Detection and Identification of Anomalous Events in Videos Using Deep Learning-Based Anomaly Detection

Alex Sander - 13521061
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jalan Ganesha 10 Bandung
E-mail: 13521061@std.stei.itb.ac.id

***Abstract*— In recent years, the growing demand for intelligent and efficient surveillance systems has driven advancements in anomaly detection technology for video analysis. This study proposes a deep learning-based approach to detect and identify anomalous events in videos, integrating anomaly detection and classification modules into a simple simulation system with a graphical user interface. The anomaly detection module employs an Inflated 3D ConvNet (I3D) architecture, while two pre-trained YOLO models (YOLOv8 and YOLOv11) are compared for anomaly classification. Experiments were conducted on the UCF-Crime dataset, using 10% for testing. Results show that the I3D model achieves a ROC-AUC of 0.73 for anomaly detection, while YOLOv11 outperforms YOLOv8 in classification across 13 anomaly classes with an accuracy of 74.75%. Future improvements include training with higher-quality datasets and larger model architectures to enhance performance.**

Keywords*— anomaly detection, anomaly type identification, deep learning, Inflated 3D ConvNet, YOLO

I. Introduction

In recent years, the demand for intelligent surveillance systems capable of automatically detecting suspicious activities has increased significantly, particularly in public surveillance domains such as train stations, airports, shopping centers, and highways. Conventional video surveillance systems that rely on direct human monitoring have proven to be limited in both response speed and accuracy. These limitations highlight the need for automated approaches that can detect and classify anomalous events effectively.

Deep learning-based anomaly detection has emerged as a promising solution, enabling systems to learn complex temporal and spatial patterns from large-scale video datasets. By leveraging such methods, surveillance systems can automatically detect unusual activities and identify the type of anomaly, thereby enhancing efficiency and reducing dependency on human operators. This research focuses on developing and integrating an anomaly detection module and an anomaly classification module into a unified system with a graphical user interface. The system aims to improve the accuracy and consistency of video-based anomaly detection and identification, providing a foundation for more advanced and scalable surveillance solutions in the future.

II. Theoretical Foundation

A. Video

A video, or moving image, is a medium for storing visual information composed of a sequence of still images called frames. These frames are displayed sequentially at a specific rate, creating the illusion of motion. The rate is measured in frames per second (fps), indicating the number of frames displayed in one second.

In a video, object motion can be detected by observing changes in the position or color of the object within a frame and comparing them with subsequent frames. If significant changes in position or color are detected, the object can be labeled as moving. This principle forms the foundation for motion detection and analysis in computer vision applications, including anomaly detection in surveillance videos.

B. UCF-Crime Dataset

The UCF-Crime dataset is a large-scale dataset developed by the Center for Research in Computer Vision (CRCV) at the University of Central Florida for research in anomaly detection in surveillance videos. It contains 1,900 long-duration videos with a total length of approximately 128 hours, capturing real-world footage from various public locations such as highways, shopping areas, and other public facilities. The dataset is divided into two main categories: anomalous event videos and normal videos. The anomalous events cover 13 classes, including Abuse, Arrest, Arson, Assault, Burglary, Explosion, Fighting, Road Accident, Robbery, Shooting, Stealing, Shoplifting, and Vandalism. Normal videos represent daily activities that do not involve criminal acts or suspicious incidents.

A key strength of this dataset lies in its untrimmed and real-world nature, which presents more complex challenges compared to datasets composed of short, pre-segmented clips. In anomaly detection research, this is highly relevant, as the model must not only recognize anomalous patterns but also distinguish critical event segments within lengthy videos that mostly contain normal activities. Additional challenges include class imbalance, where anomalous events are significantly less frequent than normal videos, as well as variations in video capture conditions such as lighting, camera angles, and crowd density.

C. Anomaly Detection and Identification

According to Foorthuis, an anomaly is an event or a group of events considered unusual or deviating from a concept of normality. Anomalies are also referred to as outliers, novelties, deviants, or discords. These events are rare and differ from common situations, encompassing various phenomena, whether static entities or time-related occurrences, in both atomic (single-case) and aggregate forms, and whether desirable or undesirable. In general, anomalies represent any deviation from typical patterns or expectations, often drawing special attention due to their rarity and departure from established norms.

