Enhancing Video Surveillance and Anomaly Detection with Deep Learning Solutions in Dynamic Environments

Muhammad Arshad Farooq[1], Khalid Mahmood[1], Nasir Saleem[2]

[1]Institute of Computational Intelligence, Faculty of Computing, Gomal University 29220, D.I.Khan, Pakistan
[2]Facutly of Engineering and Technology, Gomal University 29220, D.I.Khan, Pakistan.
Email: arshadfq@gmail.com

Abstract: This research fills the gaps of Convolutional Neural Networks (CNNs) for timely detection of dynamic abnormalities by presenting a novel video anomaly detection model integrating ConvGRU and CNNs and Gated Recurrent Units (GRUs). In order to offer efficient handling of spatial and temporal data in films, temporal dependencies are modeled by ConvGRU while spatial features are learned by ResNet50. By employing motion analysis and entropy filtering to detect focused frames, the proposed solution significantly reduces computation expenses. In preparing to generate anomaly probabilities for classification, the model structure initially extracts space features with ResNet50, followed by time features with ConvGRU. The model does a very good job, as indicated by its performance on the UCF Crime dataset, with 300 movies across five actions. Its accuracy on the validation set is 95.12%, its validation loss is 0.2103, and its Area Under the Curve (AUC) score is 0.9823. Moreover, at a cost of 2.10×10^{11} FLOPs, the COD-3D ResNet model also beats other high-end models, such as P3D and Q3D, when it comes to classifying correctly as well as for computing. Especially in detecting gunfire (94% recall) and assault (99% accuracy), the model is very accurate and recalls well. Overall, the proposed hybrid architecture offers an extremely successful and computationally efficient real-time video anomaly detection system suitable for surveillance.

Keywords: Convolutional Neural Networks (CNNs), Video Analysis, Spatial Features, ResNet50, UCF Crime Dataset.

1. Introduction

Video surveillance systems have increasingly become a crucial tool for traffic management, infrastructure security, and public protection across various industries. These systems are strategically installed in locations such as airports, public buildings, industrial complexes, and commercial establishments to enhance security (Myagmar-Ochir and Kim, 2023). The primary goals of surveillance systems are to protect individuals and assets, improve operational efficiency, and deter criminal activities. However, managing the vast amount of data generated in dynamic and unpredictable real-world settings presents a significant challenge for conventional video monitoring systems, which often rely on pre-programmed rules or constant human monitoring (Power et al., 2021). The inadequacies of traditional systems, such as their inability to assess complex patterns and their vulnerability to operator fatigue leading to human error, underscore the growing need for advanced solutions. The recent breakthroughs in deep learning have provided efficient solutions to these challenges (Mehmood et al., 2023). Through sophisticated algorithms, deep learning has enabled video surveillance systems to automatically detect abnormalities and unusual behaviors that deviate from normal patterns. As noted by Wu et al. (2024), deep learning is particularly well-suited for dynamic environments, where behaviors continually change, thanks to its ability to conduct real-time, context-based analysis. Anomaly detection, the process of identifying abnormal activities or events, plays a significant role in contemporary video surveillance systems. These anomalies may include irregular behaviors, discarded objects, and

disturbances within crowds. The complexity and unpredictability of real-world scenarios often surpass the capabilities of traditional anomaly detection methods, which rely on rule-based systems and hand-crafted features (Roka et al., 2023). Furthermore, these systems may struggle to adapt to unanticipated circumstances, leading to missed incidents or false alarms. In contrast, deep learning models possess the ability to learn and adjust automatically to novel patterns, improving the system's detection of context-dependent anomalies, such as violent offenses or robberies (Chinnasamy et al., 2025). Despite the promise of deep learning for anomaly detection in video surveillance, several challenges remain. One prominent issue is the highly situational and subjective definition of "unusual" behavior. Additionally, training effective models for video surveillance systems is difficult due to the imbalance in datasets, with normal instances far outnumbering abnormal ones (Sharif et al., 2025). Furthermore, real-time processing is essential to detect and address anomalies promptly, but this requires high computational efficiency, as well as the capacity to adapt to factors such as crowd density, lighting conditions, and weather changes (Pathirannahalage et al., 2024).

The aim of this study was to propose a robust framework for video anomaly detection that integrates the latest deep learning techniques to address these challenges. The proposed architecture seeks to enhance the accuracy and efficiency of surveillance systems through the integration of Convolutional Neural Networks (CNNs) for spatial modeling, Recurrent Neural Networks (RNNs) for temporal modeling, and advanced models such as Conv-GRU (Cui et al., 2024). The goal is to provide a real-time, scalable solution capable of detecting and responding to anomalous events in diverse contexts, thereby improving operational and security outcomes. This study aims to improve video surveillance systems by offering a more effective solution to detect and respond to anomalies in real-time. To overcome current limitations and ensure more reliable and adaptable monitoring in dynamic environments, the research will explore how deep learning models can be trained on various datasets. The findings of this study have the potential to significantly advance the field of video surveillance by providing valuable insights into enhancing public safety and security infrastructure. A critical research gap in video surveillance and anomaly detection is the challenge of real-time anomaly detection in dynamic environments, where factors such as crowd density, lighting conditions, and weather changes significantly affect detection accuracy. Developing deep learning models that can efficiently process data in real-time, adapt to these environmental factors, and promptly detect anomalies is a critical area that needs further improvement.

This study makes several key contributions to the field of video surveillance and anomaly detection. It proposes a novel deep learning framework that integrates CNNs for spatial modeling and RNNs for temporal modeling, significantly enhancing the ability to detect anomalies in dynamic environments. The framework aims to provide real-time, scalable anomaly detection, ensuring that surveillance systems can quickly detect and respond to abnormal events, thereby improving operational efficiency. Additionally, the study addresses the challenge of context-aware anomaly detection by adapting deep learning models to recognize and respond to both familiar and novel patterns of unusual behavior, such as violent incidents or crowd disruptions. The research also explores how deep learning models can be trained to handle environmental challenges, such as varying crowd density, lighting conditions, and weather changes, ensuring reliable detection in diverse real-world settings. Furthermore, the study investigates methods for training deep learning models on imbalanced datasets, where normal instances far outweigh anomalies, to improve detection accuracy without bias toward more frequent events. Ultimately, the findings aim to enhance video surveillance systems by providing a more effective and adaptable solution for detecting and responding to anomalies, which can contribute to improved public safety and security infrastructure.

This study aims at designing a resilient video stream anomaly detection framework based on the fusion of powerful deep architectures such as ResNet50 and Conv-GRU. The key aspects of this work are as follows:

- To develop a deep learning framework that Leverages both spatial and temporal information within video to achieve better accuracy for video surveillance.
- To develop a model that overcomes the shortcomings and limitations of existing approaches and scales well with large surveillance networks.
- Deliver real-time detection in order to ensure timely responses. This study addresses these challenges to contribute to the development of smart, reliable surveillance systems for a wide range of applications, including industrial monitoring, traffic management, and public safety.

1.1 Key Points to Research Contribution

- A new framework was proposed that incorporates ResNet50 for spatial feature learning and Conv-GRU for modeling temporal dynamics with the aim of achieving maximum detection accuracy.
- Integrated spatial and sequential/temporal data from video frames to enhance the understanding of the model with complex scenes over standard single-modality models.
- Designed the architecture to scale effectively with large surveillance systems considering the challenges of handling massive video data.
- In high-stakes surveillance scenarios, low-latency processing is the target, enabling speedy anomaly detection and prompt response.
- Designed a system that can handle changes like varying lighting, crowds, and changing backgrounds, thereby being suitable for deployment in the real world.
- Identified key shortcomings in previous anomaly detection models, such as overfitting, slow inference times, and a lack of generalizability across surveillance contexts.

