Naif Arab University for Security Sciences
Arab Journal of Forensic Sciences and Forensic Medicine
المجلة العربية لعلوم الأدلة الجنائية والطب الشرعي
https://journals.nauss.edu.sa/index.php/AJFSFM

Deep Learning-Based Forensic Detection of Suspicious Activities in CCTV Systems

الكشف الجنائي عن الأنشطة المشبوهة في أنظمة التلفزيون المغلق (CCTV) القائم على التعلم العميق

Qazi Emad Ul Haq[1,2*], Muhammad Imran[3], Muhammad Waqas[4], and Muhammad Hamza Faheem[1,2]

[1]Department of Cybersecurity and Digital Forensics, College of Forensics and Investigative Sciences, Naif Arab University for Security Sciences, Riyadh, Saudi Arabia

[2]Center of Artificial Intelligence for Security, Naif Arab University for Security Sciences, Riyadh, Saudi Arabia

[3]Center for Smart Analytics, Institute of Innovation, Science and Sustainability, Federation University Australia, Australia

[4]Threat Defense, Sydney, Australia

Received 16 Feb 2026; accepted 30 Mar 2026; available online 10 Sep. 2026.

Abstract

Advancements in smart surveillance and digital forensics have been significantly influenced by artificial intelligence (AI), machine learning (ML), and cloud computing, enabling real-time forensic analysis of CCTV footage to detect suspicious activities and reduce reliance on manual monitoring. While these technologies enhance public safety and policing efficiency, they also raise important concerns related to privacy, ethics, and regulation. This study proposes a deep learning-based forensic framework for real-time detection of suspicious human activity in CCTV videos, trained without relying on any external sensors. The model utilizes EfficientNet-B5 for spatial feature extraction and a Vision Transformer (ViT) that has been adapted to model both spatial and temporal dependencies between consecutive video frames. Specifically, the ViT processes frame sequences as token embeddings, enabling temporal context learning through self-attention. The EfficientNet-B5, a convolutional neural network (CNN), functions primarily as a high-level feature extractor, capturing geometric and spatial characteristics such as scale variations, aspect

Keywords: forensic sciences, artificial intelligence, closed-circuit television (CCTV), deep learning, surveillance, vision transformer (ViT)

المستخلص

لقد تأثرت التطورات في مجال المراقبة الذكية والتحقيقات الجنائية الرقمية بشكل ملحوظ بكل من الذكاء الاصطناعي (AI)، والتعلم الآلي (ML)، والحوسبة السحابية؛ مما أتاح إجراء تحليل جنائي في الوقت الفعلي لتسجيلات كاميرات المراقبة (CCTV) لاكتشاف الأنشطة المشبوهة وتقليل الاعتماد على المراقبة اليدوية. وعلى الرغم من أن هذه التقنيات تعزز السلامة العامة وكفاءة العمل الشرطي، إلا أنها تثير أيضاً مخاوف هامة تتعلق بالخصوصية والأخلاقيات والتنظيم.

تقترح هذه الدراسة إطاراً جنائياً قائماً على التعلم العميق للكشف الفوري عن الأنشطة البشرية المشبوهة في مقاطع فيديو كاميرات المراقبة، وقد تم تدريبه دون الاعتماد على أي مستشعرات خارجية. يستخدم النموذج شبكة (EfficientNet-B5) لاستخراج الميزات المكانية، إلى جانب محول بصري (Vision Transformer - ViT) تم تكييفه لنمذجة التبعيات المكانية والزمنية بين إطارات الفيديو المتتالية. وبشكل تحديد، يقوم الـ (ViT) بمعالجة تسلسلات الإطارات في هيئة ترميزات رموز (To-ken Embeddings)، مما يتيح تعلم السياق الزمني من خلال الانتباه الذاتي. وتعمل شبكة (EfficientNet-B5)، وهي شبكة عصبية تلافيفية (CNN)، بشكل أساسي كمستخرج للميزات عالية المستوى، حيث تلتقط الخصائص الهندسية والمكانية مثل اختلافات المقاييس، وننسب الأبعاد، والأنماط المرئية المرتبطة بالحركة من الإطارات الفردية.

