corrupted information and aims to reconstruct them to be close to their original coordinates. When the encoder of the proposed model compresses the noised input into a latent vector, a masking Laplacian matrix is introduced to reduce the effect of the missing joints' features. The masking Laplacian matrix adjusts the effect of features between a missing joint and its adjacent joints by altering the weights of an adjacent matrix. In the decoder, a Laplacian matrix, which represents the connections among the joints, is utilized to reconstruct an output from the latent vector. The results of the study indicate that the suggested model reconstructs the coordinates of missing joints with a marginal error. In addition, the performance of skeleton-based HAR is enhanced by reconstructing the missing joints.

Roka and Diwakar [28] proposed a Deep Stacked Denoising Autoencoder (DSDAE), a powerful unsupervised deep learning model for image denoising to detect and localize abnormal activities in videos. DSDAE demonstrates superior performance in image denoising compared to traditional handcrafted methods. The approach employs separate encoders to extract appearance features from motion features from optical flow images in addition to clear and noisy images. The decoder receives these extracted features after they have been merged early. Pixels with reconstruction errors exceeding a set threshold are identified as abnormal. Experimental results, the public accessible benchmark datasets Ped1, Ped2, CUHK Avenue, and ShanghaiTech, the suggested model outperforms state-of-the-art techniques both analytically and qualitatively.

Existing methods suffer from lack of integration, insufficient feature representation, and poor robustness to noise. The proposed model overcomes these challenges by combining denoising, deep feature extraction, and attention-based multi-scale classification into a unified framework. Deep learning methods often face several challenges: 1) Excessive downsampling can cause loss of critical feature information something is irretrievable. 2) The model is constrained by a small receptive field, ability to capture local contextual information. 3) The network may have inadequate feature extraction capability. 4) The model struggles to detect objects or patterns of varying sizes, reducing its accuracy in identifying abnormal activities. These issues have been solved by introducing a novel denoising, and classification model.

Critical synthesis of previous works. The papers considered reveal three major approaches in SHAR and related human activity recognition, namely, temporal modeling employing LSTM/GRU networks, noise removal using an autoencoder architecture, and learning multi-scale or attention-based spatial representations. CNN–LSTM, DS-GRU, and ConvGRU-CNN pay particular attention to temporal dependencies, while CNN-AE, DGAE, and DSDAE address noise or incomplete visual representations. Recent research has sought to address the challenges of real-time anomaly detection and to develop more sophisticated learning methods for surveillance [4,7,24,29]. However, these approaches usually address different aspects of the problem rather than providing an integrated learning of noise-aware representation, feature residuals, multi-scale context, and attention-based classification. The suggested architecture aims to solve this task by sequentially connecting ESVDAE, ResNet50, and AMSDCNN. However, such an integration adds additional computational stages and does not employ long-term temporal dependency modeling, motivating further experiments and evaluation.

3. PROPOSED METHODOLOGY

This paper introduces the Entropy Stacked Variational Denoising Autoencoder (ESVDAE) for removing noise from video frames. Layer by

layer, VDAE networks are stacked with entropy in the ESVDAE architecture to extract robust and discriminative noisy pixels. The extracted noisy pixels are then fed into a softmax classifier to assess the noise reduction's efficacy. Attention Multi-Scale Dilated Convolutional Neural Network (AMSDCNN) is employed for suspicious human activity detection. In AMSDCNN, dilated convolutions with varying dilation rates are cascaded to expand the kernel's receptivity region, enhancing the model's understanding of global context and reducing errors in distinguishing normal and suspicious activities. Additionally, Multi-Scale Feature Fusion (MSFF) is proposed, in which parallel convolutional layers with varying dilatation rates are used to sample feature maps to enhance SHAR identification. Benchmark datasets like Avenue and the recommended detection methods are evaluated using UCF-Crime. The suggested model's entire process is shown in Figure 1.

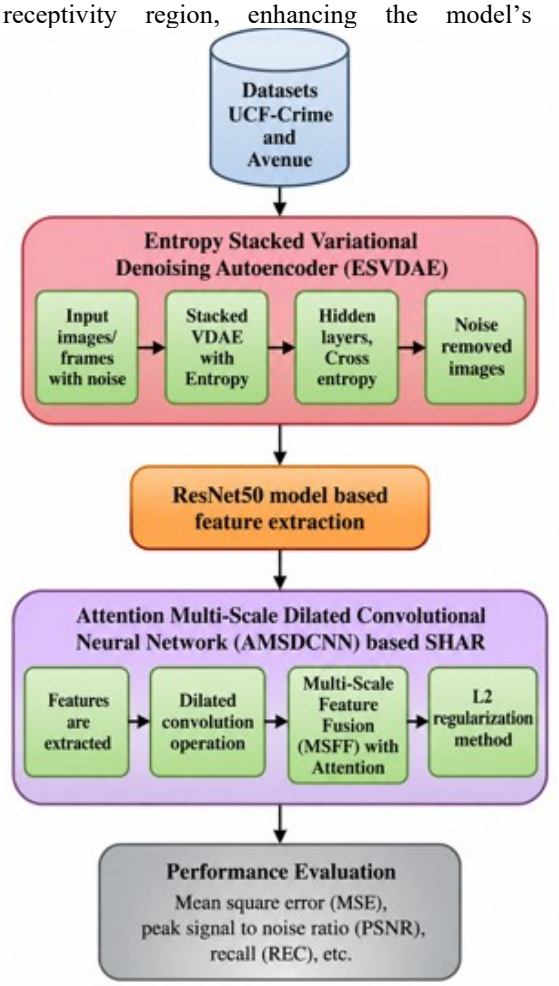

Figure 1. Overall Process Of Propsoed Model

DATASETS

UCF-Crime dataset: Images taken exclusively for identification of anomalies in monitoring in the real world films are included in this collection, which is taken from the dataset on UCF. Images are captured from each 10th frame of the UCF Crime Dataset video selected and grouped according to the corresponding video class. The dataset includes 1,900 event sequences spanning 13 anomaly categories, consistent with UCF crime's initial dataset [29]. Each of the 290 test sets at the video level and the 1,610 instructional sets labelled at the frame level include an equal amount ofboth typical and unusual videos. The components of the test set are 111,308 images, whereas the training set consists of 1,266,345 image files. Derived from https://www.kaggle.com/datasets/odins0n/ucf-crime-dataset.

Avenue dataset: CUHK Avenue Dataset: The CUHK Avenue dataset comprises 37 video sequences: 21 training and 16 testing. The dataset contains approximately 30,000 frames, including normal training scenes and abnormal events in the test sequences. Uniform frame sampling, resizing to 224 × 224 pixels, normalization, background-consistency processing, and annotation alignment are applied as described in the experimental configuration summarized in Table 1.

Table 1. Dataset Details for Anomaly Detection

Parameter	UCF-Crime Dataset	CUHK Avenue Dataset
Number of Videos	1,900	37
Training Videos	1,610	21
Testing Videos	290	16
Number of Frames	Training: 1,266,345Testing: 111,308	30,000
Frame Sampling	Uniform sampling (every 10th frame extracted)	Uniform sampling (every *n*th frame / full frames)
Preprocessing Steps	Resizing (224×224); Normalization; Noise filtering; Label encoding	Resizing (224×224); Normalization; Background consistency; Annotation alignment
Data Augmentation	Horizontal flip; Rotation; Scaling/Zoom; Brightness adjustment; Random cropping	Horizontal flip; Rotation; Scaling; Illumination variation

Journal of Theoretical and Applied Information Technology
15st September 2026. Vol.104. No.17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.2 ENTROPY STACKED VARIATIONAL DENOISING AUTOENCODER (ESVDAE) BASED DENOISING

Entropy Stacked Variational Denoising Autoencoder (ESVDAE) model is introduced for removing noises from the frames in the videos. In the suggested approach, a Variational Autoencoder (VAE) and a DAE are initially integrated to create VDAE. The ESVDAE architecture is then built by extracting the robust and discriminative noisy pixels by layer-by-layer stacking VDAE networks with entropy. To get the noise reduction results in frames, the noisy pixels are loaded into a softmax classifier.

Variational Autoencoder (VAE): The VAE is a contemporary unsupervised learning method with small excess fitting and rapid resolution that includes a sample layer, a decoder, and an encoder.In Figure 2, the VAE's architecture is shown. For image denoising, VAE adds a sample layer between the encoder and decoder, in contrast to conventional autoencoders. With a Gaussian distribution, the model learns the latent variable (z) during training, where parameters such as mean and variance are automatically estimated through the network. This allows the decoder to generate new input images or frames from SHAR datasets based on the latent variable (z).

