**Bulletin of Electrical Engineering and Informatics**
Vol. 15, No. 4, August 2026, pp. 3693~3704
ISSN: 2302-9285, DOI: 10.11591/eei.v15i4.10344


# Enhancing anomaly detection in video surveillance with spatio-temporal enhanced deep associative memory networks

**Kusuma Sriram[1], Kiran Purusotham[2], Vinutha Gurulingaiah Kaelgaerae[2]**
[1]Department of Information Science and Engineering, M. S. Ramaiah Institute of Technology (MSRIT), Affiliated to Visvesvaraya Technological University, Bengaluru, India
[2]Department of Computer Science and Engineering, R. N. Shetty Institute of Technology (RNSIT), Affiliated to Visvesvaraya Technological University, Bengaluru, India

**Article Info**

***Article history:***

Received Mar 25, 2025
Revised Jun 16, 2026
Accepted Jul 14, 2026

***Keywords:***

Deep convolutional covariance networks
Deep learning
Multiple object tracking
Similarity map function
Space-time adaptive correlation tracking

**ABSTRACT**

In the period of extensive video data generation from various surveillance sources, ensuring public safety and security is vital. Detecting unusual crowd behavior is essential, especially in scenarios with large gatherings. However, despite widespread video surveillance, incidents like vehicular accidents, stampedes, and burglaries still occur due to the limitations of traditional surveillance systems. In anomaly detection, finding the critical deviated pattern is the main and critical task. In the context of intelligent video surveillance, automated detection of abnormal behavior is achieved through computer vision analysis, eliminating the need for constant human monitoring. This paper proposes the spatio-temporal enhanced deep associative memory networks (STEAD)-network, a novel approach for anomaly detection in video sequences. The STEAD-network combines various techniques, including spatio-temporal enhancement, associative memory modules, and pattern recognition, to effectively capture and recognize abnormal events. Three benchmark datasets, UCSD Ped2, CUHK Avenue, and ShanghaiTech are used to evaluate this proposed dataset and to compare with existing state-of-the-art techniques. The results demonstrate that the STEAD-network consistently outperforms other methods, achieving significant improvements in anomaly detection accuracy across all datasets. The development of intelligent video surveillance systems is aided by this research by enhancing their ability to autonomously and accurately detect abnormal behavior in real-world scenarios.



***Corresponding Author:***

Kusuma Sriram
Department of Information Science and Engineering, M. S. Ramaiah Institute of Technology (MSRIT)
Affiliated to Visvesvaraya Technological University
Bengaluru, India
Email: kusumas_12@rediffmail.com

## 1. INTRODUCTION

Due to the increasing use of video capture devices, a substantial amount of video data has been produced. This data is crucial for ensuring the security of residents and their properties. It is crucial to recognize unusual crowd behavior, especially when there are big crowds of people. The detection of abnormalities plays a crucial role in ensuring human safety and enhancing social security. Despite the prevalent utilization of video surveillance systems in the context of public safety [1], a significant number of unfortunate incidents, such as vehicular accidents, stampedes, and burglaries, persist on an annual basis. Most surveillance equipment is not equipped with real-time notification capabilities as its main function is to capture and store video footage over extended periods of time. Anomaly detection is a widely recognized

objective in data analysis. It involves the identification of patterns in data that deviate from anticipated behavior [2]. Anomalies, outliers, discordant observations, exceptions, aberrations, surprises, oddities, or contaminants are among the commonly employed terms to describe patterns that deviate from anticipated outcomes in various contexts.

One crucial aspect of modern intelligent video surveillance systems is the automated detection and recognition of abnormal behavior. This is achieved by employing computer vision analysis techniques. The utilization of this technology enhances monitoring efficiency by eliminating the requirement for human supervision of live video streams. In a theoretical context, any atypical occurrences can be perceived as abnormal events. However, empirical evidence reveals that this assumption is invalid when implemented in practical scenarios. Cycling is considered an unconventional behavior within the context of a college campus [3]. The reason for this is that the model employed in this scenario exclusively classifies walking as a habitual task. Due to safety considerations, cycling is generally not considered a deviant pursuit. When compared to other activities, certain anomalous behaviors are more easily identifiable and explicable in real-world scenarios. Multiple instances of violent episodes have been documented [4] for illustrative purposes.

The primary objective of both conventional [5] and deep learning-based methods has been to detect abnormalities in real-time. Although the second group of methods incurs higher computing costs, it frequently produces superior outcomes. The approaches within the first group, however, exhibit superior performance but require a greater allocation of computational resources for execution. In order to facilitate real-time detection, it is imperative to decrease the complexity of the model. Several programs have been implemented to tackle the problem of low complexity. The projects involve the utilization of cascading local and global descriptors [6], the integration of low complexity features instead of high-level semantic features, the implementation of a spatio-temporal auto-encoder network for the automated extraction of abnormal behaviors, and the creation of spatio-temporal cuboids through dynamic temporal feature extraction. The issue of training 3D convolutional neural networks (CNNs) using video datasets [7] through the introduction of a novel deep learning technique. The utilization of 2D CNNs that have undergone training on image data is recommended for optimal gathering of spatial and temporal information. This technique effectively reduces the CPU and memory requirements, enabling the utilization of multiple commercially available 2D CNNs that have been trained on images.

Identifying atypical crowd behavior can pose significant challenges in practice. The realm of everyday life encompasses a wide range of abnormal behaviors, posing a difficulty in formulating specific and accurate descriptions for each unique circumstance. Moreover, instances of anomalous behavior are infrequent in real-life situations. The datasets used for identification often exhibit an imbalance, where there is a higher proportion of normal behavior frames compared to deviant ones. Moreover, it is crucial to accurately determine the specific criteria that delineate abnormal crowd behavior, while also considering individual circumstances. When the traffic light at a junction is in a flashing green signal state, a significant number of pedestrians opt to traverse the street. Nevertheless, it is deemed atypical for pedestrians to partake in identical conduct when the traffic light is exhibiting a flashing red signal. The automatic detection of abnormal activity in video sequences is a critical requirement in surveillance systems. In contrast, traditional approaches are typically associated with extended timelines and substantial resource demands, which often involve a large workforce. An urgent requirement exists for the development of a system that possesses the capability to autonomously detect and identify abnormal behavior [8].