Anomaly detection is typically formulated as an unsupervised one-class classification problem, aiming to learn the normal state of data during training and subsequently detect deviations without explicit anomaly labels. Many anomaly detection applications involve visual data, such as images or videos, and are often motivated by surveillance-related needs. Commonly evaluated datasets for visual anomaly detection are recorded using stationary cameras that observe a region where the background remains relatively static while foreground objects, such as pedestrians and vehicles, move.

Beyond detecting the presence of anomalies, some scenarios also require anomaly type identification, the process of classifying the specific category of a detected anomalous event. This allows the system not only to determine that an event deviates from normality, but also to identify whether it belongs to categories such as fighting, theft, traffic accidents, or arson. This approach generally requires datasets with class-level annotations for each anomaly type, such as the UCF-Crime dataset, which contains 13 anomaly event categories. Consequently, anomaly type identification plays a crucial role in improving surveillance system responses by enabling prioritized handling based on the severity and nature of the detected event.

D. Multiple Instance Learning

Multiple Instance Learning (MIL) is a machine learning approach designed to handle problems where labels are provided only at the aggregate or bag level, rather than at the individual instance level. In traditional machine learning settings, each data sample, such as an image or a video frame, is explicitly labeled and used directly for model training. However, in MIL, labels are assigned only to bags, each containing multiple instances with unknown labels.

A bag is labeled as positive if at least one instance within it belongs to the positive class, and negative if all instances belong to the negative class. This setup enables MIL to operate in scenarios where instance-level annotation is unavailable. The training process in MIL involves learning the relationship between a bag's label and the patterns present within its individual instances. The model aims to predict the bag label by identifying discriminative features among the instances, even though instance-level labels are unknown.

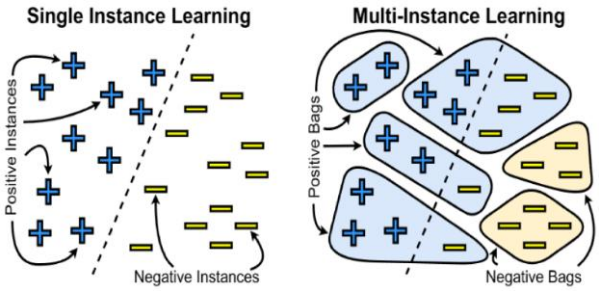

Fig 2.1. Illustration of Multiple Instance Learning Data Labeling (Fatima,S. et al, 2023)

MIL is particularly effective for tasks with limited annotations, especially in situations where obtaining instance-level labels is difficult or costly, such as large-scale video anomaly detection.

E. Convolutional 3D

Convolutional 3D (C3D) is a deep neural network architecture designed to process video data by leveraging three-dimensional convolutions. Introduced by Tran et al. [8], C3D captures both spatial and temporal information simultaneously by applying 3D filters to video blocks, where the three dimensions correspond to height, width, and time.

In C3D, a video clip is first converted into a 4D tensor representation with dimensions (number of frames × height × width × channels). A 3D convolution is then applied using a kernel/filter of size (k_t, k_h, k_w), where k_t is the temporal size, $\mathrm{k_h}$ is the height, and k_h is the width. The kernel slides across spatial and temporal positions, computing element-wise multiplications between the input pixels and kernel weights, followed by summation to produce the feature value at that position. A non-linear activation function, typically ReLU, is applied afterward to introduce non-linearity into the model.

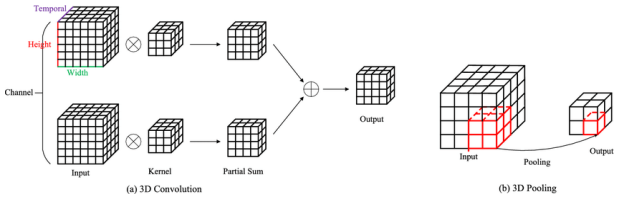

Fig 2.2. Illustration of C3D Process (Li, G. et al, 2022)

In the context of C3D, a kernel is a three-dimensional weight matrix that extracts patterns from video data, capturing both spatial and temporal features. These kernels are learned automatically during training to detect relevant features at multiple spatial and temporal scales. Following convolution, 3D pooling, commonly max pooling, is applied to reduce data dimensions while retaining key features. Pooling operates over small 3D blocks in time, height, and width, improving the model's robustness to small shifts and local variations. The general C3D process can be summarized as follows:

1. Convert the input video into a 4D tensor *(Time, Height, Width, Channel)*.
2. Perform 3D convolution with kernels (k_t, k_h, k_w) to extract spatial-temporal features.
3. Apply non-linear activation (ReLU).
4. Perform 3D pooling to reduce dimensionality and generalize features.
5. Apply subsequent layers (convolution, pooling, fully connected) until classification or detection output.