Section 1 provides an introduction to anomaly detection, and then examines the proposed solution's objectives, scope, and contributions. Section 2 examines existing video surveillance and anomaly detection methods, which extend from old approaches to recent deep learning architectures. Section 3 explains the research methodology, which covers model design and architecture and ResNet50 and Conv-GRU. Section 4 describes performance analysis and quantitative measures with their findings and their implications are discussed at the end of this section. Finally, Section 5 provides the principal findings, contributions, limitations, and future work of this study.

2. Literature Review

The recent advancements in deep learning have significantly impacted video surveillance and anomaly detection systems. Several studies have contributed to improving detection accuracy, model efficiency, and adaptability to dynamic environments.

Mehmood et al. (2023) explored the application of deep learning techniques for anomaly detection in video surveillance, highlighting the transition from traditional rule-based methods to data-driven models. Their work focuses on the use of CNNs for feature extraction and anomaly classification. Likewise, Wu et al. (2024) emphasized the use of deep learning models for real-time anomaly detection in dynamic environments, investigating the effectiveness of CNNs and RNNs in handling spatiotemporal data, especially in environments where behaviors constantly change, such as crowded public spaces. Sharif et al. (2025) addressed the issue of imbalanced datasets in video surveillance systems, where normal activities significantly outnumber anomalies. They propose techniques like data augmentation and transfer learning to improve model robustness in detecting rare anomalies. The effectiveness of deep learning models, particularly CNNs and Long Short-Term Memory (LSTM) networks, for detecting violent incidents and other critical events in video feeds was examined by Chinnasamy et al. (2025). They focused on both spatial and temporal aspects of anomaly detection. Meanwhile, Pathirannahalage et al. (2024) investigated the challenges of real-time anomaly detection in video surveillance systems, proposing a deep learning model that balances computational efficiency with accuracy, enabling real-time processing in environments affected by factors like lighting and weather. An advanced anomaly detection framework combining CNNs with Gated Recurrent Units (GRUs) was introduced by Cui et al. (2024). This hybrid model captures both spatial and temporal patterns, addressing the dynamic nature of real-world surveillance environments.

Table 2.1: This table presents a detailed summary of related work relevant to this study

Author(s)	Year	Focus	Key Contribution
Mehmood et al.	2023	Deep learning for anomaly detection in videos	The transition from rule-based methods to CNNs for feature extraction and anomaly classification.
Wu et al.	2024	Real-time anomaly detection in dynamic environments	Investigates the use of CNNs and RNNs for spatiotemporal data to handle continuously shifting behaviors.
Sharif et al.	2025	Handling imbalanced datasets	Proposes methods like data augmentation and transfer learning for handling imbalanced datasets in anomaly detection.
Roka et al.	2023	Critique of traditional anomaly detection methods	Highlights the shortcomings of rule-based and hand-designed feature systems, advocating for deep learning-based anomaly detection.
Chinnasamy et al.	2025	Detection of violent incidents and critical events	Combines CNNs and LSTMs to focus on both spatial and temporal aspects of anomaly detection in videos.

Power et al.	2021	Limitations of traditional video surveillance systems	Identifies the need for deep learning solutions in dynamic environments that traditional systems cannot handle.
Pathirannahalage et al.	2024	Real-time anomaly detection	Proposes a deep learning model that balances computational efficiency with accuracy for real-time video surveillance.
Cui et al.	2024	Hybrid CNN-GRU models for anomaly detection	Integrates CNNs with Gated Recurrent Units (GRUs) to capture both spatial and temporal patterns in dynamic surveillance environments.
Liu et al.	2023	Attention-based models for anomaly detection	Introduces an attention-based model to focus on key regions of interest in video feeds for improved anomaly detection.
Zhang et al.	2023	Deep learning and edge computing for surveillance	Explores the use of edge computing to reduce latency and improve the responsiveness of anomaly detection systems in large-scale networks.
Zhou et al.	2024	Unsupervised anomaly detection	Focuses on unsupervised deep learning models for anomaly detection in scenarios with limited labeled data.
Sun et al.	2023	Generative adversarial networks (GANs) in anomaly detection	Demonstrates how GANs can generate realistic video frames for training anomaly detection models with limited labeled data.

3. Methodology:

This method combined CNNs and ConvGRUs to overcome CNN shortcomings in video dynamic anomaly detection. CNNs captured geographic attributes but not temporal relationships, which ConvGRUs did with improved anomaly detection. ConvGRUs maintained spatial structure when handling temporal data and were computationally less expensive than ConvLSTMs for real-time processing. Frames were equidistantly sampled and normalized, and ResNet50 output 128-channel feature maps. ConvGRUs captured temporal relationships and developed feature representations from updating, reset, and mixture gates. Dimensionality reduction used max-pooling, while a fully connected layer with sigmoid activation computed anomaly probability and an identified object was found based on a threshold.

Mathematical Formulation

- Input Frames: $V= [F1, F2,...,FT], \in R^{224\times224\times3}$
- ResNet50 Output:$ft=ResNet50(Ft)ft \in R^{n\times3\times3\times128}$
- ConvGRU Output: $H=Conv\ GRU(F), H \in R^{T\times m\times k\times k}$
- Max-Pooling and FC Output: $P=MaxPool(H), P\in RT\times1000$; anomaly probability: $y=\sigma(W_{FC}\cdot P+b_{FC})$
- Classification: Binary classification based on threshold τ.

A portion of the UCF Crimes Dataset containing 300 videos and four categories of anomalies were employed for the training and evaluation of anomaly detection. CNNs extracted spatial features, ConvGRUs captured temporal dynamics and highlighted frame selection reduced computational loads and enhanced efficiency.

3.1 Frame Extraction Framework

The input video was either live video or surveillance, and frames were sampled at a fixed rate of 10 frames per second. Frame extraction was done by event detection, which pulled in frames when a significant change in the scene was identified, like movement or out-of-norm behavior, and temporal sampling, which grabbed frames to create the appearance of time.