Two different methods for crowd detection are shown in the study: an indirect method and a direct way. The direct approach is a detection method that precisely identifies and distinguishes each individual in a scene by using either segmentation or human detection. The map-based indirect technique involves the identification of visual attributes, which are then correlated with the population through the use of mapping. The application of these methodologies aims to enhance the understanding of crowd behavior and improve pedestrian safety. The machine learning algorithm developed has been specifically designed to evaluate the normality or abnormality of distinct crowd scenarios such as fear, fights, and stampedes. The authors have addressed the issue of mixed behavior problem by employing the label distribution technique. The term “mixed behavior” is employed to denote the simultaneous manifestation of multiple behaviors. As a result, attention is exclusively directed towards a specific task, resulting in the disregard of other behaviors. The simultaneous presence of aggressive behavior and disoriented or fearful behavior can result in a confluence of behavioral patterns [9].

This research is motivated by the need to harness the vast amount of video data generated by surveillance systems for enhanced public safety. With the ubiquity of cameras, there is a pressing demand for automated solutions to detect and respond to abnormal crowd behavior, which can range from accidents to criminal activities. Our study seeks to empower intelligent video surveillance systems with the capability to autonomously identify anomalies in real-time. Through the application of computer vision and deep learning, our goal is to develop a more efficient and precise approach for recognizing unusual events within video

footage. Ultimately, this research aims to advance technology that can significantly enhance security in crowded settings, benefiting society as follows:

- The proposed spatio-temporal enhanced deep associative memory networks (STEAD)-network introduces a novel architecture, by combining a spatio-temporal enhancement module, associative memory, encoders, and a decoder. This design enhances spatio-temporal feature extraction for improved anomaly detection.
- The proposed STEAD-network uses global attention for optical flow effectively distinguishing normal motion from anomalies, especially rapid movements. This enhances anomaly detection accuracy in diverse scenarios.
- The integration of a pattern recognition module (PRM) into STEAD-network enhances its adaptability and specificity. PRM learns and records normal patterns, improving anomaly detection by incorporating various patterns with appropriate weights.

The organization of this paper involves the first section provides a brief introduction that highlights the significance of automated anomaly detection in video surveillance. A review of the literature is done in the section 2, a novel STEAD-network is designed in the section 3 to improve crowd behavior anomaly detection, and a performance evaluation is given in the section 4 that shows the comparison results in the form of a graph and the section 5 gives the conclusion of the work.

## 2. RELATED WORK

Using a decision tree classification algorithm and an analysis of changes in the physical features of moving pedestrians, with an emphasis on acceleration characteristics, they created a real-time fall-down detection system. Data from 100 participants who participated in everyday activities and fall simulations were used by [10] to construct a fall detection system. The system focuses on the study of four different phases of falls and is based on the integration of multi-sensor data. Direkoglu [3] proposed a unique method for identifying aberrant crowd behavior, particularly fear and elopement behavior. CNN and motion information images (MII) are combined in this technique. Variational aberrant behavior detection (VABD) is a probabilistic approach this framework is intended to quickly identify unusual crowd behavior in video clips. Zhang *et al.* [11] suggested anomaly analysis approach is based on an improved k-means algorithm. They have constructed a simplified 2D CNN and used a pre-trained 2D CNN for motion input to get the best recognition accuracy while using the fewest amount of processing resources. Study in [12] included a thorough evaluation of the developments in crowd analysis's physical methodologies. These techniques were divided into three groups: fluid dynamics, interaction forces, and complicated crowd motion systems. Sabokrou *et al.* [13] looked at a poorly supervised framework for locating and identifying aberrant behavior (ABDL) in crowded settings in their study. Bag-of-event-models (BoEM), an effective embedding approach, was introduced by [14] to describe video clips showing both typical and aberrant behaviour. A mechanism in their study for creating synthetic anomalous events that precisely replicate certain abnormal occurrences.

Chong and Tay [15] presented a method based on scene perception that integrates the representation of fluid forces and psychological theory. Using three elements—spatial location estimation, temporal action route searching, and spatial-temporal action compensation—studied a technique for extracting actions from continuous unconstrained films. Deep learning has achieved significant advancements, notably in the fields of target tracking, facial recognition, and several other fields. The user has given a number as a starting point. Meng *et al.* [16] proposed long short-term memory networks (LSTMs) and CNNs are two essential neural network types used in deep learning. Forward and back propagation strategies are used in CNN's training process to modify the weights and thresholds. This is accomplished by feeding labels into the model as its input and video images as its output. Singh *et al.* [17] used cascaded classifiers to successively distinguish between normal and deviant pedestrian behaviour. They successfully identified aberrant areas using cascaded CNNs and cascaded autoencoders. Sikdar and Chowdhury [18] used dual-stream CNNs and the optical flow data from an input picture were used to extract elements of pedestrian behavior. The transmission of information over lengthy input sequences is a challenging task that standard recurrent neural networks are unable to perform. The LSTM neural network was created expressly to address this issue. Farooq *et al.* [19] integrated time-domain and spatial information retrieved by autoencoders with LSTM to produce an anomalous behavior detection model. Spatial-temporal CNNs and LSTM were combined to create a deep learning network that can recognize and categorize pedestrian actions. In order to spot unusual pedestrian behavior, this network was later used.

The majority of study focus has been on the development of unsupervised, training-free methods for identifying non-pedestrians and identifying escape fear behaviors. The unsupervised learning approach [16]-[18] creates rebuilt frames from a composite input made up of raw visual data. The proposed technique takes use of a convolutional autoencoder LSTM model to analyze sequences of edge photos. The model makes advantage of its algorithmic abilities to look for anomalies. The estimation of the reconstruction error

involves multiple steps. The spatial temporal adversarial auto-encoder (ST-AAE) and the spatial temporal convolutional auto-encoder (ST-CAE) in their research. Convolution and deconvolution algorithms are applied in three-dimensional space to enhance pattern identification in the temporal dimensions. Any anomalies are obvious given the context. There are two distinct tasks that must be completed. The initial step of the process involves finding cuboids that exhibit recognizable normal qualities and classifying cuboids that imply potential issues using ST-AAE. This article provides detailed information on cuboids representation in raw 3D format, with a focus on potential video applications. The technique starts with suspicious cuboids, which is the initial stage of the ST-CAE algorithm. The aberrant areas are then correctly located and detected using the approach.