الكلمات المفتاحية: علوم الأدلة الجنائية، الذكاء الاصطناعي، التلفزيون المغلق الدائرة (CCTV)، التعلم العميق، المراقبة، محول الرؤية

* Corresponding Author: Qazi Emad Ul Haq
Email: qabdulrab@nauss.edu.sa

doi: 10.26735/GGIY2111

ratios, and motion-relevant visual patterns from individual frames. The Binary UCF-Crime dataset is used for training and validation, with class rebalancing achieved through data augmentation and weighted loss adjustment to mitigate class imbalance. Performance is evaluated using accuracy, precision, F1-score, and AUC metrics, each averaged across 5-fold cross-validation. The reported accuracy is obtained through this cross-validation protocol and benchmarked against baseline models, including standalone EfficientNet-B5 and traditional machine learning approaches. The ViT-enhanced model achieves an accuracy of 87.23%, outperforming the EfficientNet-B5 baseline by a significant margin. To address ethical concerns, the framework incorporates anonymization of personal identities and local edge-based processing to prevent raw data exposure.

تم استخدام مجموعة بيانات (Binary UCF-Crime) للتدريب والتحقق، مع تحقيق إعادة توازن الفئات من خلال زيادة البيانات وتعديل الخسارة الموزونة للتخفيف من مشكلة عدم توازن الفئات. وقد تم تقييم الأداء باستخدام مقاييس الدقة (Accuracy)، والضبط (Precision)، ودرجة F1 (F1-score)، ومساحة المنحنى (AUC)، حيث تم أخذ متوسط كل منها عبر التحقق المتقاطع الخماسي (5-fold cross-validation). وتم الحصول على الدقة المُبلغ عنها من خلال بروتوكول التحقق المتقاطع هذا ومقارنتها بالنماذج الأساسية (Baseline)، بما في ذلك نموذج (EfficientNet-B5) المستقل وأساليب التعلم الآلي التقليدية.

يحقق النموذج المحسّن بـ (ViT) دقة تبلغ 87.23%، متفوقاً على النموذج الأساسي (EfficientNet-B5) بفارق ملحوظ. وللتعامل مع المخاوف الأخلاقية، يدمج الإطار تقنية إخفاء الهويات الشخصية (Anonymization) والمعالجة المحلية القائمة على الحافة (Edge-based processing) لمنع تعرض البيانات الأولية للخطر.

1. Introduction

With the emergence of a surveillance system era, the combination of artificial intelligence (AI), and machine learning (ML) have brought such a revolutionary period in the environment where law enforcement and safety of the population are addressed. A combination of these state-of-the-art technologies has made it possible to detect suspicious activities or any type of criminal behavior in real-time through reviewing real-time Closed-Circuit Television (CCTV) images, and it represents a paradigm shift in crime prevention and intervention. With these emerging technologies constantly evolving and coming into integration with each other, the technology possibilities of AI/ML-based CCTV systems are wide-reaching in their implications in matters of enabling public safety and law enforcement. But alongside these innovations there are concerns of privacy, ethical issues, governance in a transparent manner, and all these factors need to be put in place so that the development of these systems would lead to its responsible introduction in our highly monitored society. CCTV cameras have become ubiquitous and an important feature of modern urban environments, acting as all-seeing and non-problematic watcher that captures the fullest character image throughout the day. However, the manual analysis of uncountable hours of cameras footage has always posed significant challenges to law enforcement agencies [1]. The need of time is safety and security of all individuals in this century. Use of CCTV surveillance systems by high-definition cameras have been learned and implemented in many countries to make their environment secure. Therefore, automated CCTV surveillance systems may act as security agents to determine the behavior of suspicion humans or intruders basing on the acts they perform in the previous stages using the CCTV footages a difficult task [2]. Several types of suspicious activities are normally detected such as consist of killing, looting, molestation, and intensive attacks. Act of murder is an act with an intention of killing a person. Looting is another activity of stealing property of people by employing excessive physical strengths and violence. Molestation is the abuse of sexuality of individuals (man, woman, and children) without their will. This is a horrific criminal act which demonstrates significant implications. Violent fights are unlawful struggles of an individual on another

one to acquire something or to hurt others [3-5]. Detecting suspicious activities and prevention with deep learning is a catchy system. There are numerous law enforcement entities countries around the world are testing deep learning systems in order to protect the populace. The unusual operations are predictable and in need of high-volume data processing that reveals the suspicious patterns that are informative to a law enforcement department. In other cases, suspicious activity is not reported due to external influences by all verticals of the society.