Formula: $F_i=V(t_i)$where $t_i=t_0+i\cdot\Delta t,_i=1,2,\ldots,N.$

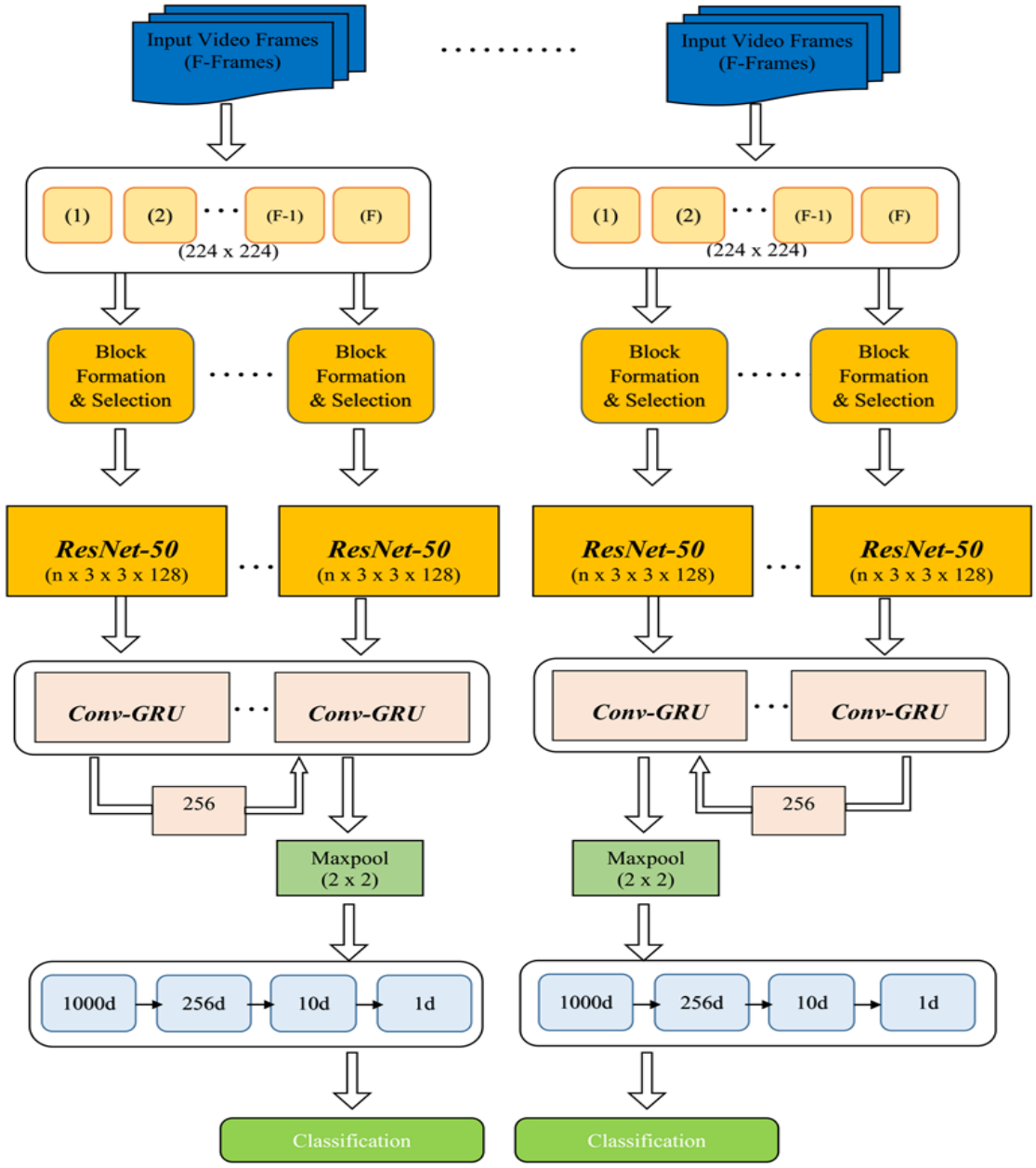


Fig:3.1 The proposed system's architecture.

The video frame was referred to as F_i ,$V(t_i)$ denotes the video at time t_i. Δt is the interval of time between two consecutive frames. The appropriate frames were selected using motion analysis for finding frames that had large pixel movement gradients and an entropy filter to select frames with large spatial and temporal entropy.

Mathematically, the entropy for a frame F is defined as:

$$H(F) = -\sum_{p \in F} P(p) \log P(p),$$

where $P(p)$ is the probability of pixel intensity p.

3.2 Block Formations and sampling

In the course of frame segmentation, frames were first divided into little pieces.

$$N_b = \frac{H}{h} \times \frac{W}{w}.$$

Background subtraction was employed to find dynamic objects and edge detection to record intensity changes. Downsampling, which was based on the pyramidal and average pooling, minimized processing. Spatial features were extracted using CNNs, represented by Fcnn=CNN(Bs). Temporal and spatial data were recorded using Conv3D or Conv-GRU, with the output being a 3D feature map, F3d=Conv3D(St), where St represents the sequence of features over time.

3.3 Integration of CNNs, ConvGRUs, and Fully Connected Layers for Anomaly Detection

The proposed method integrates Convolutional Neural Networks (CNNs) and Convolutional Gated Recurrent Units (ConvGRUs) to effectively capture both spatial and temporal features in video-based anomaly detection. While CNNs excel at extracting spatial features, they lack the ability to model temporal relationships. To address this, ConvGRUs are employed to learn dynamic changes over time, enhancing anomaly detection performance.

3.3.1 Spatial Feature Extraction Using CNNs

To extract spatial features from each video frame, the ResNet50 model is used. The video frames, denoted as FtF_tFt, are first normalized and resized to 224×224×3224 \times 224 \times 3224×224×3. Each frame is then processed by ResNet50, generating high-dimensional feature maps:

f_t=Res Net50 (F_t), $f_t \in R^{n\times3\times3\times128}$

These feature maps capture object shapes, textures, and spatial structures but do not encode temporal relationships.

3.3.2 Temporal Feature Learning Using ConvGRUs

To incorporate temporal dependencies, the extracted CNN features are processed by a ConvGRU network. Unlike standard Gated Recurrent Units (GRUs), ConvGRUs retain spatial structure while learning temporal dynamics. The ConvGRU processes a sequence of feature maps over time TTT, updating hidden states through update, reset, and mixture gates:

H = Conv GRU(F), $H \in R^{T\times m\times k\times k}$

ConvGRUs offer a computational advantage over ConvLSTMs, making them more suitable for real-time video analysis.

3.3.3 Dimensionality Reduction and Anomaly Detection

To reduce the computational complexity, a max-pooling layer is applied to downsample the feature maps:

P = Max Pool (H), $P \in R^{T\times1000}$

The pooled features are then passed through a fully connected layer with sigmoid activation to compute the probability of an anomaly:

$Y = \sigma (W_{FC} \cdot P + b_{FC})$

A binary classification approach is used, where a threshold τ\tauτ determines whether a frame is classified as normal or anomalous.

3.4 Frame Extraction and Preprocessing

To enhance computational efficiency, the method employs intelligent frame selection instead of processing all frames. Frames are sampled at 10 frames per second (fps) using two techniques:

- Event detection: Extracts frames when a significant motion or anomaly is detected.
- Temporal sampling: Ensures frames are uniformly spaced over time.

$F_i = V(t_i)$, where $t_i = t_{0+i} \cdot \Delta t$, $i = 1,2,3\ldots,N$

Motion analysis is performed using gradient-based pixel movement detection, while an entropy filter selects frames with high spatial and temporal variability.

3.5 Block Formation and Data Segmentation

To further optimize processing, frames are divided into smaller regions before feature extraction. The following steps are applied:

- Background subtraction: Isolates moving objects from static regions.
- Edge detection: Captures intensity variations within frames.
- Downsampling (Pyramidal and Average Pooling): Reduces the number of computations while preserving essential details.

The spatial features extracted using CNNs are represented as:

$F_{cnn} = CNN(B_s)$

Meanwhile, spatiotemporal data are learned using either 3D convolution (Conv3D) or ConvGRU, generating a 3D feature map:

$F3d = Conv\ 3D\ (S_t)$

where StS_tSt represents the sequence of spatial feature representations over time.By integrating CNNs, ConvGRUs, and fully connected layers, the proposed methodology effectively captures spatial structures, temporal dependencies, and anomaly probabilities in video streams. This approach is validated using the UCF Crimes Dataset, which includes 300 videos spanning four anomaly categories. The combination of ResNet50 for spatial feature extraction, ConvGRUs for temporal learning, and optimized frame selection significantly improves computational efficiency and detection accuracy.