The primary goal of this research endeavor is to create original generative models that are capable of precisely locating and identifying uncommon objects. The foreground is extracted using fully convolutional networks (FCNs) that have undergone training. The estimate of optical flow for motion characteristics is a crucial need. In order to effectively remove normal samples from the data, two neural network models are utilized to manage the data. The issue of faulty reconstruction must be acknowledged and addressed in the present circumstance. The system, designed specifically for this, will only take input that exceeds a certain threshold. A learning method is used to make the judgment regarding this threshold. The main goal of this tool is to precisely locate anomalies in a particular system or dataset. The user has entered the study as the input value. The recommended approach makes use of a non-trained, progressive spatio-temporal model. The user is the one who employs fuzzy logic as a learning tool. The concept of normalcy is continuously redefined by using CNN feature aggregation and active learning techniques. This function facilitates the monitoring of newly found abnormalities. The term “system or entity” refers to a collection of components working together to achieve a certain goal. This system or entity's activities, reactions, and interactions can be utilized to observe and analyze its behavior. Multiple factors, including as a system's or entity's architecture, inputs, environment, and internal operations, can affect how it behaves. Making informed judgments, correctly anticipating a system's performance, and recognizing potential issues all depend on understanding and evaluating a system's or entity's behavior.

## 3. METHOD

Figure 1 illustrates the primary components of our system, comprising two encoders, a spatio-temporal enhancement module, an associative memory module, and a decoder. Four successive video frames and the related optical fluxes are sent to our suggested network. In order to train the network and extract various spatio-temporal features, the feature encoder G and position embedding G are fused with the spatio-temporal enhancement module. The decoder uses the fused features to recover the memory's contents and construct the frame.

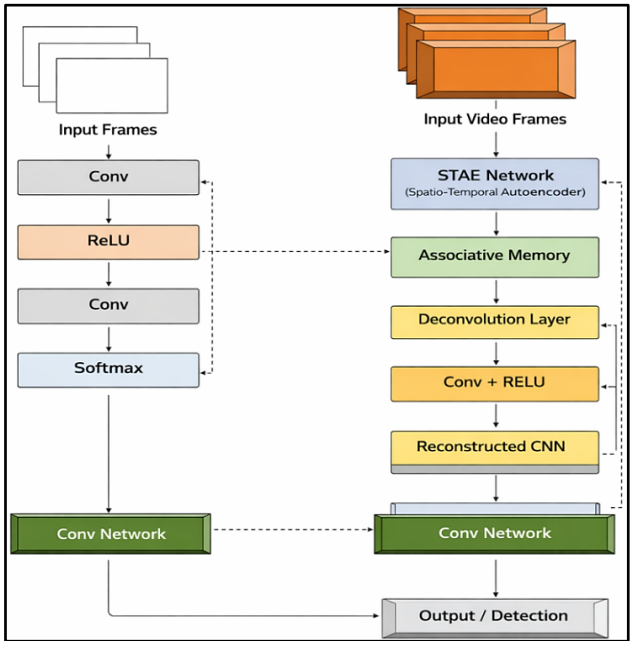



Figure 1. Proposed architecture

### 3.1. Encoder-encoder+decoder model

The spatio-temporal information is completely captured with the information appeared. The encoder-encoder model utilises the spatio-temporal information individually using a decoder for reconstruction of the frame. The encoder architecture as the encoder-decoder with two encoders share within the similar structure. The pooling operation within the stride and convolution operation within the encoder. The reshaped convolutional networks make the shape of the convolutional kernel close to the feature by addition of a direct vector. In order to provide multi-level scaled information for video frame prediction, the features from higher-level and lower-level features are combined to create offsets that support convolutional features, thereby improving prediction accuracy. Additionally, a connection between the feature encoder G_c and decoder is skipped.

### 3.2. Temporal convolutional network

The correlation in between the appearance motion of information for normal events within videos and anomalies relevant to objects moving. Suggesting to be utilized by the global attention map for optical flow of to concentrate on attention features of normal events. The global attention based optical flows a convolution based 1*1 global context modelling is used. The abnormal events record the rapid involvement by introducing a variation mechanism for rapidly moving-objects within the video. The variance-based attention mechanism estimates variance across the spatial dimension subsequently applied to generate the variance attention map as (1):

$$\alpha_{var} = \mu(| \left| \frac{h_o - 1}{F \sum_{f=1}^{F} h_o} \right| |_2^2 \quad (1)$$

where $h_o \in T^{d*f*1}$ shows the optical flow features within this position after global context modelling. Here d shows the batch size and f shows the dimension of that channel. However, $F = j * y$ shows the spatial dimension. The final convolution operation is used to aggregate the global context feature appearance on the basis of the motion representing the features within this position. The attention motion accommodates the connections amidst the crowd movement, whereas stem uses the motion features to enhance its appearance. The magnitude of the motion attention model uses magnitude and angle features within as inputs, and obtains the value of each pixel within the frame by applying softmax multiplication operation. The appearance features for optical flow shown as inputs employ convolution and variance-based attention to impose the variable constraints on the global motion appearance.

### 3.3. Pattern recognition module

The PRM consists of an item memory as $O_k \in T^{p*f}$ which are learned and recorded by the prototypes of normal prototype relations respectively. Here the term p gives the number of items and the term f provides the dimensions of the channel. The updating procedure of reading PRM is shown in:

- Read: the input data is distributed to each memory item using the cosine similarity to retrieve the pertinent memory items. The input feature z, which is represented as $O_k$ and $O_t$, is illustrated as (2) and (3):

$$O_k^1 = O_{k-1} + h_1(z)^V \otimes h_1(z) \quad (2)$$

$$Z_1 = \mu(h_3(z)^V) O_{t-1} h_2(z) \quad (3)$$

where $h_1$, $h_2$, and $h_3$ are the linear layers and $\otimes$ is the cross product.