In this paper, a deep learning-based CCTV camera footage model is proposed in order to detect and prevent suspicious activities. The real-world CCTV surveillance system is designed by implementing the EfficientNet-B5-based deep learning model to detect suspicious activities using the vision transformer (ViT) and extract the high-level features from input streams and detect suspicious activities.

2. Literature Review

Suspicious activities detection system is aimed at predicting and preventing the suspicious or abnormal (criminal) activities. Whereas, the traditional non-deep learning methods are good but they work in isolation. It would therefore be an enormous benefit to have a machine that can incorporate the relevant features of traditional methods. A study [6] has applied deep learning to Suspicious Activity Recognition (SAR) systems of the CCTV surveillance in order to minimize the manual monitoring. They employed a variety of models, such as CNN, LSTM-VGG16, ResNet50, and DenseNet121 and measured on the basis of Precision, Recall, and Accuracy.

LSTM+DenseNet121 performed the best registering 91.17% accuracy in identifying suspicious activities. However, the limitation involves false alarms, occlusion, intrusion of privacy, and scarcity of contextual knowledge. In [7], the authors have presented a deep learning-based Super-Resolution method to enhance face recognition of a surveillance video in criminal investigations. The authors used the neural network Very-Deep Super-Resolution (VDSR) to upgrade the video-based extracted low-resolution facial images. The model was trained using the CelebA dataset and tested on QMUL-SurvFace improving the PSNR by 7% and increasing SSIM by 3%. Most significantly, an improved resolution increased face recognition accuracy by 45.7%, yet there are difficulties with processing low-quality inputs of the most extreme content and with issues of different lighting sources. Another study [8], considered past developments in intelligent watchdogs to detect crimes by making use of real-time CCTV via machine learning. The paper covers the topic of object recognition, frames prioritization, and crime identification methods concerning such algorithms as YOLO and their corresponding training datasets.

It underlines the necessity of automated systems to categorize criminal actions and provide an alarm to emergency services. Despite the prospects, current methods remain limited in aspects of real-time processing, the size of data and maintaining high precision in a wide range of crime situations. A study [9], proposed a new deep learning-based surveillance scheme aimed at increasing object recognition, anomaly detection, and privacy. The test hybrid model presented a high-priced basin of 97.5%, recalling 96.3%, and F1-score 96.9%, surpassing noteworthy evaluations, such as YOLOv5, faster R-CNN, and SSS.

The system worked efficiently in recognizing unauthorized entries, suspect objects and threats with a latency of 15 ms and 96.7% real-time accuracy.

Although the system is successful, it has been challenged on scalability, ethics of data processing, and the trade-off between privacy and the effect of surveillance in real time. Another study [10], proposed

Table 1- *Summary of previous work on surveillance systems and CCTV footage.*

#Reference	Model(s) Used	Achieved Accuracy	Methodology	Dataset (if mentioned)
[6]	CNN, LSTM-VGG16, ResNet50, DenseNet121	91.17% (LSTM+DenseNet121)	Deep learning models evaluated for Suspicious Activity Recognition (SAR) to reduce manual monitoring	-
[7]	VDSR (Very-Deep Super-Resolution)	45.7% (face recognition), PSNR 7%, SSIM 3%	Super-resolution enhancement of low-quality surveillance face images	CelebA (train), QMUL-SurvFace (test)
[8]	YOLO (unspecified version)	Not quantified	Crime detection using real-time CCTV and ML for object recognition & alerting	-
[9]	YOLOv5	Not quantified	Real-time object detection with YOLO format dataset prep and fine-tuning	-
[10]	YOLO (with transfer learning & augmentation)	Not quantified	Crime detection/prevention using deep learning with lighting/camera variation handling	-

a CCTV surveillance system using AI and machine learning among other technologies to address the shortcoming of manual surveillance, including time and human error. The system also utilizes real-time analysis of videos to detect abnormal behavior, control crowds and identify patterns, providing immediate threats alerts. It brings out the necessity of high-level infrastructure and data to manage the good implementation. Although it has great potential, this method implicates ethical issues, particularly concerning cybersecurity and privacy of data and proper utilization of surveillance devices.