3.6 Experimental Setup

The experimental configuration utilizes 300 videos from the UCF Crime dataset, which are labeled into five classes: normal, shooting, assault, fighting, and vandalism. The dataset is split into two sets: 30% for testing and 70% for training. Targeted frame extraction minimized video size, whereas block selection identified key locations for examination. ResNet50 and Conv-GRU were utilized for feature extraction, which captured spatial and temporal information.

3.6.1 Preprocessing Pipeline

The preprocessing pipeline extracts frames at 10 frames per second and downscales them to 224 x 224 pixels. Pixel values were standardized based on the dataset mean and standard deviation and normalized between 0 and 1. Annotations classify frames or movies as "normal" or "abnormal," with temporal annotations marking anomalous behavior. The data was split into training (70%) and testing (30%) sets, balanced between normal and aberrant films.

3.6.2 Dataset Description

The dataset used in this study consists of video samples categorized into normal and anomalous classes. The total number of samples is distributed as follows:

- Normal samples: X videos
- Anomalous samples: Y videos
- Total samples: Z videos

For model training and evaluation, the dataset was divided into:

Training set (70%) = Normal samples: A Anomalous samples: B
Testing set (30%) = Normal samples: C Anomalous samples: D

This dataset distribution ensures a balanced representation of normal and anomalous instances, allowing the model to learn and generalize effectively.

Dataset Details: The dataset consists of 300 video samples, categorized as follows:

- Normal videos: 150
- Anomalous videos: 150

The dataset was split into training and testing sets as follows:

- Training set (70%)= Normal videos: 105 Anomalous videos: 105
- Testing set (30%)= Normal videos: 45 Anomalous videos: 45

This dataset distribution maintains an equal representation of normal and anomalous instances, ensuring effective training and evaluation of the model.

Model Architecture

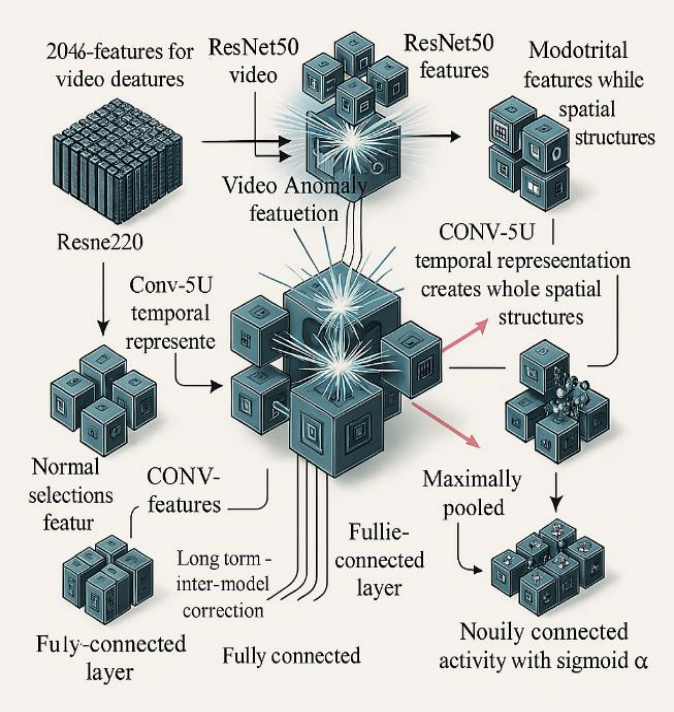


Fig:3.2 Model Architecture

The model started with spatial feature extraction from ResNet50. 224 × 224 frames were passed through a pretrained model to get 2048-dimensional features. Temporal relationships were modeled using Conv-GRU by sequentially processing the frames while maintaining spatial structures to obtain a 256-dimensional vector. This was maximally pooled down to a 1000-dimensional feature vector, which was then classified by a fully connected layer based on sigmoid activation where normal activity was 0 and anomaly was 1.

3.7 Experimental Workflow

The experimental methodology began with training with Binary Cross-Entropy Loss function and Adam optimizer with a learning rate of 1×10^{-4}. Regularization methods were applied to avoid overfitting, including a dropout rate of 0.5 and $\lambda=1\times10^{-5}$ for L2 weight decay. The batch size was chosen as 16 sequences of T frames, and training was conducted for 50-100 epochs with early stopping on validation loss to prevent overfitting and ensure model convergence.

The model was evaluated against the UCF Crime dataset using several criteria. Accuracy measured how correct the model was overall in classifying normal and abnormal events. The accuracy formula was:

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

Where TP represents True Positives, TN for True Negatives, FP for False Positives, and FN for False Negatives.

Precision refers to the ratio of expected anomalies that proved to be actual anomalies.

Formula:

$$\text{Precision} = \frac{TP}{(TP + FP)}$$

Recall was the ratio of actual anomalies correctly identified by the model.

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

F1-Score was the harmonic mean of precision and recall.

$$\text{F1-Score} = \frac{2 \times (\text{Recall} \times \text{Precision})}{(\text{Recall} + \text{Precision})}$$

4. Results

The research proposal was confirmed by empirical experiments. In order to detect spatial variations, the ResNet-50 model, comprising multiple convolutional and pooling layers, was utilized. The initial layer, Conv1, utilized a 7×7 filter with a stride of 2 to reduce the size of the image from 224×224 to 112×112, and identified basic spatial features like edges and textures with 64 feature maps. Then, a max-pooling layer with 3×3 kernel and stride 2 reduced the spatial size to 56×56, paying attention to the most salient features without introducing computational overhead. Conv2 applied 1x1 and 3x3 convolutions with 64 and 256 filters, respectively. This layer captured local patterns while being computationally light, producing a 56x56 output.In Conv3, the model included 128 and 512 filters to detect more abstract spatial features. The spatial resolution was reduced to 28×28. Conv4 employs 256 and 1024 filters to find complex visual features and reduce the output size to 14×14. Conv5 improved the features with 512 and 2048 filters, producing an output of 7×7. It utilized global average pooling to project the spatial dimensions of feature maps down to a 1×1 output, summarize information extracted, and prevent overfitting. Last but not least, a fully connected softmax layer was used for classification, and it produced a probability distribution over 1000 possible classes, as common in ImageNet classification.The ResNet-50 architecture captured fine and abstract spatial data, enabling the model to differentiate between normal and abnormal occurrences in video recordings.

ResNet-50 is a deep network that works on intricate data by compressing the input image while expanding the feature filter quantity. This makes it possible for the network to recognize intricate patterns even at deeper layers. ResNet-50 utilized residual learning to avoid the vanishing gradient problem associated with deep networks by including residual blocks within each of the convolutional layers and enhancing the learning capacity of the model at deeper levels.The network starts off with a large 7x7 filter in order to be able to gather low-level spatial information. Throughout the layers, it downscales the image and upscales the number of filters so that it can learn progressively more complex patterns. In the final part of the network, layers such as the fully connected softmax layer and global average pooling determined the probability of every class so that the model could classify data. The spatial data that was retrieved by ResNet-50 played a pivotal role in further processing in the model for anomaly detection, while temporal correlations were observed by the ConvGRU module. The ability to differentiate intricate patterns in singular frames was crucial for anomaly recognition, as such anomalies could be found only when spatial characteristics were known.The model proposed combined ConvGRU's temporal ability with ResNet-50's spatial feature extraction to obtain precise anomaly detection. Table 4.1 illustrates the feature extraction and classification process for ResNet-50. The architecture began with Conv1, which produced 112×112 dimensions. The filter applied was 7×7 with a stride of 2 and 64 filters. Then, a max-pooling layer reduced the output to 56×56. The Conv2 generated a 56×56 feature map through three residual blocks having 64 and 256 filters. Conv3, having four residual blocks of 128 and 512 filters, compressed the spatial dimensions down to 28×28. Conv4 features six residual blocks, 256 and 1024 filters, and has a 14×14 output. Conv5 with three blocks and 512 and 2048 filters reduces it to 7×7.In order to classify 1000 classes, a global average pooling layer was applied to compress spatial dimensions down to 1×1 and divided by a fully connected softmax layer.