- $O_t$ updation: however, $O_t$ tracks the relationship within normal prototype patterns within $O_k$, the spatial attention module processes the results $O_k^1$ and $Z_1$ of the PRM, to update the memory within $O_t$, this processes the results $O_k^1$ by addition of the result preceding the memory $O_{t-1}$.

$$O_t^1 = \vartheta(O_k^1 + Z_1 \otimes h_2(z)) \quad (4)$$

$$O_t = O_{t-1} + O_t^1 \quad (5)$$

$$O_o = \text{norm}(Y_n O), 0, s, m \quad (6)$$

$$C^{\otimes}(s, M, X) = \sum_{k=1}^{pmx} \gamma(s * m_k) \otimes x_k \quad (7)$$

$$\vartheta(O) = C^{\otimes}(O_s, O_m, O_x) \quad (8)$$

norm is the normalisation operation, $*$ shows the multiplication and wherein $\gamma$ is the tanh function.

- $O_k$ updation: $O_t$ is updated upon addition of a content read from previous PRM to the result which is processed by spatio-temporal information as shown: $h_4$ show linear layers.

$$O_k = O_k^1 + h_4(O_t^1) \tag{9}$$

- $O_t$ transfer: the information which is semantic is recorded via the $O_t$, converted into an associative memory and addresses in a similar fashion shown as (10) and (11):

$$O = h(O_t) \tag{10}$$

$$z = yO = \sum_{k=1}^{P} y_k o_k \tag{11}$$

$O \in T^{p*f}$ is an associated memory, feed forward network is represented by h and the updated feature using associated memory represented by z, $y_k$ is evaluated as (12) and (13):

$$y_k = \frac{\exp(f(z,o_k))}{\sum_{l=1}^{P} \exp(f(z,o_l))} \tag{12}$$

$$f(z, o_k) = \frac{z o_k^V}{||z||||o_k||} \tag{13}$$

The spatio-temporal information, there exists a variation in between a PRM and a spatio-temporal module. This implementation, within each batch corresponds to the memory indicating the relational memory dimension shown as $d * p * f * f$ and within the dimension of the actual memory as $d * f * f$, wherein d is the batch-size, p is the memory item, whereas in this implementation each batch shares one memory, the memory records the typical prototype patterns. In $O_t$ transfer the spatio-temporal module uses a neural network to, whereas PRM further extends a associative memory integrating various patterns with suitable weights.

### 3.4. Entropy and abnormality score

This training of the proposed model is done by the entropy $N_r$ denoted as the objective function, that minimizes the $n_2$ divergence amid the predicted frame $k_v$ and the corresponding value as $K_v$.

$$N_r = ||K_v - k_v||_2^2 \tag{14}$$

The testing stage involves the evaluation of the proposed technique to score anomalies. The PSNR value in between the predicted frame and the ground truth is given by:

$$X = \sum_{k=0}^{P} x_k \tag{15}$$

$$R(K_v - k_v) = 10\log_{10}(\frac{1}{X}) \tag{16}$$

here $x_k$ depicts the prediction error maximum and the k and P gives the total number of scales denoted by error. The normalization to attain PSNR values in the range of the abnormal value, $U(K_v)$ by Gaussian filter.

$$U(K_v) = \frac{R(K_v - k_v) - \min_v(R(K_v - k_v))}{\max_v(R(K_v - k_v)) - \min_v(R(K_v - k_v))} \tag{17}$$

## 4. PERFORMANCE EVALUATION

Here in this step, the comprehensive evaluation of the proposed methodology carried out on anomaly detection datasets, namely UCSD Ped2, CUHK Avenue, and the ShanghaiTech campus dataset. The methodology's effectiveness and robustness are demonstrated through a detailed analysis, comparing estimated abnormal frames with ground truth labels. The anomaly scores are visually presented, and the area under the curve (AUC) is calculated to assess the system's performance in comparison to existing state-of-the-art techniques.

### 4.1. Preliminary analysis

To evaluate the proposed method on the publicly available anomaly datasets known as UCSD Ped2, CUHK Avenue, and the ShanghaiTech campus dataset. The proposed approach shows its effectiveness and robustness of the model. A comparison is carried out in between the abnormal frames estimated and compared with the ground truth and the detailed analysis is provided. The anomaly score of these samples is visually showcased by utilizing this method. The AUC is evaluated for the proposed system by comparing it with the existing state-of-the-art techniques and hence proving that the proposed system performs efficiently in comparison with the existing system.

### 4.2. Dataset details

In order to assess the effectiveness and generalization ability of the suggested framework, experiments were carried out utilizing various publicly accessible benchmark datasets. These datasets include a range of surveillance situations, types of anomalies, and environmental conditions, allowing for a thorough evaluation of the proposed approach. A concise description of each dataset is presented below.

- UCSD Ped2 dataset: the UCSD Ped2 dataset is a widely recognized resource in the field of computer vision and anomaly detection. It is made up of video clips taken in outdoor settings by fixed security cameras. These videos showcase various scenarios involving pedestrians, including normal activities like walking and jogging, as well as abnormal behaviors like sudden falls or suspicious movements.
- CUHK Avenue dataset: the CUHK Avenue dataset is an important benchmark for evaluating the performance of algorithms in detecting abnormal events in surveillance videos. This dataset includes video clips obtained from multiple outdoor cameras, presenting a variety of scenarios that encompass both normal and abnormal events. The dataset's content ranges from crowded scenes and traffic flows to unexpected occurrences such as vehicle accidents or sudden gatherings.
- ShanghaiTech dataset: for computer vision problems involving crowd counting and crowd density estimates, the ShanghaiTech dataset is an invaluable resource. This dataset, which consists of two separate sections, Part A and Part B, contains photos taken in metropolitan environments. Part A features densely populated scenes, while Part B contains images of sparser environments. Each image within the dataset is annotated with the precise count of individuals present.