In [11], authors used deep learning and CCTV by leveraging the YOLO object detection model to enhance crime identification and crime prevention. The suggested YOLO-based framework portrayed real-time detection and categorization of dangerous violations, such as vandalism and theft, as well as effective execution of different illumination and camera settings. Processes that increased generalization were added, including transfer learning and data augmentation. Notwithstanding scalability and pace, the system has issues of privacy, moral aspects, and high-performance hardware requirements. Table 1 (Summary of Previous Work on Surveillance system and CCTV Footage) summaries the background studies with achieved accuracies.

3. Materials and Methods

The proposed method is aimed at identifying harmful human actions in closed-circuit television (CCTV) video using only the video without external sensors. The model acts on video feeds and detects any anomalous activity (e.g., theft, burglary, or assault), thus pushing an urgent message to security personnel. The framework utilizes the powerful spatial features of the EfficientNet-B5 and the ability to model the long-term temporal dependencies and contextual dependence of the Vision Transformer (ViT) to generate accurate detection of anomalies in videos.

3.1. Video Preprocessing

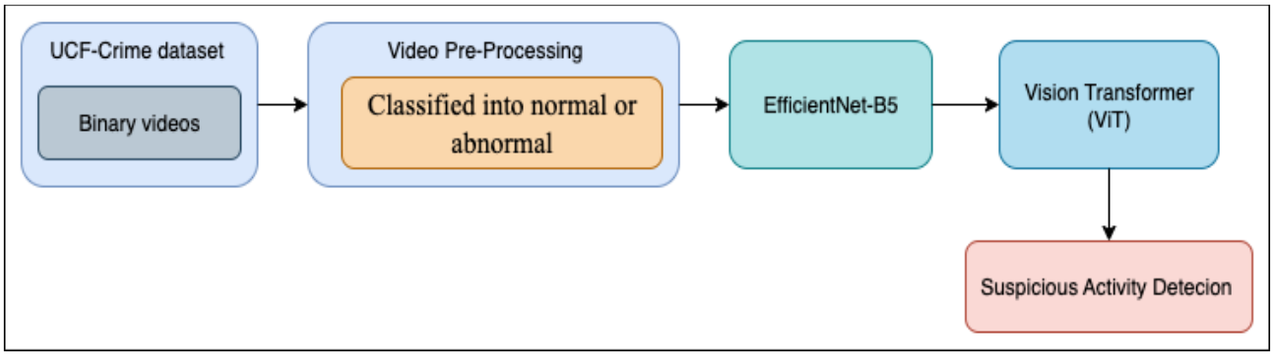

Figure 1- *Block diagram of the proposed suspicious activity detection framework using the UCF-Crime dataset, including video pre-processing, feature extraction with EfficientNet-B5, and classification using Vision Transformer (ViT).*

The CCTV videos are pre-processed by split into a constant number of frames per second. We sample in 30 frames per second experiments, so a 60 second video has 1800 frames. The UCF-Crime binary dataset videos can be classified as normal or abnormal (amalgamating 13 anomaly situations: abuse, arrest, arson, assault, burglary, explosion, fighting, robbery, road accident, shooting, shoplifting, vandalism, and others), as shown in Figure 1. To reduce computation expense and cut down on the absorption of data, we extract concentrated bags of frames with abnormal activities parameters to focus on the informative parts.

The videos will be cut and divided into fixed sized video clips, and unnecessary frames containing uninformative motion of the video will be discarded. This strategy enhances the capability of the system to discriminate against suspicious events by ensuring highest information density of the training samples. Normal data augmentation (such as random cropping, flipping, adjustments in brightness) is used to make the data more variable and eliminate overfitting.