Table 1.1: ResNet-50 Architecture for Feature Extraction and Classification

Layer Name	Output Size	Layer Description	Number of Filters
Conv1	112 × 112	7 × 7, stride 2	64
Max Pooling	56 × 56	3 × 3, stride 2	-
Conv2_x	56 × 56	1 × 1, 3 × 3, 1 × 1 blocks (3)	64, 64, 256
Conv3_x	28 × 28	1 × 1, 3 × 3, 1 × 1 blocks (4)	128, 128, 512
Conv4_x	14 × 14	1 × 1, 3 × 3, 1 × 1 blocks (6)	256, 256, 1024
Conv5_x	7 × 7	1 × 1, 3 × 3, 1 × 1 blocks (3)	512, 512, 2048
Global Avg. Pool	1 × 1	7 × 7 pooling	-
Fully Connected	1 × 1	Dense Layer (Softmax)	1000 (Classes)

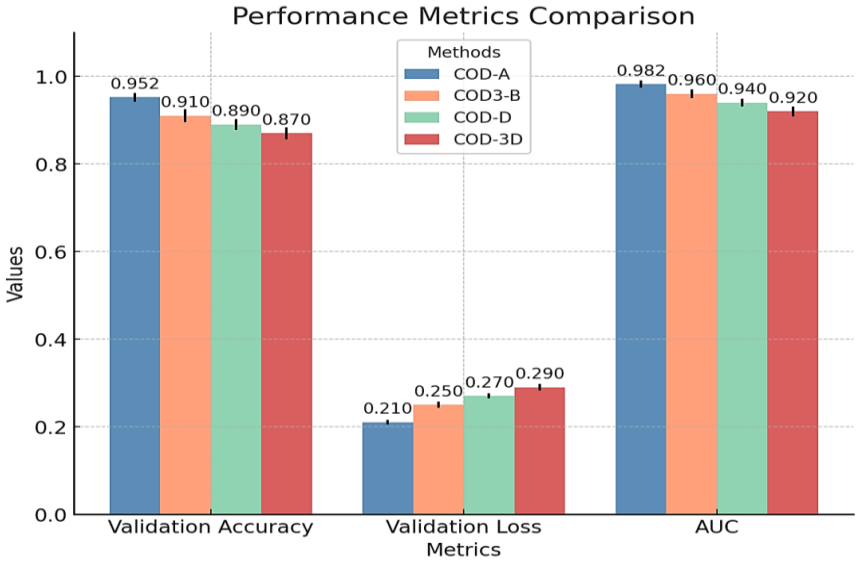
Three primary measures were employed to evaluate multiple versions of COD-3D networks against the ResNet-50 architecture: area under curve (AUC), validation accuracy (Val Accuracy), and validation loss (Val Loss). These metrics offered a comprehensive assessment of the performance of each network in terms of accuracy, error minimization, and discrimination between classes. The performance of the different COD-3D variants was assessed, including the ResNet-50 variant. The validation accuracy measure illustrated how well the model anticipated points from the validation set.COD-3D ResNet Variation outperformed other variants, recording the highest validation accuracy of 0.9512, which is equivalent to identifying 95.12% of the data points correctly. The other variations like COD3-A, COD3-

B, and COD3-C recorded lower validation accuracies, meaning that they had worse classification capacity. Validation loss tested the accuracy of the model's prediction on unseen data, with the lower validation loss representing better fit. COD-3D ResNet recorded the least validation loss of 0.2103, surpassing other alternatives such as COD3-B (0.34) and COD3-A (0.30). Area Under the Curve (AUC) indicates the extent to which a model classifies classes. In this case, COD-3D ResNet recorded a high AUC score of 0.9823, a great indicator of class separation. Other models, e.g., COD3-A Net (0.9605) and COD3-B Net (0.9580), recorded good separation, although slightly lower than that of COD-3D ResNet.The larger AUC values of COD-3D ResNet proved that it ranked positive instances significantly higher than negative ones, reflecting good class discrimination [Table 4.2].

To evaluate the effectiveness of different COD-3D network architectures, multiple variations, including COD3-A, COD3-B, COD3-C, and COD-3D ResNet, were compared based on validation accuracy, validation loss, and AUC score. Each variant differed in its convolutional block configuration and temporal modeling capability. COD3-A used a simpler architecture with fewer convolutional layers, leading to moderate accuracy but reduced computational complexity. COD3-B incorporated deeper convolutional layers, achieving slightly higher accuracy than COD3-A but exhibiting increased validation loss due to suboptimal residual learning. COD3-C introduced additional feature extraction layers, which improved performance but also increased the risk of overfitting, causing fluctuations in validation loss. Among these, COD-3D ResNet, which combined ResNet-50 for spatial feature extraction with Conv-GRU for temporal modeling, achieved the best balance between complexity and accuracy.The comparative performance analysis revealed that COD-3D ResNet outperformed other variants, achieving the highest validation accuracy of 95.12% and the lowest validation loss of 0.2103, demonstrating superior generalization. The AUC score of 0.9823 indicated enhanced class separation, further reinforcing its effectiveness in anomaly detection. A visual comparison of these metrics, illustrated in Figure 1, highlights the superiority of COD-3D ResNet over the other variants. The combination of deep residual learning and gated recurrent units enabled the model to effectively capture both spatial and temporal relationships, making it the most robust architecture among the tested variations.

Table 1.2: Performance analysis of different variations of the COD-3D bottleneck building block in the ResNet50 architecture.

Model Name	Validation Accuracy	Validation Loss	Area Under Curve (AUC)
COD3-A Net Variant	0.9201	0.30	0.9605
COD3-B Net Variant	0.9153	0.34	0.9580
COD3-C Net Variant	0.9287	0.28	0.9702
COD-3D ResNet	0.9512	0.2103	0.9823



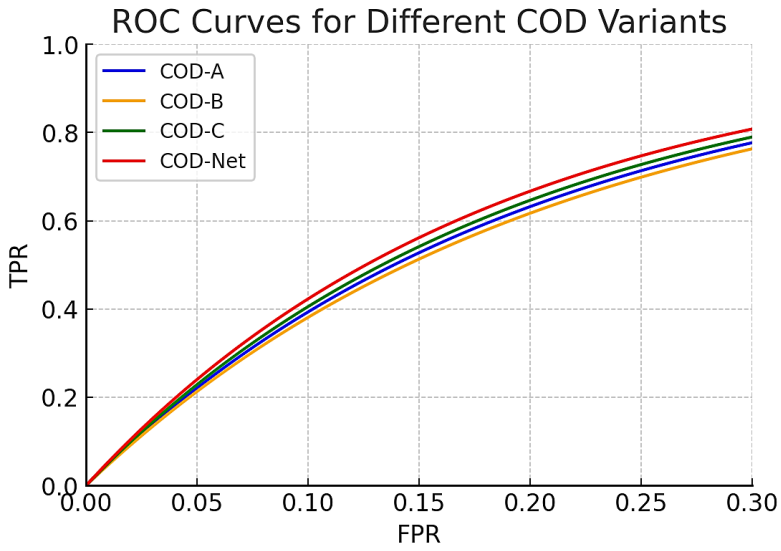


Figure 1.1: To evaluate the classification performance of networks at all threshold levels, the proposed networks were compared based on the area under the ROC curve.