### 4.3. Results

In Figure 2, a comparison is carried out between the existing state-of-the-art techniques and the proposed STEAD-network for the UCSD Ped2 dataset, and the result is shown in the form of graph. “MPPCA” at 69.30% already demonstrates a substantial leap in performance, while “Motion Influence Map” and “Unmasking” further improve with AUC scores of 77.30% and 82.20%, respectively. The subsequent methodologies, including “Deep Ordinal Regression,” “Chong,” “Ramachandra,” “Plug and play CNN,” “ConvAE,” “AMDN,” “Stacked RNN,” “ConvLSTM-AE,” “Georgescu,” and “ES,” consistently demonstrate average performance, with AUC scores ranging from 83.20% to 92.90% as tabulated in the Table 1. “Frame-Pred,” “Nguyen,” and “Ionescu”, achieve a higher AUC score of 95.40%, 96.20%, and 97.80%, respectively. Finally, the proposed “STEAD-network [PS] achieves the maximum performance for the AUC score of 99.80%.

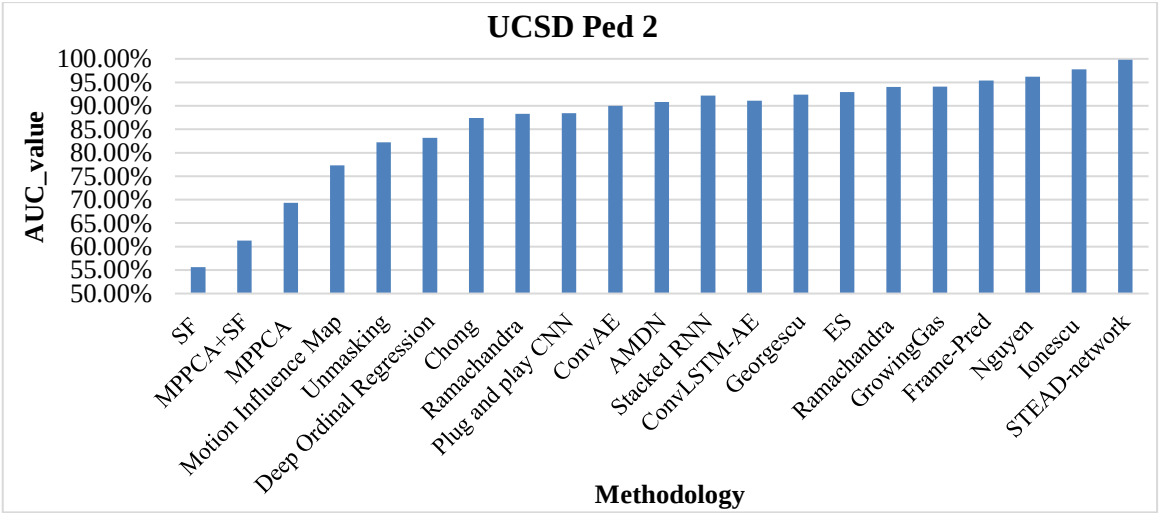



Figure 2. AUC plot for UCSD Ped2 dataset

Table 1. AUC for UCSD Ped2

| Method | AUC (%) |
|---|---|
| SF [20] | 55.60 |
| MPPCA+SF [21] | 61.30 |
| MPPCA [22] | 69.30 |
| Motion Influence Map [23] | 77.30 |
| Unmasking [24] | 82.20 |
| Deep ordinal regression [25] | 83.20 |
| Chong [15] | 87.40 |
| Ramachandra [26] | 88.30 |
| Plug and play CNN [27] | 88.40 |
| ConvAE [28] | 90.00 |
| AMDN [29] | 90.80 |
| Stacked RNN [30] | 92.20 |
| ConvLSTM-AE [31] | 91.10 |
| Georgescu [32] | 92.40 |
| ES [33] | 92.90 |
| Ramachandra [34] | 94.00 |
| GrowingGas [35] | 94.10 |
| Frame-Pred [36] | 95.40 |
| Nguyen and Meunier [37] | 96.20 |
| Ionescu [38] | 97.80 |
| STEAD-network [PS] | 99.80 |

In Figure 3, a comparison is carried out between the existing state-of-the-art techniques and the proposed STEAD-network for CUHK Avenue dataset and the result is shown in the form of graph also AUC values are tabulated in the Table 2. The "ConvAE" methodology depicts an AUC value of 70.20%, "Vu et al" showcases a value of 71.50%, and "Ramachandra et al" at 72.00% represent the baseline performance. "ConvLSTM-AE" demonstrates a significant leap with an AUC score of 77.00%, followed by "Unmasking" and "Chong" at 80.20% and 80.30%, respectively. "ES," "Stacked RNN," and "Frame-Pred" continue this upward trend, with AUC scores ranging from 80.50% to 84.90%. "Georgescu" and "Ramachandra" achieve notably strong performance with AUC scores of 86.90% and 87.20%, respectively. "Ionescu" shows a remarkable AUC score of 90.40%, highlighting its effectiveness in the task. Finally, the "STEAD-network" [PS] as gives an AUC score of 94.30%. to conclude the stead-network performs effectively in comparison with the existing state-of-the-art techniques.

