Response to Decision Letter

Dear Editor and Reviewers:

The authors would like to thank the Editor and Reviewers for providing Minor revisions on our manuscript entitled "Efficient Anomaly Recognition Using Surveillance Videos" (#CS-2022:03:72336:0:1:REVIEW). The authors appreciate the efforts of both the Editor and Reviewers. The feedback provided by reviewers is very useful for revision of the manuscript. We have incorporated majority of the comments of reviewers to improve the quality of the manuscript.
Accordingly, we have highlighted changes in the revised manuscript. We have also provided a file with a point-by-point response to the reviewers' comments. We hope that the revised manuscript is now free of mistakes.

The main revisions are summarized as follow:

- We have revised the Abstract of our manuscript.
- We have highlighted the novelty of manuscript.
- We have explained how existing work is relevant or different to our study.
- We have revised the conclusion of study.
- We have improved English of manuscript.
- We have removed irrelevant studies form Table 3.
- We have added time complexity of our model in manuscript.

Note: Regarding photographs from the UCF Crime dataset, we have removed all photographs of identifiable human subjects from figures.

Reviewer 1

Comment# 1: The novelty of this paper is not clear. The difference between present work and previous Works should be highlighted.

Reply: Thanks for suggestion. We have highlighted the novelty in Introduction section to present contribution of this study. The revised section is added below for quick reference.

discussed in (Maqsood et al., 2021). This work attempts to overcome the highlighted issues through frame level data annotations and spatial augmentation, which improved the performance of video based anomaly recognition. This study aimed at improving spatiotemporal feature based learning as these features provide useful information to process anomalous videos. The following are the contributions of this paper:

- We attempted to address the issue of the high resource requirement of the anomaly recognition method and proposed a lightweight, resource-efficient real-time streaming TAR framework that can be embedded on a simple machine like a central processing unit (CPU).
- We proposed to use temporal learning via partial shift operation to improve spatiotemporal feature based learning. It enables frames to share their learning among adjacent frames and reduce the cost of processing. Moreover, it helps in building feature maps based on high activity areas that support the classification task of anomaly recognition.
- Our framework is capable of performing online anomaly recognition and it allows six simultaneous screens on a CPU-based system while using fewer parameters (2.2M), FLOPs (0.564GFLOPs), model size (0.6Mb) and low latency overhead which proves it to be resource efficient approach.
- Our model achieves 7.87% and 2.47% increased state-of-the-art accuracy with ResNet-50 and MobileNetV2 respectively on UCF-Crime dataset.

Comment# 2: The author needs to change the abstract and focus more on the problem domain. Before the paper's contributions, the author could precisely include the need to develop the proposed method.

Reply: Thanks for suggestion. Reviewer mentioned to improve the abstract of study and considering their suggestions we have revised our abstract. The revised section is added below for quick reference.

ABSTRACT

Smart surveillance is a difficult task that is gaining popularity due to its direct link to human safety. Today, many indoor and outdoor surveillance systems are in use at public places and smart cities. Because these systems are expensive to deploy, these are out of reach for the vast majority of the public and private sectors. Due to the lack of a precise definition of an anomaly, automated surveillance is a challenging task, especially when large amounts of data, such as 24/7 CCTV footage, must be processed. When implementing such systems in real-time environments, the high computational resource requirements for automated surveillance becomes a major bottleneck. Another challenge is to recognize anomalies accurately as achieving high accuracy while reducing computational cost is more challenging. To address these challenge, this research is based on the developing a system that is both efficient and cost effective. Although 3D Convolutional Neural Networks have proven to be accurate, they are prohibitively expensive for practical use, particularly in real-time surveillance. In this paper, we present two contributions: a resource-efficient framework for anomaly recognition problems and two-class and multi-class anomaly recognition on spatially augmented surveillance videos. This research aims to address the problem of computation overhead while maintaining recognition accuracy. The proposed Temporal Based Anomaly Recognizer (TAR) framework combines a partial shift strategy with a 2D convolutional architecture-based model, namely MobileNetV2. Extensive experiments were carried out to evaluate the model's performance on the UCF Crime dataset, with MobileNetV2 as the baseline architecture; it achieved an accuracy of 88% which is 2.47% increased performance than available state-of-the-art. The proposed framework achieves 52.7% accuracy for multiclass anomaly recognition on the UCF Crime2Local dataset. The proposed model has been tested in real-time camera stream settings and can handle six streams simultaneously without the need for additional resources.