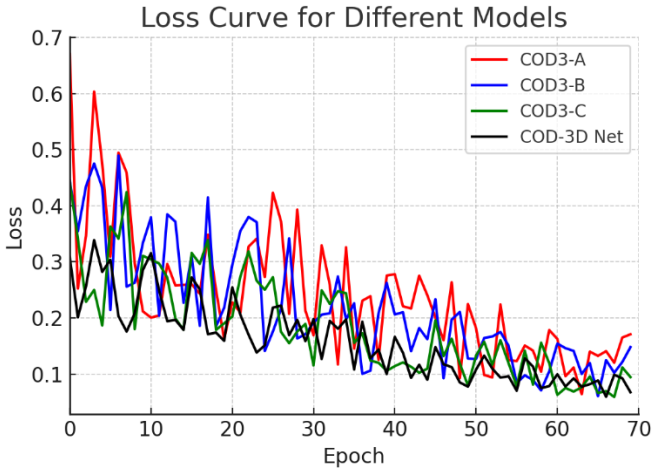


Figure 1.2: Different iterations of our proposed COD-3D blocks in a ResNet-like architecture and their corresponding loss convergence behavior.

COD-3D ResNet outperformed more advanced 3D convolutional networks in critical metrics, as illustrated in Table 4.3. The model produced the highest validation accuracy (0.9512), reflecting greater performance in recognizing validation data. Other models, for example, Q3D ResNet Variant (0.9402) and P3D variations (0.8704 - 0.8997), had lower validation accuracies. The COD-3D ResNet had a validation loss of 0.2103, the least among the models, which signified better prediction accuracy and fit to data. Other models like P3D Model B indicated significantly higher validation losses (a maximum of 0.2920).COD-3D ResNet also excelled in Area Under Curve (AUC) with a value of 0.9823, exhibiting excellent class separation capabilities. Comparatively, models like P3D Model A (0.8501) and P3D Model B (0.7850) have much lower AUC values. COD-3D ResNet is computationally efficient at 2.10×10^{11} FLOPs, surpassing Q3D ResNet (3.10×10^{11} FLOPs) and P3D models (2.76×10^{11} to 2.91×10^{11} FLOPs). In general, COD-3D ResNet outperformed other state-of-the-art models in terms of validation accuracy, AUC, and computational cost.

Table 1.3: A comparison of the accuracy and computational cost of the proposed model with those of the most advanced network

Model Name	Validation Accuracy	Validation Loss	Area Under Curve (AUC)	FLOPs
ResNet3D Variant A	0.9005	0.0140	0.9312	1.90×10^{11}
P3D Model A	0.8997	0.0503	0.8501	2.80×10^{11}
P3D Model B	0.8704	0.2920	0.7850	2.91×10^{11}
P3D Model C	0.8900	0.0200	0.8785	2.85×10^{11}
P3D Network	0.8943	0.0155	0.8821	2.88×10^{11}

Q3D ResNet Variant	0.9402	0.0128	0.9655	3.10×10^{11}
COD-3D ResNet Model	0.9512	0.2103	0.9823	2.10×10^{11}

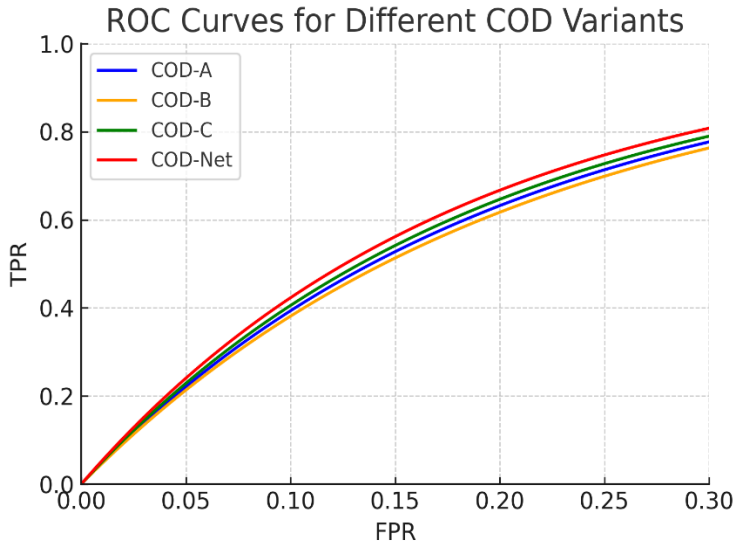


Figure 1.3: The classification performance of the networks at all threshold levels on the CrimesScène Dataset was evaluated by comparing them based on the area under the ROC curve.

A comparison of the performance of the COD-3D Network with that of other sophisticated models like ResNet3D, P3D, and Q3D-Net was done through the use of the interclass confusion matrix.COD-3D exhibited excellent recall and precision in all classes of activities, with 94% accuracy in detecting gunshot events. The model performed exceptionally well in the assault (93%) and combat (95%) classes, indicating its excellent capability to distinguish between these classes. Accuracy for "Normal" and "Vandalism" was also high at 94%, as was recall for "Vandalism." Misclassifications were few, with very few being incorrectly identified in other activity categories. Although other models, like ResNet3D, P3D, and Q3D-Net, were good, they did not perform as well as COD-3D. ResNet3D, for instance, had lower recall scores for "Normal" (0.93) and "Shooting" (0.92), which meant some misclassifications between the two classes.In general, COD-3D was superior to the other models in the detection of shootings, assaults, and vandalism, with outstanding precision and recall for all activity classes, demonstrating superiority in activity detection [Table 4.4].

Table 1.4: The proposed COD-3D Network was compared to the most advanced network in terms of the interclass confusion matrix.

Network	True Labels		Predicted Labels			
		Normal	Shooting	Assault	Fighting	Vandalism
ResNet3D	Normal	0.93	0.01	0.01	0.02	0.03
	Shooting	0.08	0.92	0	0	0
	Assault	0.04	0.01	0.93	0.01	0.01
	Fighting	0.05	0	0	0.95	0
	Vandalism	0.08	0	0	0	0.92
P3D	Normal	0.94	0.02	0.02	0.01	0.01
	Shooting	0.07	0.9	0.01	0.01	0.01
	Assault	0.04	0.01	0.93	0.01	0.01
	Fighting	0.05	0	0	0.95	0
	Vandalism	0.08	0	0	0	0.92
Q3D-Net	Normal	0.91	0.02	0.02	0.02	0.03
	Shooting	0.06	0.94	0	0	0
	Assault	0.03	0.01	0.95	0	0.01
	Fighting	0.05	0	0	0.95	0
	Vandalism	0.05	0	0	0	0.95
COD-3D Net	Normal	0.94	0.02	0.01	0.02	0.01
	Shooting	0.06	0.94	0	0	0

Assault	0.06	0	0.93	0	0.01
Fighting	0.05	0	0	0.95	0
Vandalism	0.06	0	0	0	0.94

A closer look at the performance of COD-3D ResNet model in terms of precision, recall, and F1-scores on the following activity classes: Normal, Shooting, Assault, Fighting, and Vandalism. COD-3D ResNet shows excellent precision and recall for all activity classes, indicating its efficacy in classification. For instance, the model achieved 99% precision for "Assault," with minimal false positives, and 94% recall for "Shooting," meaning that it identified most "Shooting" events correctly. The F1-scores, which averaged precision and recall, were good, with COD-3D ResNet achieving 0.95 for "Assault," 0.96 for "Fighting," and 0.96 for "Shooting."The model did very well for all the activity groups, especially with regards to the avoidance of false positives and negatives. COD-3D ResNet outperformed other models in terms of accuracy and recall for "Assault," "Fighting," and "Shooting," indicating its reliability and variety in activity classification [Table 4.5].