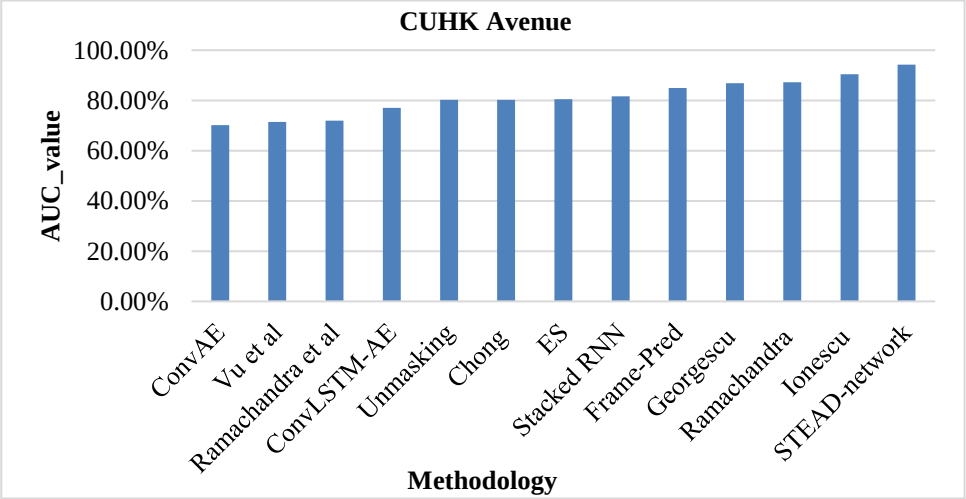



Figure 3. AUC plot CUHK Avenue dataset

In Figure 4, a comparison is carried out between the existing state-of-the-art techniques and the proposed STEAD-network for ShanghaiTech dataset and the result is shown in the form of graph. The methodology "Stacked-RNN" showcases a value of 68.00%, providing a baseline level of performance. "Frame-Pred" showcases a substantial leap with an AUC score of 72.80%, while "Morais" further enhances performance at 73.40%. "ES" provides an average performance with an AUC value of 80.30%,

demonstrating its effectiveness in the task as tabulated in the Table 3. “Georgescu” raises the bar even higher AUC score of 83.50%, indicating its potential utility in anomaly detection scenarios. Finally, the “STEAD-network” [PS] gives a better performance than the existing state-of-the-art-techniques.

Table 2. AUC value for CUHK Avenue dataset

| Method | AUC (%) |
|---|---|
| ConvAE [28] | 70.20 |
| Vu *et al.* [39] | 71.50 |
| Ramachandra [34] | 72.00 |
| ConvLSTM-AE [31] | 77.00 |
| Unmasking [24] | 80.20 |
| Chong [15] | 80.30 |
| ES [33] | 80.50 |
| Stacked RNN [30] | 81.70 |
| Frame-Pred [36] | 84.90 |
| Georgescu [32] | 86.90 |
| Ramachandra [26] | 87.20 |
| Ionescu [38] | 90.40 |
| STEAD-network | 94.30 |

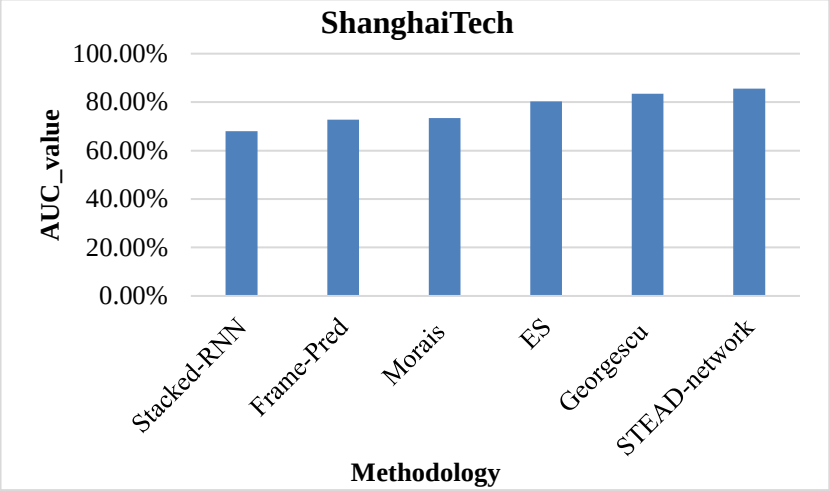



Figure 4. AUC plot for ShanghaiTech dataset

Table 3. AUC value for ShanghaTech dataset

| Method | AUC (%) |
|---|---|
| Stacked-RNN [37] | 68.00 |
| Frame-Pred [20] | 72.80 |
| Morais [40] | 73.40 |
| ES [33] | 80.30 |
| Georgescu [32] | 83.50 |
| STEAD-network | 85.60 |

### 4.4. Comparative analysis

Based on the performance of two approaches, ES and PS, improvement, a comparison analysis of three datasets—UCSD Ped2, CUHK Avenue, and ShanghaiTech as tabulated in the Table 4 offers insightful information about how well these approaches work with various datasets. In the context of UCSD Ped2, both ES and PS exhibit strong performance, with PS outperforming ES significantly with an AUC score of 99.80% compared to 92.90%. This translates to a substantial improvement of 7.16% with PS. Moving to CUHK Avenue, PS again demonstrates better performance with an AUC score of 90.30%, whereas ES achieves 80.50%, this depicts an improvement of 11.48% with PS. Finally, in the case of the ShanghaiTech dataset, both methodologies perform well, with PS slightly edging ahead at 85.60% compared to ES at 80.30%. This results in a respectable 6.39% improvement with PS. Overall, our results show that PS consistently performs better than ES in all three datasets, highlighting its efficacy in various anomaly detection contexts.

Table 4. Comparative analysis

| Dataset | ES (%) | PS (%) | Improvisation (%) |
|---|---|---|---|
| UCSD Ped2 | 92.90 | 99.80 | 7.16139 |
| CUHK Avenue | 80.50 | 90.30 | 11.4754 |
| ShanghaiTech | 80.30 | 85.60 | 6.38939 |

## 5. CONCLUSION

This paper introduces the spatio-temporal enhanced associative memory network (STEAD-network), a novel approach for anomaly detection in video sequences. The STEAD-network incorporates various components, including feature encoders, spatio-temporal enhancement modules, associative memory modules, and decoders, to effectively capture and identify anomalies in surveillance videos. Extensive evaluations on benchmark datasets such as UCSD Ped2, CUHK Avenue, and the ShanghaiTech campus dataset demonstrate the system's remarkable performance and robustness. The STEAD-network consistently outperforms existing state-of-the-art techniques, achieving better AUC scores across all datasets. This research contributes to valuable advancements to computer vision and video surveillance, offering a effective solution for enhancing anomaly detection in real-world applications, thereby strengthening safety and security measures.

## ACKNOWLEDGMENTS

We want to sincerely thank everyone who has helped and supported this study endeavor. First and foremost, we would want to express our sincere gratitude to our guide for his constant direction, priceless insights, and support during the research process.