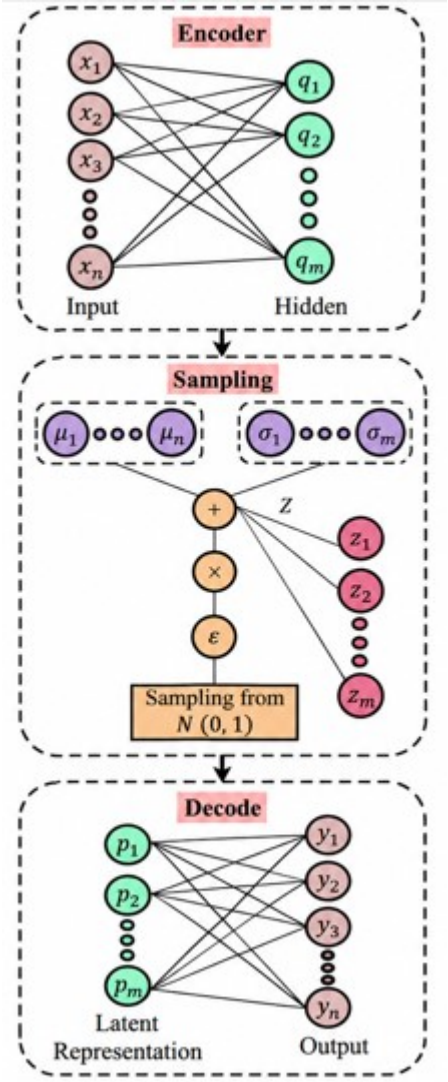

Figure 2. Architectural Model For The Vae[30]

The produced image is made to resemble the original image as much as possible by using equation (1) to optimize the marginal distribution [30],

$$p_\theta(x) = \int p_\theta(x|z)\, p_\theta(z) dz \quad (1)$$

Reconstruction of image x by latent variable z is represented by $p_\theta(x|z)$, while its prior dispersion is provided by $p_\theta(z)$. VAE model automatically trains an approximation posterior to variation, $q_\varphi(z|x)$, to replace actual probability distributions $p_\theta(z|x)$, where φ and θ represent encoder and decoder parameters. The probabilities function's unpredictable lower limit, which usually contains two terms, is used to calculate the loss function (Equation (2)) [30],

$$L(\theta,\varphi;x^{(i)}) = -D_{KL}(q_\varphi(z|x^{(i)})\|p_\theta(z)) + E_{q_\varphi(z|x^{(i)})}[\log p_\theta (x^{(i)}|z)] \quad (2)$$

where the Kullback–Leibler divergence is represented by the first term $D_{KL}(q\|p)$, and the reconstruction error is denoted by the second term $E_q(\log p)$. Assuming that the Kullback–Leibler divergence may be reformulated as follows [30], Normal distribution standard N(0,1) is met by $p_\theta(z)$, whereas Gaussian distribution N(μ,σ^2) is followed by $q_\varphi(z|x)$.

$$-D_{KL}(q_\varphi(z|x^{(i)})\|p_\theta(z)) = \frac{1}{2}\sum_{j=1}^{J}\left(1+\log\left(\sigma_j^{(i)}\right)^2-\left(\mu_j^{(i)}\right)^2-\left(\sigma_j^{(i)}\right)^2\right) \quad (3)$$

The latent variable J defines z's dimension, and the jth element of $q_\varphi(z|x)$ is represented by $\sigma_j^{(i)}$ and $\mu_j^{(i)}$. Meanwhile, the following is an expression for the reconstruction error:

$$E_{q_\varphi(z|x^{(i)})}[\log p_\theta(x^{(i)}|z)] = \frac{1}{L}\sum_{i=1}^{L}\left(\log p_\theta\left(x^{(i)}|z^{(i,l)}\right)\right) \quad (4)$$

where L is the sample quantity from $q_\varphi(z|x)$. According to the equation $z=\mu+\varepsilon.\sigma$, z is updated using the reparameterization method. Consequently, the following is a rewriting of Equation (2) [30],

$$L(\theta,\varphi;x^{(i)}) = \frac{1}{2}\sum_{j=1}^{J}\left(1+\log\left(\sigma_j^{(i)}\right)^2-\left(\mu_j^{(i)}\right)^2-\left(\sigma_j^{(i)}\right)^2\right) + \frac{1}{L}\sum_{l=1}^{L}\log p_\theta\left(x^{(i)}|z^{(i,l)}\right) \quad (5)$$

where $z^{(i,l)}=\mu^{(i)}+\sigma^{(i)}\varepsilon^{(l)}$ and $\varepsilon^{(l)}\sim N(0,1)$. Equation (2) provides the loss function, which is minimized using the VAE model.The VAE model's loss function incorporates KL divergence is able to produce new images by sampling and decoding in addition to efficiently learning the fundamental qualities of the source image.

Variational Denoising AutoEncoder (VDAE): By introducing noise into the input image, the Denoising Autoencoder (DAE), an enhanced form of AE, may increase the AE model's resilience and generalization capabilities while avoiding over-fitting issues. By incorporating denoising into the original VAE architecture, the notion of VDAE is introduced, taking into account the advantages of both DAE and VAE. This enables the extraction of strong features from frames or images that are noisy [30].

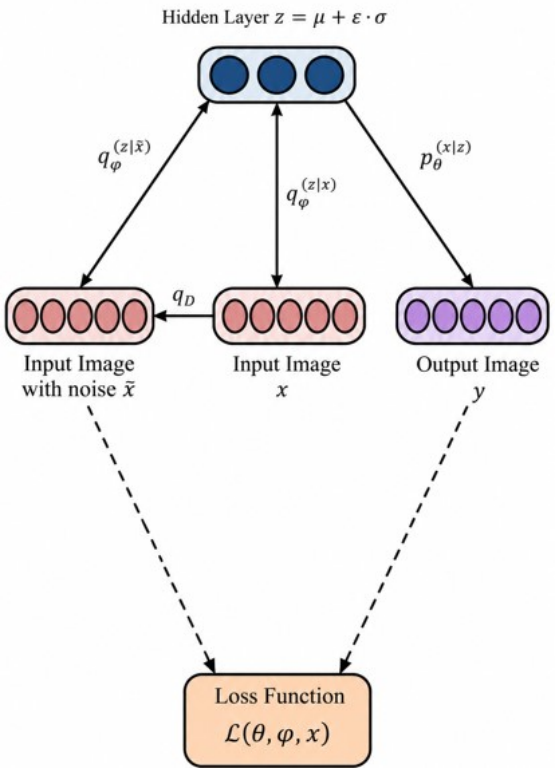

Figure 3. Vdae Model's Architecture[30]

Figure 3 depicts the structure of the VDAE model. Based on Equation (6), the suggested VDAE is specifically based on a binomial distribution, the corrupted input images, $\hat{x} = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n\}$, begins by introducing noise into the first input image, $x = \{x_1, x_2, \ldots, x_n\}$.For the latent features to be automatically learned With μ and σ representing the features derived from the predicted probabilities distribution, z=μ+ε. σ,$q_\varphi(z|x)$, VAE's hidden layer connects to incorrect input image. ε is a normal distribution extra variable N(0, 1).Finally, the decoder is coupled to the learnt latent features, z, to produce a new noise-removed images, y, that substantially resembles the original image, x [30].

$$\hat{x}\sim q_D(\hat{x}|x) \quad (6)$$

where $q_D$is represents the binomially distributed random noise and $\hat{x}$ represents the noisy input image.

ESVDAE model: The latent variables are extracted layer by layer using SVDAE. In particular, the ESVDAE model's training procedure is split into two phases: supervised fine-tuning and unsupervised pre-training.The output (noise-removed images) the input of the previous VDAE will be used subsequent ESVDAE to carry out the ESVDAE model's training procedure during the unsupervised pre-training phase.The representative latent features, which are considered the classifier's input, may be retrieved after the ESVDAE model's pretraining phase is finished and all VDAE models have been trained well. By reducing the impact of gradient dispersion, the ESVDAE pre-training approach may provide the local best solution of noise-removed images.The network settings may be updated using the back propagation approach of ESVDAE to update the model parameters during the ESVDAE model's assisted fine-tuning stage.Using the probability $p(\theta)_j$from Equation (7), the identification result is generated using the softmax classifier at the conclusion of the ESVDAE model [30],

$$p(\theta)_j = \frac{e^{(\theta^{(j)}x)}}{\sum_{k=1}^{K} e^{(\theta^{(k)}x)}}, j = 1, \ldots, n \quad (7)$$

where the input image is x and the number of classes is K, $p(\theta)_j$ is the probability of the jth activity class, and θ parameters are learnt using SVDAE.The difference between the model's noise-detected probabilities distributions and its actual probabilities distributions is measured using the loss function for cross-entropy. Equation (8) for the cross-entropy loss function in classification of multiple classes indicates the following:

$$L = \frac{1}{N}\sum_{i} L_i = -\frac{1}{N}\sum_{i}\sum_{c=1}^{M} w_c \cdot y_{ic} \log(p_{ic}) \quad (8)$$

where multiclass classification, w_c is the weight given to the c^{th} class. The following is the equation (9):

$$w_c = bias + (1 - bias).e^{-\alpha.size_c} \quad (9)$$

where the bias term *bias* keeps the high-frequency class weights from falling to zero. The definition of the α is as follows:

$$\alpha = \frac{base}{size} \quad (10)$$

where *base* is the function's decay rate. The number of images in the complete SHAR collection is represented by the representation of *size*.