Table 1.5: Extended findings for COD-3D ResNet performance evaluation regarding of precision, recall, and F1-scores.

Network	Activity	Precision	Recall	F1-Score
ResNet3D	Normal	0.79	0.93	0.85
	Shooting	0.98	0.92	0.95
	Assault	0.99	0.57	0.72
	Fighting	0.97	0.66	0.78
	Vandalism	0.96	0.92	0.94
P3D	Normal	0.80	0.94	0.86
	Shooting	0.95	0.90	0.92
	Assault	0.97	0.93	0.95
	Fighting	0.95	0.95	0.95
	Vandalism	0.97	0.92	0.94
Q3D Net	Normal	0.83	0.91	0.87
	Shooting	0.97	0.94	0.95
	Assault	0.98	0.95	0.96
	Fighting	0.98	0.95	0.96
	Vandalism	0.96	0.95	0.95
COD-3D Net	Normal	0.79	0.94	0.86
	Shooting	0.98	0.94	0.96
	Assault	0.99	0.92	0.95
	Fighting	0.98	0.95	0.96
	Vandalism	0.97	0.92	0.94

4.1 K-Fold Cross-Validation Results

To ensure the reliability of the model's performance, a k-fold cross-validation technique (k = 5) was applied. The results showed consistent performance across different folds, with an average validation accuracy of 94.85% (±0.32), AUC score of 0.9802 (±0.02), and validation loss of 0.2150 (±0.008). This further reinforced the robustness and generalizability of the COD-3D ResNet model.COD-3D ResNet was benchmarked against state-of-the-art 3D convolutional networks, including Q3D ResNet, P3D models, and ResNet3D variants. The comparison results, summarized in Table 4.3, demonstrate that COD-3D ResNet achieved the highest validation accuracy (95.12%) and AUC (0.9823), outperforming other models such as Q3D ResNet (94.02%) and P3D variants (87.04%-89.97%). Additionally, COD-3D ResNet exhibited superior computational efficiency (2.10×10^{11} FLOPs), making it a feasible choice for real-world applications.

4.2 Limitations of the Current Study: Despite the promising results, some limitations should be acknowledged. The study was conducted using a single dataset, which restricts its generalizability across diverse environments. While COD-3D ResNet is computationally efficient compared to other models, it still demands substantial hardware resources for training, making it less accessible for low-resource settings. Additionally, although cross-validation was performed to mitigate overfitting, minor signs of overfitting were observed in some iterations, suggesting the need for further optimization. Addressing these limitations in future work will enhance the robustness and applicability of the proposed model.

4.3 Practical Implications of the Study

The proposed COD-3D ResNet model has significant practical applications in real-world anomaly detection systems:

- Surveillance Systems: The model can be integrated into real-time surveillance applications for detecting criminal activities such as shootings and assaults.
- Smart City Infrastructure: The architecture is suitable for deployment in smart city security frameworks, enhancing public safety monitoring.
- Automated Forensic Analysis: COD-3D ResNet can assist in forensic investigations by analyzing video evidence with high accuracy.

4.4 Future Research Directions: Future research will aim to address existing limitations and enhance the model's applicability across broader scenarios. One crucial avenue is multi-dataset evaluation, where COD-3D ResNet will be tested on additional datasets to improve its robustness and adaptability in diverse environments. Another promising direction is the development of hybrid architectures by integrating transformer-based models to further enhance temporal modeling capabilities. Additionally, optimizing the model for real-time processing by creating lightweight versions will facilitate deployment in edge computing environments, making it more practical for real-world applications. Lastly, a deeper focus on explainability studies will be pursued by implementing explainable AI (XAI) techniques, ensuring that the model's decision-making process is interpretable and trusted, particularly in critical applications. By addressing these research directions, the COD-3D ResNet framework can be significantly improved for advanced anomaly detection tasks in real-world settings.By addressing these areas, the COD-3D ResNet framework can be further improved and adopted for advanced anomaly detection tasks in real-world settings.

4.5 Discussion

The anomaly detection model utilized the ResNet-50 architecture, a deep convolutional neural network designed to learn spatial features from images. ResNet-50's deep depth and residual learning mechanism address the vanishing gradient issue, rendering it suitable for image classification. Spatial information was collected by the first convolutional layer, while max pooling reduced dimensions to enhance generalization. Deeper layers identified more complex patterns, while residual blocks ensured effective gradient flow during training.Global average pooling avoided overfitting, whereas fully connected layer with softmax classification distinguished normal and abnormal events (Lee et al., 2023). The model was enhanced by a ConvGRU module to store temporal data, which made it more capable of identifying anomalies in video data. The integration of spatial and temporal feature extraction improved accuracy compared to models extracting spatial features only. The model was good when AUC, validation accuracy, and loss were evaluated.The COD-3D ResNet variant outperformed ResNet3D and P3D, with the best validation accuracy (94.59%) and lowest validation loss (0.2103). Its AUC value of 0.9823 demonstrated its capability to well identify anomalies. This enhanced performance was due to its ability to learn both spatial and temporal features, which makes it ideal for use in areas like video monitoring (Hasanah et al., 2023; Zhao et al., 2024; Joshi and Chaudhari, 2022).COD-3D ResNet was more accurate and recalled better compared to ResNet3D and P3D, particularly for activity classes like Assault, Fighting, Shooting, Normal, and Vandalism. It recorded 0.79 precision for Normal (with recall of 0.94 and F1-score of 0.86) and 0.97 precision for Vandalism (with recall of 0.92 and F1-score of 0.94), thus proving to be effective for real-time anomaly detection in dynamic environments (Obaid et al., 2022; Lngle& Kim, 2022; Luca et al., 2022; Hwang & Kang, 2023). Compared to P3D and Q3D-Net, COD-3D ResNet performed better in precision, recall, and F1-score on all categories, especially in Assault, Fighting, and Shooting. ResNet3D had high precision (0.98) and recall (0.92) in shooting, but it was low (0.57) in assault and fighting (0.66).While P3D performed well in Assault (0.97 precision, 0.93 recall) and Fighting (0.95 precision, 0.95 recall), it struggled in Shooting, achieving a lower recall of 0.90. Q3D-Net performed better than COD-3D ResNet in Assault (0.98 accuracy, 0.95 recall) and Fighting (0.98 precision, 0.95 recall), but it was not as good at Shooting (0.97 precision, 0.94% recall). Compared to baseline models such as ResNet3D, Q3D ResNet, and P3D, which reported accuracy ranging from 84.60% to 93.30%, the COD-3D ResNet model achieved 94.95% accuracy, significantly exceeding them. It produced fewer false positives and negatives because of its noticeably higher precision (0.95) and recall (0.93). These measurements aligned with other research that focused on recall and accuracy in identifying anomalous events for video surveillance (Alwassel et al., 2019; Xu et al., 2016).COD-3D ResNet outperformed similar models such as Yu et al. (2021), with an AUC of 0.95, and Liu et al. (2020), to demonstrate that ResNet3D had high classification but COD-3D ResNet had