## FUNDING INFORMATION

No funding is raised for this research.

## AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

| Name of Author | C | M | So | Va | Fo | I | R | D | O | E | Vi | Su | P | Fu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Kusuma Sriram | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | ✓ | |
| Kiran Purusotham | ✓ | ✓ | | | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | | |
| Vinutha Gurulingaiah | | ✓ | | | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | | | |
| Kaelgaerae | | | | | | | | | | | | | | |

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

## CONFLICT OF INTEREST STATEMENT

Author declares no conflict of interest.

## DATA AVAILABILITY

Dataset is utilized in this research mentioned in reference [23].

## REFERENCES

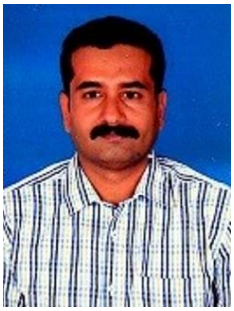
[1] N. Li, H. Yi-bin, and H. Zhang-qin, "Implementation of a real-time fall detection algorithm based on body's acceleration," *Journal of Chinese Computer Systems*, vol. 33, no. 11, pp. 2410–2413, 2012.
[2] D. Pan, H. Liu, D. Qu, and Z. Zhang, "Human Falling Detection Algorithm Based on Multisensor Data Fusion with SVM," *Mobile Information Systems*, vol. 2020, pp. 1–9, Oct. 2020, doi: 10.1155/2020/8826088.

[3] C. Direkoglu, "Abnormal Crowd Behavior Detection Using Motion Information Images and Convolutional Neural Networks," *IEEE Access*, vol. 8, pp. 80408–80416, 2020, doi: 10.1109/ACCESS.2020.2990355.
[4] J. Li, Q. Huang, Y. Du, X. Zhen, S. Chen, and L. Shao, "Variational Abnormal Behavior Detection With Motion Consistency," *IEEE Transactions on Image Processing*, vol. 31, pp. 275–286, 2022, doi: 10.1109/TIP.2021.3130545.
[5] S. Guo, Q. Bai, S. Gao, Y. Zhang, and A. Li, "An Analysis Method of Crowd Abnormal Behavior for Video Service Robot," *IEEE Access*, vol. 7, pp. 169577–169585, 2019, doi: 10.1109/ACCESS.2019.2954544.
[6] A. Mehmood, "Efficient Anomaly Detection in Crowd Videos Using Pre-Trained 2D Convolutional Neural Networks," *IEEE Access*, vol. 9, pp. 138283–138295, 2021, doi: 10.1109/ACCESS.2021.3118009.
[7] X. Hu *et al.*, "A weakly supervised framework for abnormal behavior detection and localization in crowded scenes," *Neurocomputing*, vol. 383, pp. 270–281, Mar. 2020, doi: 10.1016/j.neucom.2019.11.087.
[8] X. Zhang, Q. Yu, and H. Yu, "Physics Inspired Methods for Crowd Video Surveillance and Analysis: A Survey," *IEEE Access*, vol. 6, pp. 66816–66830, 2018, doi: 10.1109/ACCESS.2018.2878733.
[9] C. S., D. K., and V. L.K.P., "Bag-of-Event-Models based embeddings for detecting anomalies in surveillance videos," *Expert Systems with Applications*, vol. 190, p. 116168, Mar. 2022, doi: 10.1016/j.eswa.2021.116168.
[10] W. Lin, J. Gao, Q. Wang, and X. Li, "Learning to detect anomaly events in crowd scenes from synthetic data," *Neurocomputing*, vol. 436, pp. 248–259, May 2021, doi: 10.1016/j.neucom.2021.01.031.
[11] X. Zhang, D. Ma, H. Yu, Y. Huang, P. Howell, and B. Stevens, "Scene perception guided crowd anomaly detection," *Neurocomputing*, vol. 414, pp. 291–302, Nov. 2020, doi: 10.1016/j.neucom.2020.07.019.
[12] R. Nayak, U. C. Pati, and S. K. Das, "A comprehensive review on deep learning-based methods for video anomaly detection," *Image and Vision Computing*, vol. 106, p. 104078, Feb. 2021, doi: 10.1016/j.imavis.2020.104078.
[13] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette, "Deep-Cascade: Cascading 3D Deep Neural Networks for Fast Anomaly Detection and Localization in Crowded Scenes," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1992–2004, Apr. 2017, doi: 10.1109/TIP.2017.2670780.
[14] T. M. Yang, Z. Chen, and W. J. Yue, "Spatio-temporal two-stream human action recognition model based on video deep learning," *Journal of Computer Applications*, vol. 38, no. 3, pp. 895–899, 2018, doi: 10.11772/j.issn.1001-9081.2017071740.
[15] Y. S. Chong and Y. H. Tay, "Abnormal Event Detection in Videos Using Spatiotemporal Autoencoder," in International *Symposium on Neural Networks*, 2017, pp. 189–196, doi: 10.1007/978-3-319-59081-3_23.
[16] B. Meng, X. Liu, and X. Wang, "Human action recognition based on quaternion spatial-temporal convolutional neural network and LSTM in RGB videos," *Multimedia Tools and Applications*, vol. 77, no. 20, pp. 26901–26918, Oct. 2018, doi: 10.1007/s11042-018-5893-9.
[17] G. Singh, R. Kapoor, and A. Khosla, "Optical Flow-Based Weighted Magnitude and Direction Histograms for the Detection of Abnormal Visual Events Using Combined Classifier," *International Journal of Cognitive Informatics and Natural Intelligence*, vol. 15, no. 3, pp. 12–30, Jul. 2021, doi: 10.4018/IJCINI.20210701.oa2.
[18] A. Sikdar and A. S. Chowdhury, "An adaptive training-less framework for anomaly detection in crowd scenes," *Neurocomputing*, vol. 415, pp. 317–331, Nov. 2020, doi: 10.1016/j.neucom.2020.07.058.
[19] M. U. Farooq, M. N. M. Saad, and S. D. Khan, "Motion-shape-based deep learning approach for divergence behavior detection in high-density crowd," *The Visual Computer*, vol. 38, no. 5, pp. 1553–1577, May 2022, doi: 10.1007/s00371-021-02088-4.
[20] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 935–942, doi: 10.1109/CVPR.2009.5206641.
[21] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2010, pp. 1975–1981, doi: 10.1109/CVPR.2010.5539872.
[22] J. Kim and K. Grauman, "Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 2921–2928, doi: 10.1109/CVPR.2009.5206569.
[23] D.-G. Lee, H.-I. Suk, S.-K. Park, and S.-W. Lee, "Motion Influence Map for Unusual Human Activity Detection and Localization in Crowded Scenes," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 10, pp. 1612–1623, Oct. 2015, doi: 10.1109/TCSVT.2015.2395752.
[24] R. T. Ionescu, S. Smeureanu, B. Alexe, and M. Popescu, "Unmasking the Abnormal Events in Video," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2914–2922, doi: 10.1109/ICCV.2017.315.
[25] G. Pang, C. Yan, C. Shen, A. van den Hengel, and X. Bai, "Self-Trained Deep Ordinal Regression for End-to-End Video Anomaly Detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 12170–12179, doi: 10.1109/CVPR42600.2020.01219.
[26] B. Ramachandra and M. J. Jones, "Street Scene: A new dataset and evaluation protocol for video anomaly detection," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2020, pp. 2558–2567, doi: 10.1109/WACV45572.2020.9093457.
[27] M. Ravanbakhsh, M. Nabi, H. Mousavi, E. Sangineto, and N. Sebe, "Plug-and-Play CNN for Crowd Motion Analysis: An Application in Abnormal Event Detection," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2018, pp. 1689–1698, doi: 10.1109/WACV.2018.00188.
[28] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning Temporal Regularity in Video Sequences," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 733–742, doi: 10.1109/CVPR.2016.86.
[29] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," *Computer Vision and Image Understanding*, vol. 156, pp. 117–127, Mar. 2017, doi: 10.1016/j.cviu.2016.10.010.
[30] W. Luo, W. Liu, and S. Gao, "A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 341–349, doi: 10.1109/ICCV.2017.45.
[31] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional LSTM for anomaly detection," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, Jul. 2017, pp. 439–444, doi: 10.1109/ICME.2017.8019325.
[32] M.-I. Georgescu, A. Barbalau, R. T. Ionescu, F. Shahbaz Khan, M. Popescu, and M. Shah, "Anomaly Detection in Video via Self-Supervised and Multi-Task Learning," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 12737–12747, doi: 10.1109/CVPR46437.2021.01255.
[33] S. Zhang *et al.*, "Influence-Aware Attention Networks for Anomaly Detection in Surveillance Videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5427–5437, Aug. 2022, doi: 10.1109/TCSVT.2022.3148392.
[34] B. Ramachandra, M. J. Jones, and R. R. Vatsavai, "Learning a distance function with a Siamese network to localize anomalies in videos," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2020, pp. 2587–2596, doi: 10.1109/WACV45572.2020.9093417.