C3D typically processes short video clips to learn motion changes and visual patterns effectively. Its main advantages are architectural simplicity and inference efficiency. However, due to its relatively shallow depth, its ability to capture complex motion patterns remains limited.

F. *Inflated 3D ConvNet*

The Inflated 3D ConvNet (I3D), introduced by Carreira and Zisserman, extends the 2D kernels and pooling operations of the Inception architecture into three dimensions, a process referred to as "inflation." This approach enables the use of pre-trained weights from large-scale image datasets, such as ImageNet, and subsequently fine-tunes them on video datasets. By inflating 2D filters into 3D, I3D can capture both spatial and temporal representations in a richer and more structured manner.

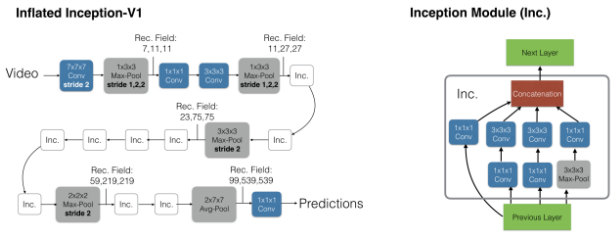

Fig 2.3. I3D Architecture (Carriera, J. & Zisserman, A., 2017)

The I3D architecture divides processing into multiple branches within each Inception module, each employing different 3D kernel sizes to extract features at varying spatial-temporal scales. The outputs of these branches are concatenated to form a richer feature representation. Batch normalization and ReLU activation are applied after each convolution to maintain training stability. Dimensionality reduction is performed progressively using 3D pooling, which compresses both spatial resolution and temporal information.

Compared to C3D, I3D generally achieves superior performance in video activity recognition tasks due to its ability to model more complex inter-frame relationships. However, this advantage comes at the cost of higher computational complexity and memory requirements. Prior studies have reported that I3D achieves a ROC-AUC of 0.8403, outperforming the C3D Two-Stream model (0.7541). Furthermore, the Two-Stream I3D variant achieves an even higher ROC-AUC of 0.8445, demonstrating that combining information from both RGB and optical flow streams yields richer spatial-temporal representations, albeit with further increases in computational cost.

In the context of video anomaly detection, I3D can be integrated with Multiple Instance Learning (MIL) to handle cases where only video-level labels are available without precise temporal annotations. In this scenario, a video is treated as a bag containing multiple instances, represented as short temporal segments. I3D is used to extract spatial-temporal features from each segment, while MIL identifies which segments contribute most to the anomaly classification. This integration enables weakly supervised learning of anomaly patterns, maintaining effectiveness even in the absence of detailed temporal segmentation labels.

G. *YOLO*

YOLO (You Only Look Once) is a computer vision method for real-time object detection based on deep learning. According to Redmon et al., YOLO formulates the object detection task as a single regression problem, directly mapping the entire input image to bounding box coordinates and class probabilities, enabling object detection in a single inference pass.

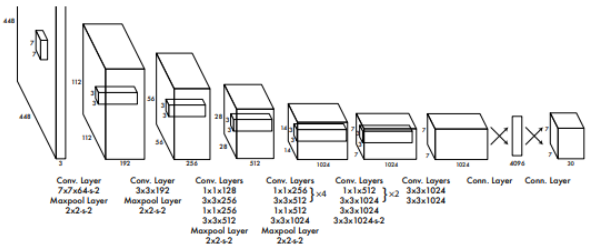

Fig 2.4. YOLO General Architecture (Redmon, J. et al, 2015)

The YOLO architecture divides the input image into a grid, with each grid cell responsible for predicting the presence of an object and its associated bounding box parameters. This design results in a fast and efficient end-to-end convolutional network. Later developments, as described by Redmon and Farhadi, such as YOLOv3, introduced multi-scale detection and residual connections to improve accuracy for small objects without sacrificing speed.