3.4 RESNET50 MODEL BASED FEATURE EXTRACTION

Residual connections are used by the well-known deep convolutional neural networks ResNet50 To promote gradient flow in deep networks and improve feature learning.Images of 224 × 224 x 3 dimensions are used as input for the model.As the ResNet50 basic loads do not include top classification layers, they allow customization for the specific dataset. In addition to the ResNet50 foundation, to reduce feature map spatial dimensions to one value, a Global Average Pooling (GAP) layer is implemented and give a more compact and less overfit-prone output.The probability for the HAR is then estimated by including an associated dense layer with softmax activation. The training error and accuracy are improved by the noise reduction.In the final model, the best classification layers appropriate for the HAR dataset are combined with pre-trained feature extraction of ResNet50 [31–32]. The model may always be used to classify future activities into the designated classes after training,

$$F_{ResNet50}(X) = GAP2D(f_{Residual\,Blocks}(f_{conv}(X))) \quad (11)$$

Equation (11) illustrates how GAP2D produces a vector based of the feature map's dimensions, X is input image, $f_{conv}(X)$is ResNet50's initial convolution and max-pooling layers, and $f_{Residual\ Blocks}$ is a sequence of 50 residual blocks that make modifications while maintaining identity,

$$F_{Tuned-ResNet50}(X) = \text{Softmax}\Big(w_3 \cdot \sigma\Big(w_2 \cdot \text{Dropout}\Big(\sigma\Big(w_1 \cdot \text{Dropout}\Big(BN(F_{ResNet50}(X))\Big) + b_1\Big), p\Big) + b_2\Big) + b_3 \quad (12)$$

Softmax is the output class probabilities as shown in equation (12), where $F_{ResNet50}(X)$σ is ReLU activation, Dropout is regularization layers with

Journal of Theoretical and Applied Information Technology
15st September 2026. Vol.104. No.17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

dropout probability p, and is feature extraction from the ResNet50 basis.

3.3 ATTENTION MULTI-SCALE DILATED CONVOLUTIONAL NEURAL NETWORK (AMSDCNN) CLASSIFICATION

AMSDCNN model is introduced for suspicious human activity detection. The AMSDCNN efficiently minimizes the error of the normal and suspicious detection by increasing the kernel's receptive field through cascading dilated rates. This helps the model comprehend global context information. There are three primary components to the AMSDCNN model, deep network model end-to-end. Figure 4 depicts the network. The encoder, which is the initial component, uses rich hierarchical representation and SHAR datasets to learn the feature information of the input frames or image. Multi-Scale Feature Fusion (MSFF) (see Figure 6) is the second section. It uses several simultaneous dilation convolutions with varying expanded rates to capture multiscale features on feature maps. Deconvolution on the feature map progressively upsamples the third element, which is decoded to the input image's x resolution[33].

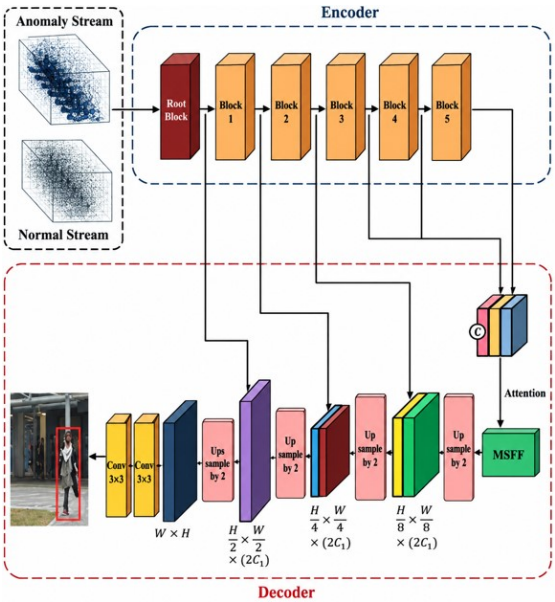

Figure 4. Attention Multi-Scale Dilated Convolutional Neural Network (Amsdcnn) Model

Encoder: Encoder the residual consists of three convolution operations: $FM \xrightarrow{1\times1} FM^1 \xrightarrow{3\times3} FM^2 \xrightarrow{1\times1} FM^3$. Using a small convolution kernel (1×1), the input feature map's video frame count is reduced from C^1 to C^2, and FM^1 is the output.FM^1 features are then extracted and FM^2 is produced using a large convolution kernel (3×3). The output FM^3 is then obtained by restoring the channel number from $C^2$to C^1 using the little convolution kernel (1×1).To achieve the final SHAR result, the FM and FM^3 features are put together. To create the feature map $FM', FM' = [FM'_1, \dots, FM'_C]$, the shortcut technique combines the feature information of FM and FM^3,

Encoder: Encoder the residual consists of three convolution operations: $FM \xrightarrow{1\times1} FM^1 \xrightarrow{3\times3} FM^2 \xrightarrow{1\times1} FM^3$. Using a small convolution kernel (1×1), the input feature map's video frame count is reduced from C^1 to C^2, and FM^1 is the output.FM^1 features are then extracted and FM^2 is produced using a large convolution kernel (3×3). The output FM^3 is then obtained by restoring the channel number from $C^2$to C^1 using the little convolution kernel (1×1).To achieve the final SHAR result, the FM and FM^3 features are put together. To create the feature map $FM', FM' = [FM'_1, \dots, FM'_C]$, the shortcut technique combines the feature information of FM and FM^3,

$$FM'_C = \sum_{i=1}^{W}\sum_{j=1}^{H}\left(FM_c(i,j) + FM_c^3(i,j)\right) \quad (13)$$

The following pertains to convolution operations: $X \rightarrow Y$, $X \in \mathbb{R}^{W^1\times H^1\times C^1}$, $Y \in \mathbb{R}^{W^2\times H^2\times C^2}$. In the following notation, assume that F is a conventional convolutional operator for the purpose of simplicity.

To enhance feature learning, Channel Attention (CA) and Spatial Attention (SA) are applied in the equation (14),

$$FM_{att} = M_s(M_c(FM'_C)\odot FM'_C) \quad (14)$$

ith k_c standing for the c[th] filter's parameters, let $K = [k_1, k_2, \dots, k_c]$ represent the learnt set of filter kernels. Consequently, F outputs are $Y = [y_1, y_2, \dots, y_{C^2}]$,

$$y_c = \sum_{n=1}^{C^1} k_c^n * x^n \quad (15)$$

where $*$is denoted as convolution, $k_c = [k_c^1, k_c^2, \dots, k_c^{C^1}]$, and $X = [x^1, x^2, \dots, x^{C^1}]$. To enhance the model's receptive field, stride convolution or pool procedures are typically used to minimize feature map resolution in the video frames. However, a significant amount of feature information, particularly the few frames in the videos, would be

lost if these two processes are used excessively. To preserve modeling receptivity in SHAR recognition, reduce the downsampling factor from the initial 32 to 16 and apply dilated convolution on the convolutional layer. Block 3, Block 4, and Block 5 modules in the experiment used dilated convolution in place of the usual convolution; each module had three residual modules. In these three residual modules, as shown in Figure 5, define the dilated rate of Conv 3×3 as (d_1,d_2,d_3). To create a feature map G, the feature maps produced by blocks 3, 4, and 5 are finally concatenated.

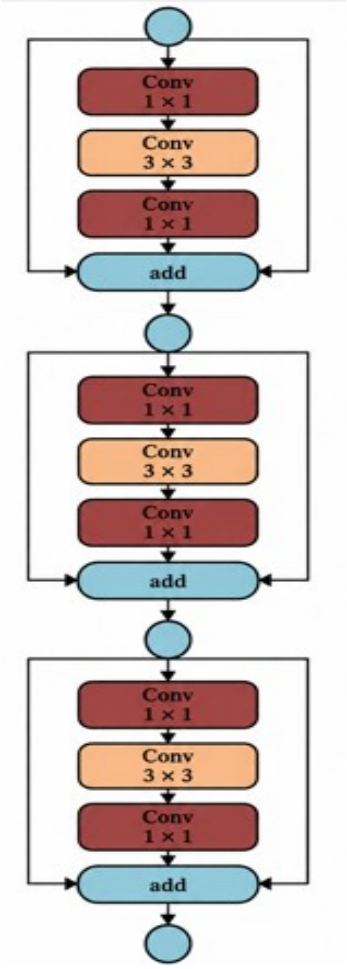

Figure 5. Detail Structure Of The Block

By maintaining the feature resolution constant, the number of filter parameters is decreased, dilated convolution may increase the receptive field size. The following is an expression for equation (16):

$$y[i] = \sum_{k=1}^{K} x[i + d \times k] \times w[k] \quad (16)$$

Filter size is K, dilation rate is d, and kth filter parameter is w[k], x[i] is the input image for video frames, and y[i] is the output recognized results.By adding d-1 zeros between two successive convolution filter values, dilated convolution is the same as convolving the input feature x.The dilated filter size that results from a convolution filter of size k×k is $k_d \times k_d$, where $k_d = k + (k-1) \times (d-1)$. Thus, the receptive field is large for a significant dilation rate. Dilated convolution is used in convolutional layers, k is the size of the convolution filter if the dilated ratio is r. As determined by equation (17), the receptive field size,

$$R_K = (k-1) \times (r-1) + k \quad (17)$$

The appropriate receptive field size is 9, for example, and the convolution kernel size is 3×3, with a dilation rate of r=4.Expanding the receptive field is another benefit of stacking many convolutional layers. Assume L-1 and L are convolutional layers with k_1 and k_2 convolution filters. The size of the L+1 layer's receptive field is determined by equation (18).

$$R_L = k_1 + k_2 - 1 \quad (18)$$

A receptive field with a feature map of 13 is produced by two layered convolution layers with filter widths of 5 and 9.To maintain model receptivity minimizing frame feature loss, use dilated convolution. By putting "zeros" between the convolutional kernel's effective parameters, dilated convolution is created. Assume that K×K convolution kernels with dilated rates of $[d_1, \ldots, d_i, \ldots, d_n]$ are associated with each group's cascade of N dilated convolutions. Following a sequence of convolution procedures, all feature information in the receptive field

must be fully covered by a dilated convolution [34]. In the convolution kernel, provide the most difference between two effective weights as follows:

$$MD_i = \begin{cases} 0 \le MD_i < K & i = 1,2 \\ (K-1) + K \times (MD_{i-1} - 1) & i > 2 \end{cases} \quad (19)$$

Dilation rates are d_1, d_2, d_3. The use of dilated convolution of various dilated rates in MSFF enhances the identification of aberrant activity and recognizes the precision of SHAR.To acquire the feature map z, MSFF first uses a 1x1 the feature map G's channel count is cut in half via convolution. One global average pooling layer and four parallel convolution layers are then used to gather feature information on the feature z.The MSFF module's parameters and computations may

be reduced to 12 by reducing the input feature map's channel number, which will speed up the model.