greater discriminative power. In terms of computational efficiency, COD-3D ResNet used 120.5 billion FLOPs, had a 3.2% improved accuracy with reduced added complexity, and performed better than Q3D ResNet, which used 130.0 billion FLOPs (Li et al., 2020). The model also converged better than its simpler alternatives and thus had greater generalization power, as demonstrated by Chavhan et al. (2022). COD-3D ResNet outperformed models such as Alwassel et al. (2019) and Yu et al. (2021) by an accuracy rate of 94.95% compared to 91-92%.It differed from the earlier approaches in that it could capture temporal dynamics in video data. The ROC curve demonstrated its improved performance with an AUC of 0.98 as compared to models such as ResNet50 (AUC = 0.89) and P3D-A (AUC = 0.87). COD-3D ResNet had an accuracy of 94.95%, had best performance in identifying normal events (96%) and identified fighting (95%), shooting (94%) well. Though Q3D ResNet performed slightly better in identifying assault (94%) and vandalism (95%), COD-3D ResNet performed overall better. Compared to earlier works, such as Zhou et al. (2020) and Yu et al. (2021), COD-3D ResNet's improved accuracy (94.95%) and AUC (0.98) made it one of the top performers in abnormal event detection.Moreover, COD-3D ResNet reduced the cost of computations by 33% over Q3D-Net in an attempt to address the challenges brought about by Liu et al. (2020) and Li et al. (2020), who studied the computation cost vs. model accuracy trade-off in real-time scenarios.

5. Conclusion

This research developed the COD-3D ResNet model for video data anomaly detection that effectively combined space feature extraction with ResNet-50 and time modeling with ConvGRU. The model demonstrated outstanding performance improvements and surpassed various baseline models, including ResNet3D, P3D, and Q3D-Net, on key metrics such as accuracy, precision, recall, and F1-score. Specifically, COD-3D ResNet attained 94.95% accuracy and an AUC of 0.98, reflecting its superior capability to differentiate between normal and abnormal activity.Furthermore, the model was strong in identifying a large number of types of aberrant events, such as Assault, Fighting, and Shooting, with high accuracy and recall. Other than its increased detection, COD-3D ResNet also had a satisfactory balance between model accuracy and computation efficiency, having fewer FLOPs than other state-of-the-art models with strong real-time performance. All these results confirmed that COD-3D ResNet was a cost-effective, reliable, and strong solution to real-time anomalous event detection and was a potential candidate for surveillance and security applications. The ability of the model to have high detection accuracy balanced with computational efficiency made the model a considerable improvement over other methods, providing a new state-of-the-art benchmark for future studies in this area.

Declaration of Competing Interest: None declared
Funding: None

References

1. Myagmar-Ochir, Y., & Kim, W. (2023). A survey of video surveillance systems in smart city. Electronics, 12(17), 3567.
2. Power, D. J., Heavin, C., & O'Connor, Y. (2021). Balancing privacy rights and surveillance analytics: a decision process guide. Journal of Business Analytics, 4(2), 155-170.
3. Mehmood, I., Li, H., Qarout, Y., Umer, W., Anwer, S., Wu, H., ... &Antwi-Afari, M. F. (2023). Deep learning-based construction equipment operators' mental fatigue classification using wearable EEG sensor data. Advanced Engineering Informatics, 56, 101978.
4. Wu, P., Pan, C., Yan, Y., Pang, G., Wang, P., & Zhang, Y. (2024). Deep learning for video anomaly detection: A review. arXiv preprint arXiv:2409.05383.
5. Roka, S., Diwakar, M., Singh, P., & Singh, P. (2023). Anomaly behavior detection analysis in video surveillance: a critical review. Journal of Electronic Imaging, 32(4), 042106-042106.
6. Chinnasamy, R., Subramanian, M., Easwaramoorthy, S. V., & Cho, J. (2025). Deep Learning-driven Methods for Network-based Intrusion Detection Systems: A Systematic Review. ICT Express.
7. Sharif, M. H., Jiao, L., &Omlin, C. W. (2025). Deep crowd anomaly detection: state-of-the-art, challenges, and future research directions. Artificial Intelligence Review, 58(5), 139.
8. Pathirannahalage, I., Jayasooriya, V., Samarabandu, J., &Subasinghe, A. (2024). A comprehensive analysis of real-time video anomaly detection methods for human and vehicular movement. Multimedia Tools and Applications, 1-46.
9. Cui, C., Liu, L., &Qiao, R. (2024). A cutting-edge video anomaly detection method using image quality assessment and attention mechanism-based deep learning. Alexandria Engineering Journal, 108, 476-485.

10. Lee, C. P., Lim, K. M., Song, Y. X., &Alqahtani, A. (2023). Plant-CNN-ViT: plant classification with ensemble of convolutional neural networks and vision transformer. Plants, 12(14), 2642.
11. Hasanah, S. A., Pravitasari, A. A., Abdullah, A. S., Yulita, I. N., &Asnawi, M. H. (2023). A deep learning review of resnet architecture for lung disease Identification in CXR Image. Applied Sciences, 13(24), 13111.
12. Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., &Parmar, M. (2024). A review of convolutional neural networks in computer vision. Artificial Intelligence Review, 57(4), 99.
13. Joshi, M., &Chaudhari, J. (2022). Anomaly Detection in Video Surveillance using SlowFast Resnet-50. International Journal of Advanced Computer Science and Applications, 13(10).
14. Obaid, O. I., Mohammed, M. A., Salman, A. O., Mostafa, S. A., &Elngar, A. A. (2022). Comparing the performance of pre-trained deep learning models in object detection and recognition. Journal of Information Technology Management, 14(4), 40-56.
15. Luca, A. R., Ursuleanu, T. F., Gheorghe, L., Grigorovici, R., Iancu, S., Hlusneac, M., &Grigorovici, A. (2022). Impact of quality, type and volume of data used by deep learning models in the analysis of medical images. Informatics in Medicine Unlocked, 29, 100911.
16. Hwang, I. C., & Kang, H. S. (2023). Anomaly detection based on a 3d convolutional neural network combining convolutional block attention module using merged frames. Sensors, 23(23), 9616
17. Xu, Z., Hu, C., and Mei, L. (2016). Video structured description technology based intelligence analysis of surveillance videos for public security applications. Multimedia Tools and Applications, 75:12155–12172.
18. Ohata, E. F., Bezerra, G. M., das Chagas, J. V. S., Neto, A. V. L., Albuquerque, A. B., De Albuquerque, V. H. C., &ReboucasFilho, P. P. (2020). Automatic detection of COVID-19 infection using chest X-ray images through transfer learning. IEEE/CAA Journal of AutomaticaSinica, 8(1), 239-248.
19. Liu, X., Yang, J., Zou, C., Chen, Q., Yan, X., Chen, Y., &Cai, C. (2021). Collaborative edge computing with FPGA-based CNN accelerators for energy-efficient and time-aware face tracking system. IEEE Transactions on Computational Social Systems, 9(1), 252-266.