Multi-scale detection allows predictions at multiple feature resolutions, for example, large feature maps for small objects and smaller feature maps for large objects, enhancing the model's ability to capture information at various spatial detail levels. Residual connections directly link the output of earlier layers to deeper layers, facilitating gradient flow during training and reducing the risk of vanishing gradients. This mechanism enables the network to learn more complex representations without performance degradation in deeper architectures.

For anomaly type identification tasks, recent publicly available YOLO variants such as YOLOv8 and YOLOv11 can be employed due to their improved detection accuracy, scalability, and computational efficiency.

III. Problem analysis and Solution

A. Problem Identification

Anomalous event detection in videos is a critical challenge in various applications, including security monitoring, behavioral analysis, and automated surveillance systems. Anomalous events refer to unusual or suspicious motion patterns in videos that deviate from normal activities. In video processing, traditional approaches such as statistical analysis or template matching are often insufficient to address the complexity of motion dynamics. These methods typically rely solely on spatial information and fail to capture temporal dependencies across frames.

Deep learning technologies, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have shown great potential in video analysis for tasks such as object detection and event recognition. However, their application to anomalous event detection still faces significant challenges, including data imbalance and limited availability of accurately labeled datasets. Data imbalance is a major obstacle, as normal samples often far outnumber anomalous ones, causing models to favor normal pattern recognition while struggling to detect rare anomalies. Furthermore, accurate annotation of anomalous events is challenging due to high variability in motion patterns across diverse scenarios.

In this context, weakly supervised or unsupervised approaches are highly relevant, as they reduce the dependency on time-consuming manual annotations. Deep learning-based anomaly detection offers a promising solution by enabling models to automatically learn normal motion patterns and detect deviations that signify anomalous events. Although prior studies have applied deep learning to video anomaly detection, many still rely heavily on labeled data and do not fully exploit the potential of anomaly detection for visually subtle anomalies.

From the above problem identification, three main conclusions can be drawn. First, anomalous event detection in videos continues to face significant challenges in handling data imbalance and complex motion patterns. Second, traditional anomaly detection methods are less effective for dynamic and complex video data. Third, while deep learning holds great potential, weakly supervised or unsupervised approaches are essential to reduce reliance on extensive manual annotations. Therefore, a deep learning-based anomaly detection method that can detect anomalous events automatically, efficiently, and accurately is required to enhance the capabilities of automated surveillance systems across various applications.

B. Solution

Based on the problem analysis, the proposed solution is the development of a deep learning–based anomaly detection system for surveillance videos. The system aims to automatically detect unusual or suspicious movements with high accuracy, leveraging methods capable of handling the variability and complexity of real-world video data. By utilizing deep learning, the solution addresses the limitations of conventional methods that rely on manual monitoring, while improving both efficiency and detection accuracy in surveillance contexts.

The proposed system tackles key challenges such as the difficulty of detecting rare anomalies, the need for robust model generalization across various surveillance scenarios, and the limitations of models that fail to capture both spatial and temporal relationships in video data. To address these challenges, the system integrates advanced video understanding architectures, namely Inflated 3D ConvNets (I3D), which have proven effective in modelling temporal and spatial dynamics in surveillance footage.

The system architecture consists of several primary stages:

1. Preprocessing**:** Surveillance videos are resized and adjusted to meet the input requirements of the deep learning models.
2. Feature Extraction**:** When using I3D, the system benefits from deeper 3D ConvNet architectures with inflated filters and transfer learning from large-scale video datasets such as Kinetics, enhancing anomaly detection accuracy.
3. Anomaly Detection**:** The extracted features are analysed to identify deviations from learned normal motion patterns. Significant deviations trigger

anomaly detection, producing a binary output (True/False) indicating the presence of suspicious activity.

4. Anomaly Classification**:** For frames flagged as anomalous, a fine-tuned YOLO classification model is employed to identify the specific type of anomaly within the detected frame.

By combining the strengths of I3D for anomaly detection and YOLO for anomaly classification, the proposed solution is expected to outperform traditional approaches in both detection accuracy and robustness. The system's ability to jointly capture spatial and temporal patterns, while accurately classifying detected anomalies, enables a more efficient, automated, and reliable surveillance solution applicable to domains such as public security, behavioral analysis, and industrial monitoring.