3.2. EfficientNet-B5: Spatial Feature Extraction

EfficientNet-B5, a CNN that favors accuracy-to-complexity trade-offs, is used to extract spatial and visual features of each of the input frames. To pre-train the model with a transfer learning approach, EfficientNet-B5 is first trained on ImageNet to initialize parameters and subsequently fine-tuned using UCF-Crime dataset. The input frames resized to 240 x 240 x 3 are then passed through the convolutional and pooling layers of the EfficientNet-B5, which generates deep spatial feature maps. They are rearranged into a 4-D tensor (n x h x w x c) and forwarded to the Vision Transformer (ViT) to model time.

3.3. Vision Transformer (ViT): Temporal and Contextual Modeling

Although CNNs are very effective at learning spatial patterns, they struggle to learn effects of long-range temporal dependencies. Where the Vision Transformer is different is that it circumvents this shortcoming by viewing the series of frames as a collection of patches, which are, in effect, tokens in Natural Language Processing. The representation of every patch is enhanced by positional embeddings, and as a consequence, ViT can learn spatial-temporal relationships between frames. Self-attention layers with multiple heads extract the motion dynamics and contextual dependencies and perform the idea distinction between normal and abnormal patterns within activity, successfully recognizing them.

Table 2- *Total number of video for binary dataset*

Anomalies	Binary Videos	Anomalies	Binary Videos
Abuse	50	Road Accident	150
Arrest	50	Robbery	150
Arson	50	Shooting	50
Assault	50	Shoplifting	50
Burglary	100	Stealing	100
Explosion	50	Vandalism	50
Fighting	50	Normal	950
Total	400	Total	1500

Table 3- *Evaluation results of the proposed models using 5-fold cross-validation.*

Model	Fold	Accuracy	Precision	F1 Score	AUC
EfficientNet B5	Fold 1	85.80185553	85.80185553	85.80185553	85.80185553
EfficientNet B5	Fold 2	85.90954274	85.91549296	85.90837545	85.90954274
EfficientNet B5	Fold 3	86.05864811	86.06462303	86.05749317	86.05864811
EfficientNet B5	Fold 4	86.21603711	86.21603711	86.21603711	86.21603711
EfficientNet B5	Fold 5	85.89297548	85.89892295	85.89180681	85.89297548
Vision Transformer (ViT)	Fold 1	87.25977469	87.25977469	87.25977469	87.25977469
Vision Transformer (ViT)	Fold 2	87.3674619	87.37365369	87.36641538	87.3674619
Vision Transformer (ViT)	Fold 3	87.17693837	87.17693837	87.17693837	87.17693837
Vision Transformer (ViT)	Fold 4	87.06096753	87.06096753	87.06096753	87.06096753

The output of ViT is flattened and topped with fully connected batch-normalization (BN) and a ReLU activation layer. The binary classification (normal vs. abnormal) objective uses a sigmoid activation on the output that is optimized using binary cross-entropy loss.

3.4. Training Strategy

The specified model was trained on the Binary UCF-Crime dataset with 5-fold cross-validation to ensure such robustness and generalizability. It divided the dataset into training and testing sets of 80% and 20%, respectively, which is in line with similar researches. Learning rate, batch size, and dropout are hyperparameters that were tuned using experiments so as to avoid overfitting. The measurement of the model was precision, accuracy, F1-score, AUC, and precision, as well as a confusion matrix (TP, TN, FP, FN). The total number of videos per-class in the Binary dataset described in the Table 2 (Total number of Video for Binary Dataset). Once trained, the system can process live CCTV streams by segmenting the incoming frames, extracting spatial features using EfficientNet-B5, and analyzing temporal dependencies with ViT. When an abnormal activity was detected, the system triggers an immediate alert