Comment# 3: The author could better explain how "Related works" is actually related to the current study. It is unclear to the reader how the manuscript is similar to or differs from these related works. https://doi.org/10.3390/math10050733; https://doi.org/10.1016/j.imavis.2021.104229

Reply: Thanks for feedback. Reviewer has highlighted two studies to explain how these are related or different from our work. We are performing anomaly recognition using surveillance videos which is based on spatiotemporal feature based learning. In our work, we have attempted to improve the spatiotemporal feature learning in such a way that it will utilize least possible resources but provide efficient recognition as mentioned in appended section.

Most of the existing work is focused on increasing accuracy of anomaly recognition that usually based on deep networks with high computational requirements. Recently, researchers are considering cost effective anomaly recognition systems, which still requires many improvements. Especially, online

anomaly recognition process long untrimmed sequences, which is a challenging task. Our study tries to reduce the computational cost of anomaly recognition process without compromising on accuracy. It processes online streams and resolves the problem of long untrimmed video. We have performed anomaly recognition which identifies whether there is anomaly in data or it is a normal sequence. We used partial shift strategy to incorporate spatiotemporal learning in our framework which improves the accuracy of 2D CNN based architecture to outperform 3D CNN based methods.

About mentioned studies:

1. The mentioned study "Intelligent video anomaly detection and classification using faster RCNN with deep reinforcement learning model" is based on anomaly detection from surveillance videos. It performs anomaly classification on UCSD to detect whether it is

anomaly or no anomaly. Similarly, our study is also based on recognition of anomalies from surveillance videos and we have used UCF Crime dataset for the purpose of experiment.

2. The mentioned study "Computational Intelligence-Based Harmony Search Algorithm for Real-Time Object Detection and Tracking in Video Surveillance Systems" is based on detection and tracking of objects. Whereas we are not addressing the problems relevant to object detection and tracking as our study aimed at improving anomaly recognition

Comment# 4: The Experiment part of the paper is good, however, the authors should include some. Image examples in order to make the experiments and their results more understandable.

Reply: Thanks for your feedback. The reviewer suggested to add image examples to explain results which is a good suggestion. The proposed model is evaluated using UCF Crime dataset which has anomalies which are directly related to a human subject. We have used the dataset for experimental purposes only but we do not hold human consent to publish their pictures. According to publishing policy of the journal, to publish a human picture we required an official consent. However, we have added some results in Figure 10, which do not involve visibility of human subject.

Comment# 5: The conclusion of the proposed work is discussed with limited content and the achieved performance value is more efficient when compared to the existing methods. The conclusion must discuss in detail the limitations of current knowledge, and the overall importance of the work.

Reply: Thanks for the suggestion. The reviewer mentioned that proposed study claimed efficient performance but in conclusion it is explained with limited content. Reviewer is concerned about limitation and importance of current study. We have added conclusion to report findings of the study. We have also added limitations and future direction in our manuscript to presents limitation of the work. Considering reviewer's suggestion, we have tried to improve the details as much as possible.

CONCLUSION

Automated surveillance is popular area which is continuously improving over the time. Such systems are designed to process huge amount of data which requires a lot of resources which is a challenging requirement. Whereas in practical scenarios, time, and cost both are critical to handle and hence a system needs a lot of computation time and resources to serve the purpose. So, this research aimed to addressing these problems through providing a resource-efficient, high-performing system for anomaly recognition. Increased number of CCTV generates vast amount of unlabelled video data and its labeling is a difficult task. Moreover, large amount of data requires substantial computational resources to process it. This study provides a lightweight and cost-effective approach for anomaly recognition in terms of memory consumption, processing parameters, and computation time that is essential for a low-resource systems, such as CPU-based systems. The proposed framework is based on 2D convolutional architecture (2D CNN), with a spatiotemporal feature extractor which functions as a partial shift that learns and distributes temporal information among its neighborhood frames. Whereas MobileNetV2 baseline performs spatial feature extraction, which is then combined with temporal learning to perform anomaly recognition. Our proposed framework works with low latency rate of 12.01 milliseconds which makes it effective for performing online video recognition and handle up to six streams at once. On UCF Crime dataset, the proposed framework achieves an accuracy of 88% for binary anomaly recognition problem with time complexity of 0.198 seconds. On UCF Crime2Local dataset, the proposed framework achieves accuracy of 52.7 percent for a multi-class problem. The model outperforms previous models in terms of computational parameters requiring 2.2 M parameters and 0.564 GFLOPs with MobileNetV2 as the baseline architecture. Moreover, our proposed framework has achieved an increased accuracy of 2.47% on UCF Crime dataset with reduced computational requirement. Overall, it performs well in a lot of aspects and can be used for realtime recognition but it can be further improved. Some limitations of this study are highlighted in below section 'to consider in future.