IV. Evaluation

A. Anomaly Detection

The performance evaluation of the anomaly detection module was conducted by measuring the True Positive Rate (TPR) and False Positive Rate (FPR) obtained from the detection results of the implemented model. These two variables were then used to construct a Receiver Operating Characteristic (ROC) curve. Subsequently, the Area Under the Curve (AUC) was calculated from the constructed ROC curve, which serves as a key performance metric for this module.

The anomaly detection performance was evaluated using the Inflated 3D ConvNet (I3D) model, which had been trained with the dataset. The testing was performed on a test dataset consisting of 150 videos labeled as Normal and 140 videos labeled as Anomalous (non-Normal).

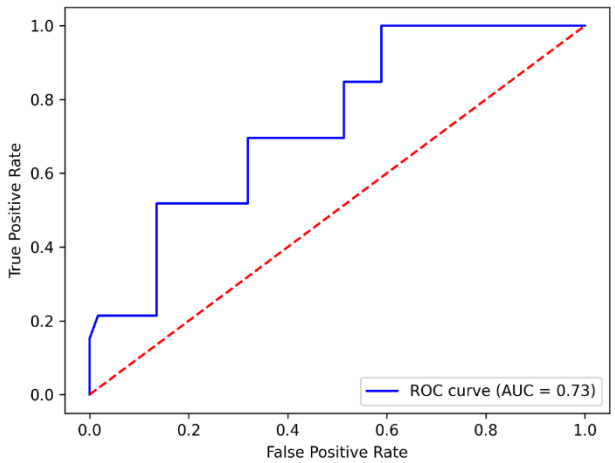

Fig 4.1. I3D ROC-AUC Graph

Fig. 4.1 presents the ROC curve generated from the evaluation of the I3D model. The resulting AUC value is 0.73, indicating that the model demonstrates a reasonably good capability in detecting the presence of anomalous events. However, this performance is lower than the result reported in previous studies, where an AUC value of 0.8403 was achieved.

B. Anomaly Identification

The performance evaluation of the anomaly type identification module was conducted by measuring the classification accuracy for each anomaly class as well as the overall accuracy. The evaluation was performed on an annotated dataset in which frames containing anomalies were labeled. The dataset consisted of 140 videos and included 13 anomaly classes with varying numbers of frames per class. The accuracy was calculated by dividing the number of correctly identified frames by the total number of frames tested, thus providing a quantitative measure of the module's performance.

Two classification models were evaluated: YOLOv8s-cls and YOLO11s-cls. The results of both models were compared to determine the model with the best performance for integration into the system.

1. YOLO11s-cls Model Testing

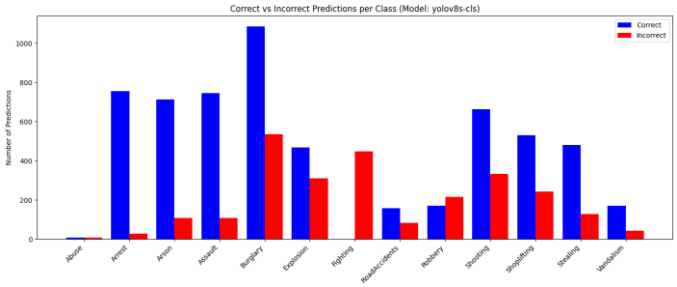

Fig 4.2. YOLO11s-cls Testing Results

The YOLO11s-cls model achieved an overall accuracy of 74.75%, correctly identifying 6,390 out of 8,548 tested frames. This performance indicates that the model is highly capable of identifying anomaly types from incoming frames. Further analysis per class accuracy is required to better understand strengths and weaknesses.

2. YOLOv8s-cls Model Testing

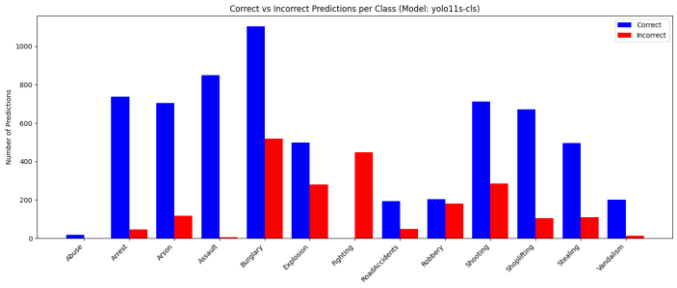

Fig 4.3. YOLOv8s-cls Testing Results

The YOLOv8s-cls model achieved an overall accuracy of 69.63%, correctly identifying 5,952 out of 8,548 tested frames. The model showed good capability in identifying classes such

as *Abuse*, *Arrest*, *Arson*, *Assault*, *Road Accidents*, *Shooting*, *Stealing*, *Shoplifting*, and *Vandalism*, and reasonable performance for *Burglary*, *Explosion*, and *Robbery*. However, its performance on *Fighting* was notably poor, with an accuracy of 0.22%.