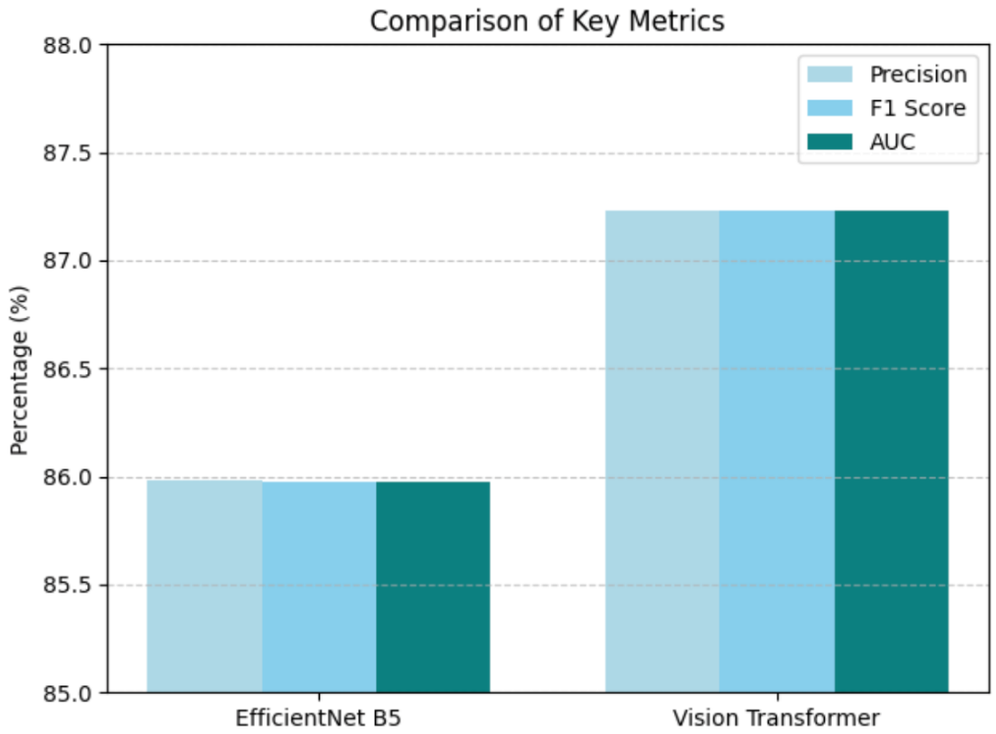

Figure 2- *Comparison of key performance metrics (Precision, F1-score, and AUC) for the base models EfficientNet-B5 and Vision Transformer.*

to security personnel. The architecture is optimized for high throughput, making it suitable for real-time surveillance applications. The model is trained using a binary cross-entropy loss and the model is tested on the standard classification metrics like accuracy, precision, recall, AUC, and F1-score.

4. Results

Evaluation of the proposed model was carried out on dataset UCF-Crime divided into two categories namely binary classification. To guarantee the stability and applicability of the model, 5 folds cross-validation method was used. EfficientNet-B5 and Vision Transformer (ViT) were two models of Deep learning used. The metrics of standard performance would be applied to evaluate performance such as accuracy, precision, F1-score, AUC, as well as confusion matrix True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). EfficientNet-B5 showed similar results in all five folds with average accuracy of 85.97%, precision of 85.97%, F1-score of 85.97%, and AUC of 85.97%. Comparatively, the Vision Transformer performed better than the EfficientNet-B5 with an average of 87.23% accuracy, precision of 87.2%, F1-score of 87.23%, and AUC of 87.23%, as provided in the Table 3 (Evaluation Results of the Proposed Model Using 5-Fold Cross-Validation) and shown in Figure 2. Also, ViT displayed higher values of TP and TN as well as lesser FP and FN implying better classification of both normal and abnormal events.