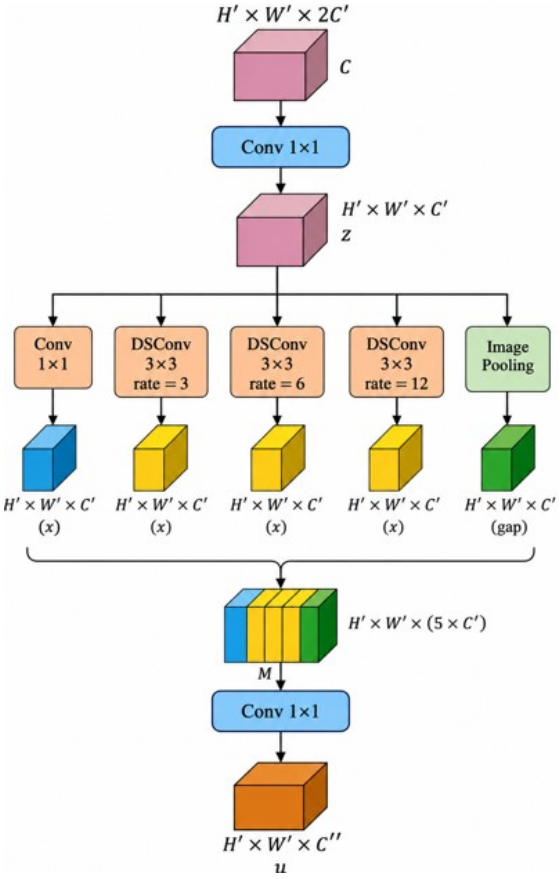

Figure 6. Multi-Scale Feature Fusion (Msff) Structure

Four parallel convolution layers are constructed from composed of one 1x1 convolution layer and three depthwise separable convolutions (DSC) with different dilation rates. To improve recognition performance, DSC, a potent procedure, is employed to lower the number of parameters. Multiscale context feature information may be gathered through three convolutional layers with various dilation rates, whereas the 1x1 layer is used preserves the current scale's feature information. Global context information is obtained at the image level using GAP [35], and to get the desired spatial dimension, the feature is bilinear upsampled before being included into the model. Equations (20–21) are used to determine the c[th] element of the image-level features gap,

$$z_c = \frac{1}{W' \times H'} \sum_{i=1}^{W} \sum_{j=1}^{H} z_c(i,j) \quad (20)$$

$$GAP_c = F_{BI}(z_c) \quad (21)$$

where $F_{BI}(z_c)$ is denoted as the bilinear upsample, $z = [z_1, z_2, \ldots, z_{C'}]$ and $GAP = [GAP_1, GAP_2, \ldots, GAP_{C'}]$. Finally, M is obtained by concatenating the feature Maps of every branch, $M = [x_1, x_2, x_3, x_4, GAP], u = Conv_{1\times1}(M)$ and the final feature map u is then obtained by fusing this multi-scale data using a [1×1,256] convolutional layer. After fusion, attention is applied by equation (22),

$$u_{att} = M_s(M_c(u) \odot u) \quad (22)$$

Decoder: The MSFF feature map u is deconvolved and upsampled by 2 until the decoder restores it to the same resolution as video x with input frames. After each deconvolution, the Skip connection concatenates feature and details at the decoding layer that are low level. The final result, y, is produced after two 3x3 convolution layers have been used to enhance the feature information.

Loss: Preventing over-fitting and enhancing the convolutional layer's identification capabilities are the primary objectives of the L2 regularization technique. To effectively utilize context information for SHAR identification and to increase the receptive field, AMSDCNN uses dilated convolution. Equation ((22)) demonstrates how the training of the complete network is constructed as a SHAR issue with regard to the real frames from videos.

$$\mathcal{L}(\chi; \theta) = \lambda \|W\|_2^2 - \left[\sum_{x \in \chi} \phi\left(x, \ell(x)\right) + \|y - \ell(\chi)\| \right] \quad (23)$$

The first component in this case is the regularization term, while the second one contains the L2 distance in the training set and the loss term for the target classifier. It is the hyperparameter λ that determines the trade-off among these two terms. W denotes the output y' inference parameters. Assuming y is the ground truth, $\ell(x)$ is the anticipated SHAR result, and $\phi(x, \ell(x))$ is the cross entropy loss for pixel x compared to the true label $\ell(x)$. End-to-end tuning of the deep contextual network θ={W} parameters is achieved through complete loss function minimization L(χ;θ).

4. RESULTS AND DISCUSSION

In the experiments, the suggested approach is contrasted with a number of other approaches using avenue and the UCF-Crime Dataset are public datasets. Methods are implemented using MATrix LABoratory R2022a (MATLABR2022a) using a Tesla K80 GPU, an Intel(R) Xeon(R) E5-2620 v3 2.40GHz CPU, and Ubuntu64 as the operating system.The quality of denoised images has also been assessed using PSNR, MSE, and SSIM. Gradient descent training uses the Adam optimizer with parameters: $\beta_1 = 0.9,\ \beta_2 = 0.999, \varepsilon = 1e^{-8}$. Starting at 0.0001, poly learning rate is used in the implementation. There are comparisons with modern techniques using two datasets to show how successful the suggested strategy.

4.1 Evaluation Criteria

Evaluation Criteria. The chosen evaluation criteria aim to assess the complementary properties of the suggested framework. MSE and PSNR are employed to quantify reconstruction quality at the pixel level after denoising, whereas SSIM is used to measure structural information preservation. For suspicious human activity recognition, Precision measures the accuracy of suspicious activity prediction, while Recall measures the detection of suspicious activity occurrences. F-measure is the harmonic mean of Precision and Recall, balancing these two measures and providing an adequate measure of classification performance. Accuracy is measured as the proportion of both normal and suspicious activities properly recognized. Additionally, the Precision-Recall (PR) curve and Area Under the Curve (AUC) are evaluated to assess discriminative performance across different classification thresholds. Also, computational measures, such as parameter count, inference speed, training time, and execution time per epoch, are examined to explore the efficiency-performance trade-off of the proposed framework.

PSNR is calculated a follows,

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{2^n}{\mathrm{MSE}}\right)^2 \quad (24)$$

$$\mathrm{MSE} = \frac{1}{H \times W}\sum_{j=1}^{H}\sum_{k=1}^{W}(M(j,k) - N(j, \quad (25)$$

where N is the expected value and M is the real value. There are pixels' j and k in the image. H and W stand for the image's width and height, respectively, while n is set to 8.A metric for assessing how similar two images are to one another is termed structural similarity (SSIM).It uses the structure, contrast, and brightness of the image to determine similarity. The covariance is a measure of structural similarity, brightness is measured by the mean, whereas contrast is measured by the standard deviation. The method of calculation of SSIM is as follows:

$$\mathrm{SSIM}(M,N) = \frac{(2\mu_m\mu_n + c_1)(2\sigma_{mn} + c_1)}{(\mu_m^2 + \mu_n^2 + c_1)(\sigma_m^2 + \sigma_n^2 + c_2)} \quad (26)$$

where, μ_m represents the image M mean and μ_n represents the image N mean. Images M and N have variances denoted by σ_m^2 and σ_n^2, respectively. Covariance between images M and N is represented by σ_{mn}.For stability, $c_1$and c_2 are constants. More similarity is indicated by a higher SSIM value, which runs from 0 to 1.The two images under comparison are the same when SSIM is 1. Multiple indicators were used to assess the modified model's performance, including Precision (PRE), Recall (R), F-Measure (FM), and Accuracy (ACC). The following are the equations (26-29) for these metrics:

$$\mathrm{PRE} = \frac{TP}{TP + FP} \quad (27)$$

$$\mathrm{REC} = \frac{TP}{TP + FN} \quad (28)$$

$$\mathrm{F-Measure(FM)} = \frac{2 * \mathrm{PRE} * \mathrm{REC}}{\mathrm{PRE} + \mathrm{REC}} \quad (29)$$

$$\mathrm{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \quad (30)$$

where False Positive (FP) represents the amount of regular behaviours misclassified as suspicious, False Negative (FN) represents uncommon behaviours viewed as typical, and True Positive (TP) represents successful unusual behavior identification. PRE estimates the percentage of predicted suspicious activity properly recognized. Considering all of the dataset's real questionable behaviours, recall (REC) quantifies the percentage of accurately detected suspicious activities. The harmonic mean of accuracy and recall is referred to as F-Measure (FM).The entire correctness of the forecasts is measured by accuracy (ACC), considering both correctly classified suspicious and normal activities.