C. Prototype

The prototype testing aimed to verify the successful integration of the anomaly detection module and the anomaly type identification module into the application interface. The complete interface of the application during the testing process is shown in Fig. IV.5.

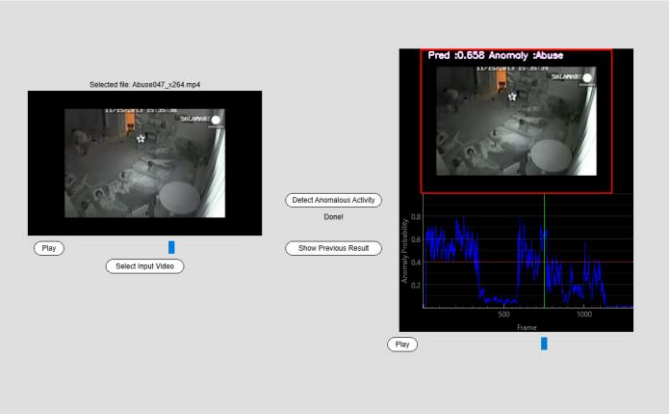

Fig. 4.4. Graphical User Interface of the Prototype

As illustrated in Fig. 4.4, the application accepts a video input for anomaly detection and identification. Once the video is loaded, the application provides a built-in media player to display the input video. The user can initiate the main process by selecting the "Detect Anomalous Activity" button. Upon completion, the resulting processed video is displayed in a separate media player, accompanied by a probability graph showing the likelihood of anomaly occurrence throughout the video.

D. Analysis

As shown in Fig. *4.2* and *4.3*, the identification performance varied significantly across classes. The trained models demonstrated excellent identification capability for classes such as *Arrest*, *Arson*, *Assault*, *Stealing*, and *Vandalism*, and good performance for *Abuse*, *Burglary*, *Explosion*, *Road Accidents*, *Shooting*, and *Shoplifting*. However, the models performed poorly for *Fighting* and *Robbery*.

The results show that both models can identify anomaly types effectively; however, both struggle with certain classes, particularly *Fighting* and *Robbery*. Overall, YOLO11s-cls outperformed YOLOv8s-cls and was therefore selected as the anomaly type identification model for the proposed anomaly detection and identification system.

Based on the evaluation of the anomaly detection module, anomaly type identification module, and the prototype application, the developed system exhibits certain limitations in accurately detecting and identifying anomalies. The system is still unable to correctly detect and classify certain types of anomalous events, indicating that the implemented modules are not yet fully optimized and require further training.

The system also experiences a decline in performance when processing videos with very long durations. This is due to the anomaly detection model's reliance on temporal context to determine the presence of anomalies within a video. While incorporating temporal information allows the model to more accurately determine the presence of anomalies in specific frames, it also increases the processing time for long input videos.

Considering the implications of these results and the identified limitations, several potential improvements can be made to the proposed system. For the anomaly detection module, training the model with a larger and more diverse dataset could enhance its performance. For the anomaly type identification module, higher-quality datasets and classification models that leverage temporal context for frame-level classification are recommended. Furthermore, improving the dataset to achieve a more balanced class distribution could enhance the performance of both modules.

V. Conclusion and Future Works

A. Conclusion

An anomaly detection and identification system was developed to identify the presence of anomalous events in videos and determine their specific types using deep learning-based methods. Based on the implementation and evaluation results, the following conclusions can be drawn:

1. Anomaly detection in videos was performed using an Inflated 3D ConvNet model integrated into the anomaly detection module. The developed module achieved good detection performance, with an ROC-AUC score of 0.73 on the test dataset. However, the model's accuracy can be further improved through additional training, and its computational time remains relatively high.
2. Anomaly type identification was conducted using a pre-trained YOLOv11 classification variant integrated into the anomaly classification module. This module achieved a classification accuracy of 0.7475 on the test dataset. Nevertheless, the model still faces challenges in identifying certain anomaly types accurately.