The proposed model was compared with some of the conventional and modern-day models such as SVM, AuroEncD, MIL, TSN, and UGD-KN particularly the ViT model that significantly surpassed all other models upon benchmarking, as shown in Figure 3. SVM and UGD-KN maintained

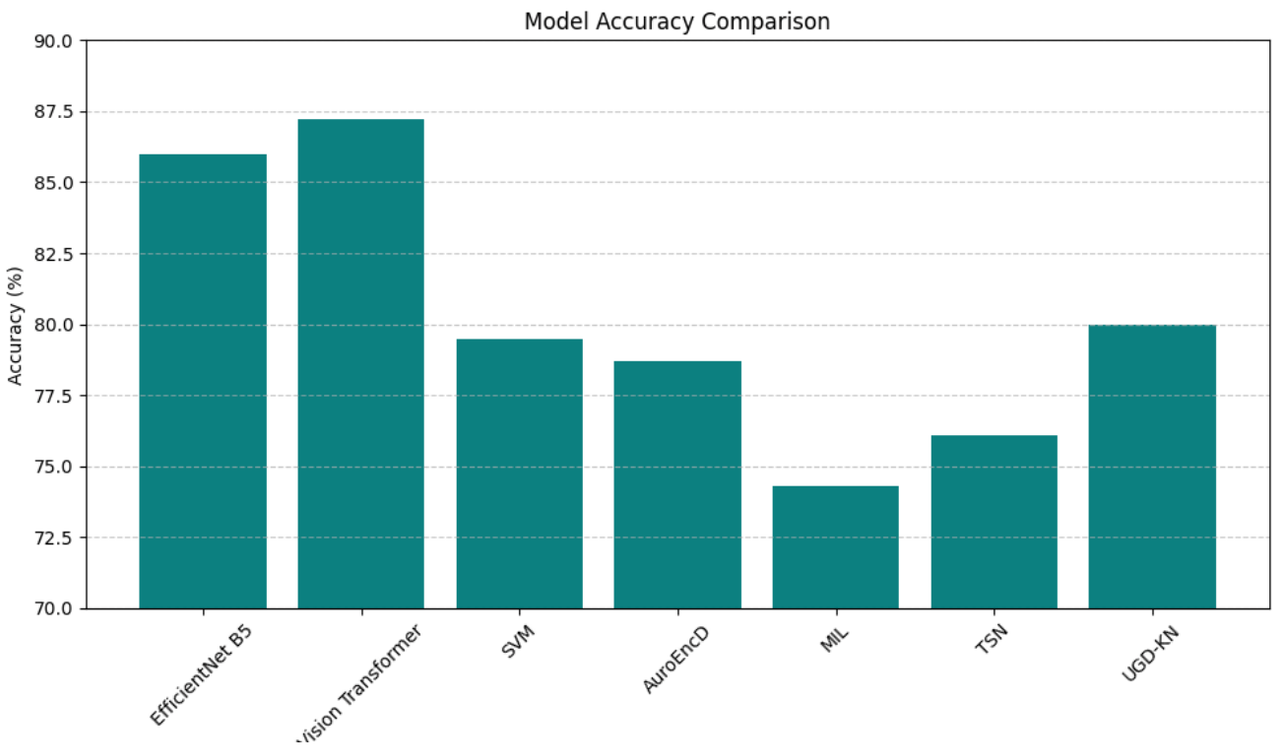

Figure 3- *Comparison of classification accuracy (%) across different models, including EfficientNet-B5, Vision Transformer, SVM, Autoencoder, MIL, TSN, and UGD-XN.*

Table 4- *Models accuracy comparison summary*

Models	Accuracy
EfficientNet B5	85.9
(Vision Transformer (ViT	87.2
SVM	79.5
AuroEncd	78.7
MIL	74.3
TSN	76.1

an average accuracy of 79.5%, 80.00% respectively with other algorithms such as MIL and TSN still trailing behind with 74.3% and 76.1% accordingly, as provided in the Table 4.

5. Discussion

The experimental findings indicate the superiority of the proposed deep learning architecture of real-time recognition of abnormal activities in CCTV surveillance streams. Vision Transformer (ViT), featuring attention-based framework, showed enhanced performance in terms of long-range spatial, contextual relations across video frames resulting in better performance (detection accuracy) than the convolution-based EfficientNet B5 model. The output of the models is consistent across folds and had significant improvements over classical algorithms such as SVM and MIL verify the practicability of using the current deep learning architectures on the tasks of surveillance. Moreover, a very low false positive and false negative of ViT imply that it is neither inaccurate nor unreliable in the parse of reducing false alarms and threats. Along with the great performance, the Vision Transformer necessitates greater computer resources in terms of both training and inference. This trade-off ought to be sought in practice, when the deployment environment does not support much hardware. Still,

its capacity to handle the surveillance tracks in real time and warn of suspicious actions proves the high potential of this system to be applied in practice in smart surveillance and monitoring of general safety.