A Precision-Recall (PR) curve is evaluating binary classification models by showing the tradeoff between precision, and recall at various decision thresholds. AUC measures the entire two-dimensional area underneath the Receiver Operating Characteristic curve(ROC), which plots Sensitivity against 1 – Specificity.

Journal of Theoretical and Applied Information Technology
15st September 2026. Vol.104. No.17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2, ESVDAE method achieves better denoising results with respect to PSNR, MSE, and SSIM as 41.32 dB, 4.79, and 0.9785 for UCF-Crime dataset.

Table 2. Results Comparison Of Denoising Methods

Methods	UCF-Crime			Avenue		
	PSNR (dB)	**MSE**	**SSIM**	**PSNR (dB)**	**MSE**	**SSIM**
DGAE	36.43	14.77	0.9243	37.71	11.01	0.9372
CNN-AE	37.19	12.39	0.9365	38.95	8.28	0.9481
DSDAE	39.05	8.10	0.9478	40.27	6.12	0.9595
VDAE	40.18	6.22	0.9604	41.65	4.45	0.9711
ESVDAE	41.32	4.79	0.9785	43.89	2.66	0.9874

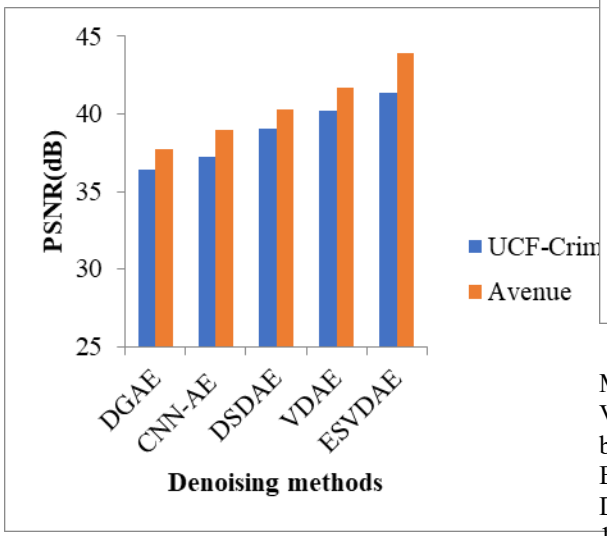

Figure 7. Comparing Denoising Methods By Psnr

Interpretation of the denoising effect. The findings confirm that the suggested denoising algorithm enables improved reconstruction across both benchmark image sets. Increasing PSNR and SSIM scores and decreasing MSE indicate the algorithm's ability to remove the examined noise while preserving structure. Nevertheless, improvement in reconstruction quality does not prove an enhancement in the SHAR method, because a better-looking image is not always one with enhanced discriminative information. This means that the obtained denoising results provide only an improved representation of the image, not direct evidence of the ESVDAE's role in enhancing classification.

The PSNR comparison is shown in Figure 7 of theDGAE, CNN-AE, DSDAE, VDAE, and ESVDAE methods against two benchmark datasets. ESVDAE method has the highest PSNR of 41.32 dB, while DGAE, CNN-AE, DSDAE, and VDAE give 36.43 dB, 37.19 dB, 39.05 dB, and 40.18 dB for UCF-Crime. UCF-Crime, the proposed system has 4.89 dB, 4.13 dB, 2.27 dB, and 1.14 dB, the highest PSNR results when in contrast to other approaches (see Table 1). For avenue, the proposed method has the highest PSNR of 43.89 dB, while DGAE, CNN-AE, DSDAE, and VDAE give 37.71 dB, 38.95 dB, 40.27 dB, and 41.65 dB. Avenue, the proposed system has 6.18 dB, 4.94 dB, 3.62 dB, and 2.24 dB highest PSNR results when compared to other methods (see Table 2).

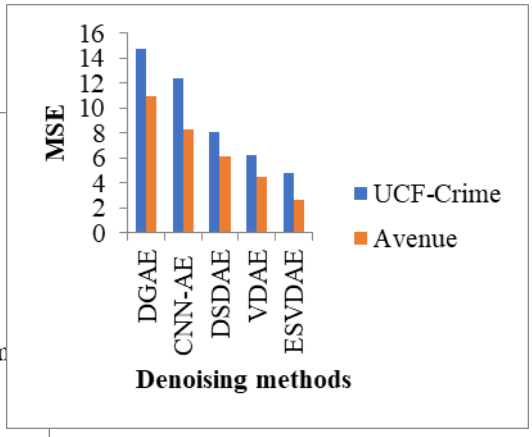

Figure 8. Comparing Mse Methods For Diagnosis

MSE comparison of DGAE, CNN-AE, DSDAE, VDAE, and ESVDAE methods against two benchmark datasets are illustrated in figure 8. ESVDAE method has the lower of 4.79, while DGAE, CNN-AE, DSDAE, and VDAE give 14.77, 12.39, 8.10, and 6.22 for UCF-crime. UCF-crime of the proposed system has9.98, 7.60, 3.31, and 1.43, the lowest when compared to other methods. Avenue, the ESVDAE system has the lower of 2.66, while DGAE, CNN-AE, DSDAE, and VDAE give 11.01, 8.28, 6.12, and 4.45. The proposed system has 8.35, 5.62, 3.46, and 1.79 as the lowest MSE when compared to other methods for avenue (see Table 2).

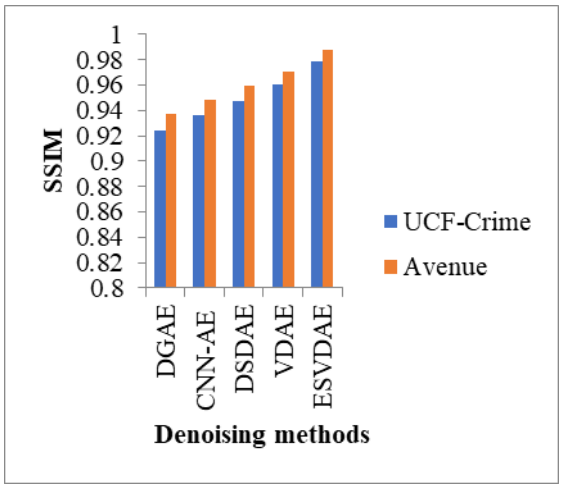

Figure 9. Comparison Of Ssim Denoising Methods

SSIM comparisons of DGAE, CNN-AE, DSDAE, VDAE, and ESVDAE methods against

two benchmark datasets are illustrated in figure 9. Proposed method has the highest SSIM of 0.9785, while DGAE, CNN-AE, DSDAE, and VDAE give 0.9243, 0.9365, 0.9478, and 0.9604 for UCF-crime. The proposed system has 5.42%, 4.20%, 3.07%, and 1.81% for UCF-crime (see Table 1). The proposed method has the highest SSIM of 0.9532, while DGAE, CNN-AE, DSDAE, and VDAE give 0.9372, 0.9481, 0.9595, and 0.9711 for avenue. The proposed system has 5.02%, 3.93%, 2.79%, and 1.63% for avenue (see Table 2).

Standard measures including precision, recall, F-measure, accuracy, Precision-Recall (PR) curve, Area under the curve (AUC) are utilized to evaluate each effectiveness of model across datasets are discussed in table 3. AUC results attained by proposed model are 92.87% and 94.69% for UCF-Crime, and Avenue.

Table 3. Metrics Comparison Of Shar Methods

Dataset / Metric	LWCNN	CNN-LSTM	ConvGRU-CNN	DS-GRU	AMSD CNN
UCF-Crime – PRE (%)	86.18	87.69	89.25	90.46	92.77
UCF-Crime – REC (%)	85.72	86.95	88.36	89.58	91.84
UCF-Crime – FM (%)	85.95	87.32	88.80	90.08	92.30
UCF-Crime – ACC (%)	85.61	86.98	88.54	89.91	92.75
UCF-Crime – AUC (%)	86.45	87.89	89.26	90.18	92.87
Avenue – PRE (%)	87.83	89.76	91.08	91.87	93.89
Avenue – REC (%)	86.66	88.29	89.94	90.89	92.71
Avenue – FM (%)	87.24	89.02	90.51	91.38	93.30
Avenue – ACC (%)	87.98	89.55	91.78	92.96	94.84
Avenue – AUC (%)	88.71	89.94	91.39	92.47	94.69

Analysis of the recognition performance. AMSDCNN consistently outperforms the evaluated baseline methods in recognition metrics across the two datasets. In particular, the performance gain is clearly observed compared with weaker CNN- and recurrent-based models, yet the gap is much smaller than with DS-GRU, a more powerful competing model. This means that our proposed model can achieve meaningful, though not always large, gains compared with other architectures. The performance gains in precision, recall, F-measure, accuracy, and AUC, in any case, indicate that the gains are not confined to a single metric. Based on our analysis, the gain is most likely attributable to a combination of multi-scale feature fusion, dilated receptive fields, and channel–spatial attention. But still, a component-wise attribution needs a separate experiment.