B. Future Works

The developed anomaly detection and classification system successfully identified anomalous events in input videos using deep learning-based anomaly detection. However, several improvements can be pursued in future research:

1. Increasing both the quantity and quality of anomaly data in the datasets used for training the anomaly detection and classification models. Enhancing dataset

quality and diversity can help improve model performance, which in this study remains suboptimal for certain anomaly cases.

2. Employing deep learning models that incorporate temporal context for the anomaly classification module, such as integrating YOLO with Long Short-Term Memory (LSTM) networks. Utilizing models capable of processing temporal information can enhance the accuracy of anomaly type classification.

REFERENCES

[1] W. Sultani, C. Chen, and M. Shah, "Real-World Anomaly Detection in Surveillance Videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 6479–6488, doi: 10.1109/CVPR.2018.00678.

[2] H. Singh, E. M. Hand, and K. Alexis, "Anomalous Motion Detection on Highway using Deep Learning," Univ. of Nevada, 2020. [Online]. Available: https://arxiv.org/pdf/2006.08143v1

[3] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation," in *AI 2006: Advances in Artificial Intelligence*, vol. 4304, L. S. et al., Eds. Berlin, Heidelberg: Springer, 2006, pp. 1015–1021, doi: 10.1007/11941439_114.

[4] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: An overview," *arXiv preprint arXiv:2008.05756*, 2020, doi: 10.48550/arXiv.2008.05756.

[5] X. Sheng, L. Li, D. Liu, and H. Li, "Spatial Decomposition and Temporal Fusion Based Inter Prediction for Learned Video Compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, pp. 6460–6473, 2024. [Online]. Available: https://arxiv.org/pdf/2401.15864

[6] A. M. Tekalp, *Digital Video Processing*. Upper Saddle River, NJ, USA: Prentice Hall, 1995.

[7] R. Foorthuis, "On the nature and types of anomalies: A review of deviations in data," *Int. J. Data Sci. Anal.*, vol. 12, pp. 297–331, 2020. [Online]. Available: https://www.researchgate.net/publication/353701611

[8] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," *arXiv preprint arXiv:1611.07004*, 2018. [Online]. Available: https://arxiv.org/abs/1611.07004

[9] W. Lotter, G. Kreiman, and D. Cox, "Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning," *arXiv preprint arXiv:1605.08104*, 2017. [Online]. Available: https://arxiv.org/abs/1605.08104

[10] Y. Tian et al., "Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, pp. 4955–4966, doi: 10.1109/ICCV48922.2021.00493.

[11] J. Park, J. Kim, and B. Han, "Learning to Adapt to Unseen Abnormal Activities Under Weak Supervision," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2020, vol. 12626, pp. 611–627, doi: 10.1007/978-3-030-69541-5_31.

[12] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 4489–4497.

[13] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 6299–6308.

[14] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple-instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, no. 1–2, pp. 31–71, 1997.

[15] X. Liu, F. Dai, J. Han, and J. Yang, "Multiple instance learning: A survey," *Comput. Intell.*, vol. 28, no. 4, pp. 436–461, 2012.

[16] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2009.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 779–788.

[18] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv preprint arXiv:1804.02767*, 2018. [Online]. Available: https://arxiv.org/abs/1804.02767

[19] J. Tandel, S. Darak, K. Desai, M. Desai, and M. Kothari, "Human Anomaly Detection System Using YOLOv8 and LSTM," *Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET)*, vol. 12, no. 11, pp. 814–822, 2024, doi: 10.22214/ijraset.2024.57248.

[20] S. Fatima, S. Ali, and H.-C. Kim, "A Comprehensive Review on Multiple Instance Learning," *Electronics*, vol. 12, no. 20, p. 4323, 2023, doi: 10.3390/electronics12204323.

[21] G. Li, M. Zhang, Q. Zhang, and Z. Lin, "Efficient binary 3D convolutional neural network and hardware accelerator," *J. Real-Time Image Process.*, vol. 19, pp. 71–84, 2022, doi: 10.1007/s11554-021-01161-4.