6. Conclusion

In this paper, we presented a hybrid deep learning model based on EfficientNet-B5 and Vision Transformer (ViT) in a joint effort to improve the autonomous detection of abnormal activities in CCTV footage. The model can detect and label suspicious behaviors in real time and only with video information. By extensive testing on the Binary UCF-Crime dataset, the ViT model performed better, with regard to all the performance measures, suggesting that it is efficient in comprehending the spatial and temporal information that is present in surveillance streams. Though costly in resources, precision and low false alarm rates demonstrate the potential of ViT in real-world application in smart security monitoring applications.

Acknowledgments

The authors would like to express their deep thanks to the Vice Presidency for Scientific Research at Naif Arab University for Security Sciences for their kind encouragement of this work.

Conflict of Interest

The author declares no conflicts of interest.

Source of Funding

The author did not receive financial support for the research, authorship, or publication of this article.

References

1. Murthy JS, Siddesh GM. AI-based criminal detection and recognition system for public safety and security using novel CriminalNet-228. In: Proceedings of the 4th International Conference on Frontiers in Computing and Systems (COMSYS 2023). 2024. p. 1–3.
2. Ali MM, Noorain S, Qaseem MS, ur Rahman A. Suspicious human behaviour detection focusing on campus sites. In: Emerging IT/ICT and AI technologies affecting society. 2022. p. 171–183.
3. Selvaraj A, Selvaraj J, Maruthaiappan S, Babu GC, Kumar PM. L1 norm-based pedestrian detection using video analytics technique. Comput Intell. 2020;36(4):1569–1579.
4. Joshi KA, Thakore DG. A survey on moving object detection and tracking in video surveillance system. Int J Soft Comput Eng. 2012;2(3):44–48.
5. Pustokhina IV, Pustokhin DA, Vaiyapuri T, Gupta D, Kumar S, Shankar K. An automated deep learning-based anomaly detection in pedestrian walkways for vulnerable road users' safety. Saf Sci. 2021;142:105356. doi:10.1016/j.ssci.2021.105356.
6. Saluja D, Kukreja H, Saini A, Tegwal D, Nagrath P, Hemanth J. Analysis and comparison of various deep learning models to implement suspicious activity recognition in CCTV surveillance. Intell Decis Technol. 2023;17(4):917–942. doi:10.3233/IDT-220066.
7. Alkanhal L, Alotaibi D, Albrahim N, Alrayes S, Alshemali G, Bchir O. Super-resolution using deep learning to support person identification in surveillance video. Int J Adv Comput Sci Appl. 2020;11(7):1–10.
8. Singla S, Chadha R. Detecting criminal activities from CCTV by using object detection and machine learning algorithms. In: Proceedings of the 3rd International Conference on Intelligent Technologies (CONIT 2023). 2023. p. 1–6.
9. Sudha I, Ramesh PS, Narang S, Ponmalar A. Deep learning for image and video processing in surveillance systems: advancing security with AI-driven insights. In: Proceedings of the 3rd International Conference on Communication, Security, and Artificial Intelligence (ICCSAI 2025). 2025. p. 492–496.
10. Basthikodi M, Vidya B, Pinto EM, Basith M, Rao SA. AI-based automated framework for crime detection

and crowd management. In: Proceedings of the Second International Conference on Advances in Information Technology (ICAIT 2024). 2024. p. 1–6.

11. Kumar R, Rajpurohit DS, Hanief M, Sharma S, Choudhury T, Sar A. Detection of criminal activities through CCTV by live footage analysis. In: Proceedings of the OPJU International Technology Conference on Smart Computing for Innovation and Advancement in Industry 4.0 (OTCON 2024). 2024. p. 1–6.
12. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2009. p. 248–255.