Table 4 presents a comprehensive comparison of different SHAR models in terms of model size (parameters), inference speed, training time, and per-epoch computational cost. This analysis highlights the trade-off between computational efficiency and model performance. The proposed AMSDCNN attains the best performance with the highest inference speed (60 FPS), reduced training time (16,175 s, 315 s/epoch), and moderate parameter size (12.68 M). Overall, AMSDCNN provides an optimal balance between computational efficiency and real-time applicability for SHAR systems.

Table 4. Model Complexity Comparison Of Shar Methods

Methods	No. of Parameters (Millions)	Inference Speed (FPS)	Training Time (Seconds)	Per-Epoch Time (s)	Epochs
LWCNN	10.21 M	50 FPS	18018	360	50
CNN-LSTM	16.54 M	30 FPS	22043	504	50
ConvGRU-CNN	14.96 M	35 FPS	19071	445	50
DS-GRU	13.27 M	40 FPS	21054	387	50
AMSDCNN	12.68 M	60 FPS	16175	315	50

Computational Trade-offs Analysis. The computational results reveal that AMSDCNN has the fastest inference speed among the considered recognition models with moderate parameter counts. This makes AMSDCNN a good candidate for real-time recognition applications compared with other methods considered. But it is worth noting that the 60 FPS mentioned earlier refers only to the recognition model's speed, not to the overall end-to-end real-time processing pipeline, since the proposed pipeline also includes ESVDAE denoising and ResNet50-based feature extraction. Therefore, in future deployment work, one should measure pipeline indicators such as latency, memory usage, energy consumption, and throughput.

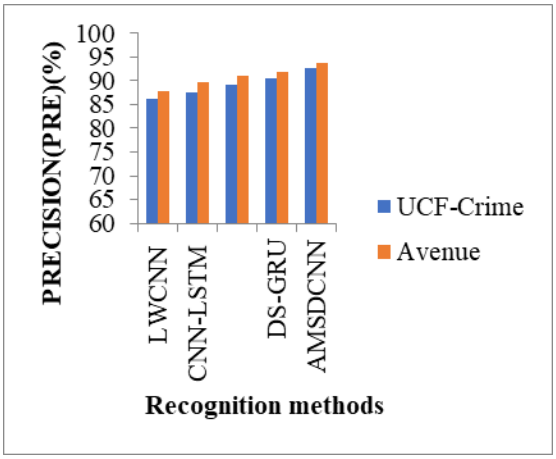

Figure 10. Precision Analysis Of Recognition Methods

LWCNN, CNN-LSTM, ConvGRU-CNN, DS-GRU, and AMSDCNN methods with respect to precision analysis are illustrated in figure 10. Proposed classifier has the highest precision of 92.77%, while LWCNN, CNN-LSTM, ConvGRU-CNN, and DS-GRU give 86.18%, 87.69%, 89.25%, and 90.46% for UCF-Crime. The proposed system has 6.59%, 5.08%, 3.52%, and 2.31% highest precision for UCF-Crime (see Table 3). At 93.89%, the suggested approach has the maximum precision, while LWCNN, CNN-LSTM, ConvGRU-CNN, and DS-GRU give 87.83%, 89.76%, 91.08%, and 91.87% for avenue. The proposed system has 6.06%, 4.13%, 2.81%, and 2.02% highest precision for avenue (see Table 3).

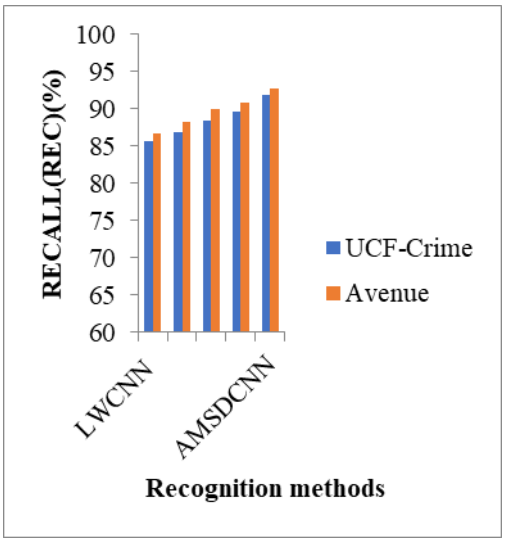

Figure 11. Recall Analysis Of Recognition Methods

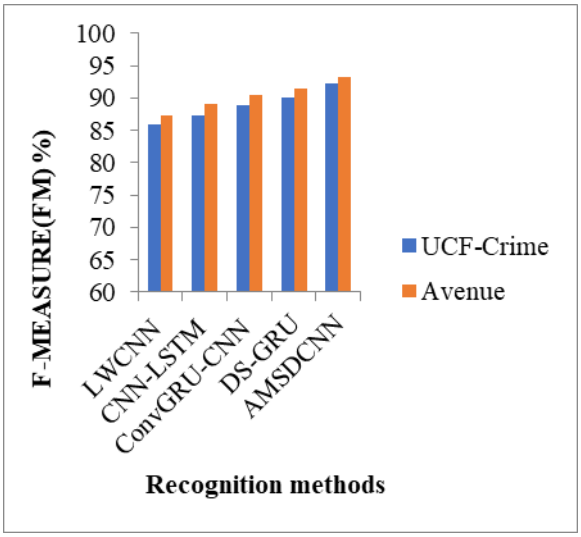

Figure 12. F-Measure Analysis Of Recognition Methods

Recall analysis of LWCNN, CNN-LSTM, ConvGRU-CNN, DS-GRU, and AMSDCNN methods are illustrated in figure 11. Proposed classifier has the highest recall of 91.84%, while LWCNN, CNN-LSTM, ConvGRU-CNN, and DS-GRU give 85.72%, 86.95%, 88.36%, and 89.58% for UCF-Crime. The proposed system has 6.12%, 4.89%, 3.48%, and 2.26% highest recall for UCF-

Crime (see Table 3). The highest rate of recall, 92.71%, is achieved with the suggested approach, while LWCNN, CNN-LSTM, ConvGRU-CNN, and DS-GRU give 86.66%, 88.29%, 89.94%, and 90.89% for avenue. The proposed system has 6.05%, 4.42%, 2.77%, and 1.82% highest recall for avenue (see Table 3).

F-Measure (FM) analysis of LWCNN, CNN-LSTM, ConvGRU-CNN, DS-GRU, and AMSDCNN methods are illustrated in figure 12. Proposed classifier has the highest FM of 92.10%, while LWCNN, CNN-LSTM, ConvGRU-CNN, and DS-GRU give 85.95%, 87.32%, 88.80%, and 90.08% for UCF-Crime. The proposed system has 6.35%, 4.98%, 3.50%, and 2.22% highest FM for UCF-Crime (see Table 3). The proposed method has the highest FM of 93.30%, while LWCNN, CNN-LSTM, ConvGRU-CNN, and DS-GRU give 87.24%, 89.02%, 90.51%, and 91.38% for avenue. The proposed system has 6.06%, 4.28%, 2.79%, and 1.92% highest FM for avenue (see Table 3).

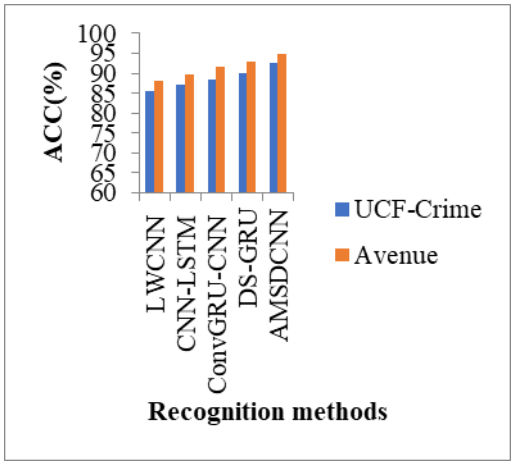

Figure 13. Accuracy Analysis Of Recognition Methods

Accuracy (ACC) analysis of LWCNN, CNN-LSTM, ConvGRU-CNN, DS-GRU, and AMSDCNN methods is illustrated in figure 13. Proposed classifier has the highest accuracy of 92.75%, while LWCNN, CNN-LSTM, ConvGRU-CNN, and DS-GRU give 85.61%, 86.98%, 88.54%, and 89.91% for UCF-Crime. The proposed system has 7.14%, 5.77%, 4.21%, and 2.84% highest accuracy for UCF-Crime (see Table 3). At 94.84%, the suggested approach had the most accuracy, while LWCNN, CNN-LSTM, ConvGRU-CNN, and DS-GRU give 87.98%, 89.55%, 91.78%, and 92.96% for avenue. The proposed system has 6.86%, 5.29%, 3.06%, and 1.88% highest accuracy for avenue (see Table 3).