[35] Q. Sun, H. Liu, and T. Harada, "Online growing neural gas for anomaly detection in changing surveillance scenes," *Pattern Recognition*, vol. 64, pp. 187–201, Apr. 2017, doi: 10.1016/j.patcog.2016.09.016.
[36] W. Liu, W. Luo, D. Lian, and S. Gao, "Future Frame Prediction for Anomaly Detection - A New Baseline," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 6536–6545, doi: 10.1109/CVPR.2018.00684.
[37] T. N. Nguyen and J. Meunier, "Anomaly Detection in Video Sequence With Appearance-Motion Correspondence," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 1273–1283, doi: 10.1109/ICCV.2019.00136.
[38] R. T. Ionescu, F. S. Khan, M.-I. Georgescu, and L. Shao, "Object-Centric Auto-Encoders and Dummy Anomalies for Abnormal Event Detection in Video," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 7834–7843, doi: 10.1109/CVPR.2019.00803.
[39] H. Vu, T. D. Nguyen, T. Le, W. Luo, and D. Phung, "Robust Anomaly Detection in Videos Using Multilevel Representations," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 5216–5223, Jul. 2019, doi: 10.1609/aaai.v33i01.33015216.
[40] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh, "Learning Regularity in Skeleton Trajectories for Anomaly Detection in Videos," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 11988–11996, doi: 10.1109/CVPR.2019.01227.

## BIOGRAPHIES OF AUTHORS

**Dr. Kusuma Sriram** received a B.E. degree from VTU in 2004, and an M.Tech. degree from RVCE in the year 2010. She is an active academician having 20+ years of teaching experience. She is currently an Assistant Professor in the Department of Information Science and Engineering at Ramaiah Institute of Technology, Bengaluru. She pursued a Ph.D. in the area of video processing and machine learning. She is a trained faculty of Infosys InfyTQ to motivate students for Infosys job opportunities. She has worked as trainer for Infosys Campus Connect Program. She can be contacted at email: kusumas_12@rediffmail.com.



**Dr. Kiran Purusotham** is currently working as Head of the Department (HOD) of the Department of Computer Science and Engineering, with total experience of 23+ years. He has done research on privacy preserving data mining with focus on detection of sensitivity patterns and was awarded Ph.D. from VTU in 2014. His research interests include cryptography, randomization, anonymization methods for generalization, indexing techniques, and design patterns. He has guided several M.Tech. and B.E. projects and internships and is currently guiding four research scholars at VTU. He can be contacted at email: kiranpmys@gmail.com.

**Vinutha Gurulingaiah Kaelgaerae** is working in the Department of Information Science and Engineering, with over 12 years of teaching experience at RNS Institute of Technology (2014–present) and Global Academy of Technology (2011–2012). She holds a B.E. and an M.Tech. degree and is currently pursuing a Ph.D. from VTU. Her areas of expertise include data structures, artificial intelligence, DBMS, Python, Java, web technologies, and software engineering. Her research interests include machine learning and deep learning. She can be contacted at email: vinutha.gk@rnsit.ac.in.