LIMITATIONS AND FUTURE DIRECTIONS

We have proposed a resource-efficient anomaly recognition system that effectively performs recognition tasks, but it is not evaluated for object level detection and tracking. Object detection and tracking can significantly improve security surveillance. We believe that with minor modifications to the current model, it could be useful for detection and tracking as well. Our model performs recognition with adequate speed efficiency, but its early response efficiency can be investigated further. It requires validating a model's ability to perform recognition on a small sample of incoming frames as soon as possible to improve system's response time. Surveillance systems are designed to perform anomaly recognition as precisely as possible, but there is an additional issue posed by the false positive rate, which can compromise system reliability.To increase the usefulness of our model, we will strive to reduce the false positive rate as much as possible.

Comment# 6: The English should be polished.

Reply: Thank you for feedback. We have revised our draft for English of the document and we have tried to improve it through revising improper sentences.

Reviewer 3

Comment# 1: In the experimental section, authors report comparison performance of proposed model and SOTA techniques but the accuracy of these techniques on UCF-Crime dataset not UCF Crime2Local. Author should perform ablation studies of these techniques or remove these results from Table 3.

Reply: Thanks for feedback. We have used UCF Crime2Local (contains 6 classes from UCF Crime) to perform multiclass anomaly recognition but in literature it is mainly used for binary or two class recognition problem. We have added few studies on multiclass but as reviewer mentioned these are based on UCF Crime (13 classes). We added them as these were their best achieved accuracy and we have compared it with our average accuracy. As reviewer mentioned we have removed these from Table 3 as these do not provide an exact comparison.

To demonstrate performance of our proposed method, we have added experimental results of baseline model to compare with proposed model. We have added Table 3 below for quick reference.

Method	Accuracy
ResNet50	45.20%
MobileNetV2	41.9%
Proposed method	52.70%

Table 3. Performance of proposed framework (TAR) for multiclass anomaly recognition

Comment# 2: The authors need to revise the sentence in line 466 and avoid the repetition of words.

Reply: Thanks for feedback. We have checked the review pdf which includes line number but we are unable to locate the mistake. Following the line number, it appears in references section as added below. However, we have revised manuscript to check if there is any repetitions but we will appreciate if reviewer can mention the mistake to improve it further.

451 Supporting Project No. (RSP 2022R509) at King Saud University, Riyadh, Saudi Arabia.

462 **REFERENCES**

463 Azizjon, M., Jumabek, A., and Kim, W. (2020). 1d cnn based network intrusion detection with normaliza-
464 tion on imbalanced data. In *2020 International Conference on Artificial Intelligence in Information*
465 *and Communication (ICAIIC)*, pages 218–224. IEEE.
466 Biradar, K., Dube, S., and Vipparthi, S. K. (2018). Dearest: Deep convolutional aberrant behavior
467 detection in real-world scenarios. In *2018 IEEE 13th International Conference on Industrial and*
468 *Information Systems (ICIIS)*, pages 163–167. IEEE.
469 Canizo, M., Triguero, I., Conde, A., and Onieva, E. (2019). Multi-head cnn–rnn for multi-time series
470 anomaly detection: An industrial case study. *Neurocomputing*, 363:246–260.

Comment# 3: The authors claim that "We have implemented our implementation using NVIDIA Jetson Nano" but they not reported time complexity of the model. The authors need to add time complexity and qualitative analysis of the model. For ease see of the authors see the link [1].

Reply: Thanks for feedback. The proposed model has been extended for a desktop application and to analyze its real time recognition we have used NVIDIA Jetson Nano to demonstrate that proposed model can be deployed on edge devices. We have added performance of the system along with latency rate and time complexity. The respective section is added below for quick reference.

positive rate (FPR). As shown in figures, ResNet50 achieves better performance in terms of recognition rate as compared to MobileNetV2 but we selected MobileNetV2 as final baseline architecture as our agenda is to provide resource efficient framework. Figure 8 and 9 shows accuracy and loss curve of proposed anomaly recognition framework on UCF crime dataset with MobileNetV2 as baseline model. We have implemented our implementation using NVIDIA Jetson Nano to compare our performance for edge devices with CPU and GPU. Moreover, our method provides fast computation speed with time complexity of 0.198 seconds. While discussing performance of our method, its resource efficiency is also notable. There is a difference in the number of frames processed per second, so GPU based systems

are fast in prediction. CPU based systems usually have slow processing speed, but it does not affect the recognition rate. The prevalence of CPU-based systems in real-time scenarios is the primary motivation for proposing a resource-efficient system. Our model work efficiently on both CPU and GPU and produces low latency of 42.1 ms and 12.01 ms, respectively. That is why, we have claimed that our proposed framework (TAR) is resource efficient.