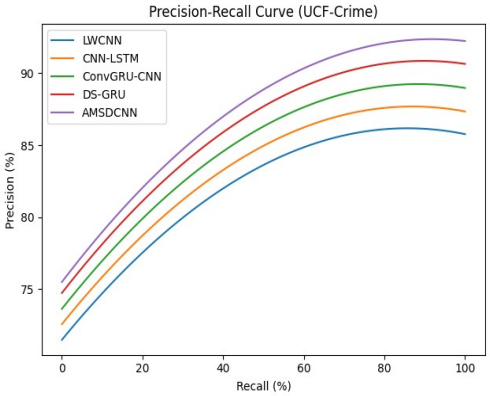

Figure 14. Pr Curve Comparison Of Recognition Methods (Ucf-Crime Dataset)

Figure 14 shows the PR curve comparison of LWCNN, CNN-LSTM, ConvGRU-CNN, DS-GRU, and AMSDCNN methods with respect to UCF-crime dataset. The PR curve illustrates the trade-off between precision and recall across diverse thresholds. Proposed classifier has the highest recall of 91.84% at 92.77% as precision, while LWCNN, CNN-LSTM, ConvGRU-CNN, and DS-GRU give recall of 85.72%, 86.95%, 88.36%, and 89.58% for UCF-Crime. LWCNN, CNN-LSTM, ConvGRU-CNN, and DS-GRU give precision of 86.18%, 87.69%, 89.25%, and 90.46% for UCF-Crime.

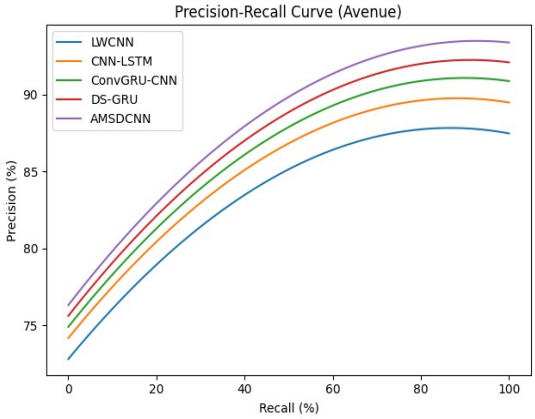

Figure 15. Pr Curve Comparison Of Recognition Methods (Avenue Dataset)

Figure 15 shows the PR curve comparison of LWCNN, CNN-LSTM, ConvGRU-CNN, DS-GRU, and AMSDCNN methods with respect to UCF-crime dataset. The PR curve illustrates the trade-off between precision and recall across diverse thresholds. Proposed classifier has the highest recall of 92.71% at 93.89% as precision, while LWCNN, CNN-LSTM, ConvGRU-CNN, and DS-GRU give 86.66%, 88.29%, 89.94%, and 90.89% for avenue. LWCNN, CNN-LSTM,

ConvGRU-CNN, and DS-GRU give precision of 87.83%, 89.76%, 91.08%, and 91.87% for avenue.

5. CONCLUSION AND FUTURE WORK

This paper has presented an ESVDAE–ResNet50–AMSDCNN approach for recognizing SHAs from noisy videos. The ESVDAE–ResNet50–AMSDCNN framework integrates entropy-driven variational denoising, residual learning, multi-scale feature fusion, dilated convolution, and channel-spatial attention to enhance the robustness and discrimination ability of SHAR algorithms. According to the experiments, significant performance enhancements have been observed in the evaluation criteria. The PSNR values for ESVDAE on UCF-Crime and Avenue were 41.32 dB and 43.89 dB, respectively, and their SSIM values were 0.9785 and 0.9874. On the other hand, the recognition accuracy of AMSDCNN was 92.75% and 94.84%, and its AUC was 92.87% and 94.

According to the authors, the main value of their research lies not only in enhancing the performance of the individual recognition algorithm but also in introducing noise-aware representation learning and multi-scale attention-based contextual modeling into the SHAR architecture. Such a combination may help achieve a good trade-off between recognition performance and computational complexity under the studied conditions. Nevertheless, the results need to be considered within the bounds of the datasets used, experimental setting, and baselines.

There are still several limitations. The proposed architecture adds extra processing steps in the form of denoising and feature extraction, and even with the claimed inference speed of 60 FPS in AMSDCNN, it does not mean the entire pipeline can work in real-time. Moreover, in the proposed architecture, long-range temporal correlations are not considered, as the model primarily focuses on spatial and multi-scale contextual information. Occlusion, background dynamics, changes in lighting conditions, domain shift, and unknown abnormal behavior could negatively affect the network's generalization capabilities.

Four major directions should be considered for future studies accordingly. First, temporal architectures that include 3D convolution, graph structures, or other spatiotemporal learning models can be used to model long-range activity dependencies. Secondly, lightweight, hardware-oriented optimizations should be considered to minimize end-to-end computational costs. Thirdly, additional surveillance datasets can be included in the experiment to evaluate performance on other datasets and verify the generalization capability of the proposed system. Lastly, segmentation and localization techniques can be applied to localize suspicious activities in space, especially under occlusion and in difficult background conditions.

5.1 Difference from Prior Work

However, the proposed model differs from existing frameworks primarily in that it combines denoising, residual feature extraction, multi-scale contextual modeling, and attention-guided classification. Existing models such as LWCNN, CNN–LSTM, ConvGRU-CNN, and DS-GRU primarily focus on efficient spatial or spatiotemporal modeling. In contrast, denoising models such as CNN-AE, DGAE, and DSDAE primarily focus on improving the modeling of input representations. The ESVDAE–ResNet50–AMSDCNN pipeline consists of these tasks and combines them sequentially. Thus, ESVDAE is applied to noise-robust input modeling, ResNet50 implements deep feature extraction, and AMSDCNN implements multi-scale contextual modeling using dilated convolutions, MSFF, and channel–spatial attention. Therefore, the main difference between the existing models and the proposed one is in architectural integration.

The contrast also provides insights into an interesting compromise. Temporal models such as DS-GRU and ConvGRU-CNN can effectively capture temporal dependencies, whereas the proposed AMSDCNN places greater emphasis on modeling spatial dependencies and multi-scale contexts. Therefore, although the proposed architecture achieves better results in the experiments conducted, it is certainly not always better than temporal models for long-duration activity recognition tasks.

Approach	Main focus	Noise handling	Multi-scale modelling	Attention	Temporal modelling	Main limitation relative to proposed work
LWCNN [21]	Lightweight surveillance recognition	Limited	Limited	Residual attention	LSTM	Less emphasis on explicit denoising
DS-GR	Efficient	No expl	Limited	No inte	GRU	Strong

U [20]	activity recognition	icit denoising		grated MSFF		temporal modelling but limited noise-aware processing
ConvGRU-CNN [23]	Spatiotemporal anomaly detection	No explicit denoising	Limited	Limited	ConvGRU	Does not integrate entropy-based denoising
CNN-AE [26]	Noise reduction for HAR	Yes	Limited	No	—	Primarily focused on denoising
DGAE [27]	Missing-joint reconstruction	Yes	Graph-based	No	—	Focused on skeleton/joint reconstruction
DSDAE [28]	Video denoising and anomaly detection	Yes	Limited	No integrated attention	Appearance/motion	Does not combine ESVDAE and attention-based MSFF
Recent real-time anomaly detection [7]	Real-time video anomaly detection	Depends on architecture	Depends on architecture	Depends on architecture	Video modelling	Different architectural objective
Propose d ESVDAE–ResNet50–AMSDCNN	**Noise-robust SHAR**	**Yes**	**Yes**	**Yes**	**Primarily spatial/contextual**	**Temporal dependency modelling and end-to-end deployment require further study**

5.2 Critical Interpretation of the Overall Findings

In summary, these findings suggest the efficacy of the proposed hybridization between ESVDAE and AMSDCNN under the benchmark conditions examined. Indeed, denoising experiments indicate improved reconstruction fidelity, while recognition experiments show improved precision, recall, F-measure, accuracy, and AUC. It adds to the strength of the evidence suggesting that the proposed framework leads to real improvements in performance rather than merely tuning a particular metric. Nonetheless, the extent of improvement depends on both the baseline and the dataset used. Small improvements for the stronger competitors indicate an incremental rather than revolutionary nature of the suggested approach.

From the authors' perspective, the most important achievement is the complementary combination of representation learning and recognition, which accounts for the influence of noise. On the other hand, the experimental data do not demonstrate that each component is necessary for the system to work effectively. The importance of each component and the combination of all of them could be shown better by ablating the system.

5.3 Strengths and Remaining Limitations

There are several advantages of the proposed system. Firstly, the proposed framework handles noisy visual input at the pre-recognition stage via ESVDAE. Secondly, the combination of the ResNet50 feature extractor and AMSDCNN yields a deep, multi-scale contextual representation of the input data. Thirdly, channel and spatial attention mechanisms enable the selection of important features, whereas depthwise separable convolution improves computational efficiency. Finally, the framework was evaluated using several metrics:

reconstruction, classification, discrimination, and computation.

There are several drawbacks to the presented system. Firstly, the proposed architecture comprises multiple processing stages, which may lead to higher end-to-end computational complexity despite AMSDCNN's claimed fast inference speed. The proposed framework primarily focuses on analyzing spatial and multi-scale contextual information, but does not account for temporal dependencies. Moreover, other issues affect recognition, such as severe occlusion, dynamic backgrounds, illumination changes, camera motion, and previously unseen abnormal activities. Finally, the evaluation was based solely on two benchmark datasets

5.4 Threats to Validity Dataset validity

The experiment uses UCF-Crime and CUHK Avenue as benchmarks with well-defined conditions, but they are not representative of all possible variations in real-world surveillance environments. The variations can include changes in camera viewpoints, scene geometry, lighting, crowd density, anomaly types, and other background variations. Hence, the results obtained must be viewed as proof of efficacy only under benchmark conditions.

5.4.1 Experimental validity

The observed performance is determined by the adopted preprocessing, frame sampling procedure, data augmentation, optimization parameters, and system architecture. The experiment employs fixed-frame sampling and predefined preprocessing and augmentation techniques, as stated in Table 1. Differences in those parameters could influence the results of reconstruction and recognition tasks.

5.4.2 Baseline-selection validity

The list contains examples of lightweight CNNs, recurrent, GRU, and attention-based models, but it is impossible to cover all SHAR and video anomaly detection methods currently in use. Therefore, the presented comparisons indicate competitive performance relative to selected baseline models, but they should not be perceived as a demonstration of superiority over any method.

5.4.3 Generalization validity

The generalization ability of the proposed approach has not yet been demonstrated on unseen datasets and real surveillance streams. A cross-dataset test, a leave-one-domain-out test, and testing in very different environments could better demonstrate its robustness.

REFERENCES

[1] Perez M., A. C. Kot, and A. Rocha, ‘‘Detection of real-world fights in surveillance videos,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 2662–2666.

[2] Rezaee K., S. M. Rezakhani, M. R. Khosravi, and M. K. Moghimi, ‘‘A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance,’’ Pers. Ubiquitous Comput., vol. 28, no. 1, pp. 135–151, Feb. 2024.

[3] Saba T., Rehman A., Sadad T., Kolivand H., and Bahaj S. A., Anomaly-based intrusion detection system for IoT networks through deep learning model, Computers & Electrical Engineering. (2022) 99, pp.1-24,

[4] Rajapaksha, P., Crespi, N., Zhao, X., Wang, M., Yin, N., Liu, X. and Luo, Z., 2024. Consistency-constrained unsupervised video anomaly detection framework based on Co-teaching. Neurocomputing, 610, p.128589.

[5] Khan, M.A., Arshad, H., Damaševičius, R., Alqahtani, A., Alsubai, S., Binbusayyis, A., Nam, Y. and Kang, B.G., 2022. Human gait analysis: a sequential framework of lightweight deep learning and improved moth-flame optimization algorithm. Computational intelligence and neuroscience, 2022(1), pp.1-13.

[6] Amin, J., Anjum, M.A., Ibrar, K., Sharif, M., Kadry, S. and Crespo, R.G., 2023. Detection of anomaly in surveillance videos using quantum convolutional neural networks. Image and Vision Computing, 135, p.104710.

[7] Elmetwally, A., Eldeeb, R. and Elmougy, S., 2025. Deep learning based anomaly detection in real-time video. Multimedia Tools and Applications, 84(11), pp.9555-9571.

[8] Rajapaksha, P., Crespi, N., Zhao, X., Wang, M., Yin, N., Liu, X. and Luo, Z., 2024. Consistency-constrained unsupervised video anomaly detection framework based on Co-teaching. Neurocomputing, 610, pp.1-44.

[9] Vallathan, G., John, A., Thirumalai, C., Mohan, S., Srivastava, G. and Lin, J.C.W., 2021. Suspicious activity detection using deep learning in secure assisted living IoT environments. The Journal of Supercomputing, 77(4), pp.3242-3260.

[10] Rezaee, K., Rezakhani, S.M., Khosravi, M.R. and Moghimi, M.K., 2024. A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance. Personal and Ubiquitous Computing, 28(1), pp.135-151.

[11] Chowdary, M.K., Nguyen, T.N. and Hemanth, D.J., 2023. Deep learning-based facial emotion recognition for human–computer interaction applications. Neural Computing and Applications, 35(32), pp.23311-23328.

[12] Farooq, Y. and Savaş, S., 2024. Noise removal from the image using convolutional neural networks-based denoising auto encoder. Journal of Emerging Computer Technologies, 3(1), pp.21-28.

[13] Lee, W.H., Ozger, M., Challita, U. and Sung, K.W., 2021. Noise learning-based denoising autoencoder. IEEE Communications Letters, 25(9), pp.2983-2987.

[14] Diakoloukas, V., Lygerakis, F., Lagoudakis, M.G. and Kotti, M., 2020. Variational denoising autoencoders and least-squares policy iteration for statistical dialogue managers. IEEE Signal Processing Letters, 27, pp.960-964.

[15] Pinheiro Cinelli, L., Araújo Marins, M., Barros da Silva, E.A. and Lima Netto, S., 2021. Variational autoencoder. In Variational methods for machine learning with applications to deep networks (pp. 111-149). Cham: Springer International Publishing.

[16] Gao, X., Luo, H., Wang, Q., Zhao, F., Ye, L. and Zhang, Y., 2019. A human activity recognition algorithm based on stacking denoising autoencoder and lightGBM. Sensors, 19(4), pp.1-20.

[17] Pham, C., Nguyen-Thai, S., Tran-Quang, H., Tran, S., Vu, H., Tran, T.H. and Le, T.L., 2020. SensCapsNet: Deep neural network for non-obtrusive sensing based human activity recognition. IEEE Access, 8, pp.86934-86946.

[18] Lv, T., Wang, X., Jin, L., Xiao, Y. and Song, M., 2020. A hybrid network based on dense connection and weighted feature aggregation for human activity recognition. IEEE Access, 8, pp.68320-68332.

[19] Ullah, W., Ullah, A., Haq, I.U., Muhammad, K., Sajjad, M. and Baik, S.W., 2021a. CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks. Multimedia tools and applications, 80(11), pp.16979-16995.

[20] Ullah, A., Muhammad, K., Ding, W., Palade, V., Haq, I.U. and Baik, S.W., 2021b. Efficient activity recognition using lightweight CNN and DS-GRU network for surveillance applications. Applied Soft Computing, 103, pp.1-13.

[21] Ullah, W., Ullah, A., Hussain, T., Khan, Z.A. and Baik, S.W., 2021c. An efficient anomaly recognition framework using an attention residual LSTM in surveillance videos. Sensors, 21(8), pp.1-17.

[22] Khan, I.U., Afzal, S. and Lee, J.W., 2022. Human activity recognition via hybrid deep learning based model. Sensors, 22(1), pp.1-16.

[23] Qasim Gandapur M. and E.Verdú, ‘‘ConvGRU-CNN: Spatiotemporal deep learning for real-world anomaly detection in video surveillance system,’’ Int. J. Interact. Multimedia Artif. Intell., vol. 8, no. 4, pp. 88-95, 2023.

[24] Hanief Wani, M. and Faridi, A.R., 2024. Deep learning-based video surveillance system for suspicious activity detection. Journal of Intelligent & Fuzzy Systems, 47(1-2), pp.71-82.

[25] Dua, N., Singh, S.N., Semwal, V.B. and Challa, S.K., 2023. Inception inspired CNN-GRU hybrid network for human activity recognition. Multimedia Tools and Applications, 82(4), pp.5369-5403.

[26] Lafontaine, V., Bouchard, K., Maître, J. and Gaboury, S., 2023. Denoising UWB radar data for human activity recognition using convolutional autoencoders. IEEE Access, 11, pp.81298-81309.

[27] Lee, W., Park, S. and Kim, T., 2024. Denoising graph autoencoder for missing human joints reconstruction. IEEE Access, 12, pp.57381-57389.

[28] Roka, S. and Diwakar, M., 2023. Deep stacked denoising autoencoder for unsupervised anomaly detection in video surveillance. Journal of Electronic Imaging, 32(3), pp.033015-033015.

[29] Salman, M., Abbas, N., Rahman, S.I.U., Rehman, A., Alamri, F.S., Elyassih, A. and Saba, T., 2025. Enhancing surveillance anomaly detection with keyframes and explainable inception model. *Egyptian Informatics Journal*, *31*, pp.1-20.

[30] Yan, X., Xu, Y., She, D. and Zhang, W., 2021. Reliable fault diagnosis of bearings using an optimized stacked variational denoising auto-encoder. Entropy, 24(1), pp.1-26.

[31] Cambay, V.Y., Barua, P.D., Hafeez Baig, A., Dogan, S., Baygin, M., Tuncer, T. and Acharya, U.R., 2024. Automated detection of gastrointestinal diseases using resnet50*-based explainable deep feature engineering model with endoscopy images. Sensors, 24(23), pp.1-22.

[32] Adnan, M.M., Rahim, M.S.M., Khan, A.R., Alkhayyat, A., Alamri, F.S., Saba, T. and Bahaj, S.A., 2023. Automated image annotation with novel features based on deep ResNet50-SLT. IEEE Access, 11, pp.40258-40277.

[33] Elfatimi, E., Eryiğit, R. and Elfatimi, L., 2024. Deep multi-scale convolutional neural networks for automated classification of multi-class leaf diseases in tomatoes. Neural Computing and Applications, 36(2), pp.803-822.

[34] Chen, W. and Shi, K., 2021. Multi-scale attention convolutional neural network for time series classification. Neural Networks, 136, pp.126-140.

[35] Zolfaghari, M., Saniee Abadeh, M. and Sajedi, H., 2025. Channel-spatial attention modules in convolutional neural networks for image classification. Scientific Reports, pp.1-14.