East African Journal of Information Technology
eajit.eanso.org
Volume 7, Issue 1, 2024
Print ISSN: 2707-5346 | Online ISSN: 2707-5354
Title DOI: https://doi.org/10.37284/2707-5354

EAST AFRICAN NATURE & SCIENCE ORGANIZATION

Original Article

Enhancing Public Safety Through Advanced Video Analysis: A Conv-LSTM-SVM Model for Violence Detection in Surveillance Footage

Samuel Muigai Muiruri[1] Dr. Mark Okong'o, PhD[2] & Dr. David Mwathi, PhD[2]*

[1] Tharaka Nithi County Government, P. O. Box 10, 60406, Kathwana, Kenya.
[2] Chuka University, P. O. Box 109 – 60400, Chuka, Kenya.
* Correspondence ORCID ID: https://orcid.org/0000-0003-4775-4486; Email: smuigai@tharakanithi.go.ke

Article DOI: https://doi.org/10.37284/eajit.7.1.2117

Date Published:

16 August 2024

Keywords:

Violence Detection, Video Analysis, Surveillance Systems, Deep Learning, Computer Vision, Artificial Intelligence, Machine Learning.

ABSTRACT

This study pioneers a new method for detecting violence in surveillance videos, addressing a major challenge in public safety and video analysis. The study presents a hybrid model that uses Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and Support Vector Machines (SVM) to detect violent incidents in video data. The Convolutional Long-Short-Term Memory and Support Vector Machines (Conv-LSTM-SVM) model combines CNN spatial feature extraction, LSTM temporal dependency modelling, and SVM classification power. A pre-trained DenseNet121 model extracts spatial information efficiently via transfer learning from large datasets in the proposed architecture. An LSTM layer captures temporal dynamics needed to understand video sequence activity development, while an SVM with a Radial Basis Function (RBF) kernel creates complex decision boundaries in the feature space with many dimensions for the final categorization. The model was developed, trained and tested using the Keras library running on TensorFlow, using an experimental research design. The model is tested using two well-known datasets: the UCF Crime dataset, which contains 1900 surveillance clips of 13 classes of violent situations, and the RWF-2000 dataset, which analyses real-world fighting. The proposed model is the best in its class, outperforming CNN, LSTM, and Conv-LSTM models with 97.3% accuracy on the UCF Crime dataset. Cross-dataset validation yielded 92.5% accuracy on the RWF-2000 dataset without changes, demonstrating robust generalization. The study also considers how public safety could be improved by processing several video streams in real time and reducing false alarms. An examination of the ethical challenges and restrictions of automated surveillance systems, such as privacy, biases, and human supervision was also done. This research uses advanced video analysis to improve public safety by creating more efficient and adaptive surveillance systems.

APA CITATION

Muiruri, S. M., Okong'o, M. & Mwathi, D. (2024). Enhancing Public Safety Through Advanced Video Analysis: A Conv-LSTM-SVM Model for Violence Detection in Surveillance Footage. *East African Journal of Information Technology*, *7*(1), 202-214. https://doi.org/10.37284/eajit.7.1.2117

CHICAGO CITATION

Muiruri, Samuel Muigai, Mark Okong'o and David Mwathi. 2024. "Enhancing Public Safety Through Advanced Video Analysis: A Conv-LSTM-SVM Model for Violence Detection in Surveillance Footage". *East African Journal of Information Technology* 7 (1), 202-214. https://doi.org/10.37284/eajit.7.1.2117.

HARVARD CITATION

Muiruri, S. M., Okong’o, M. & Mwathi, D. (2024) “Enhancing Public Safety Through Advanced Video Analysis: A Conv-LSTM-SVM Model for Violence Detection in Surveillance Footage”, *East African Journal of Information Technology*, 7(1), pp. 202-214. doi: 10.37284/eajit.7.1.2117.

IEEE CITATION

S. M. Muiruri, T. Okong’o & D. Mwathi “Enhancing Public Safety Through Advanced Video Analysis: A Conv-LSTM-SVM Model for Violence Detection in Surveillance Footage”, *EAJIT*, vol. 7, no. 1, pp. 202-214, Aug. 2024.

MLA CITATION

Muiruri, Samuel Muigai, Mark Okong’o & David Mwathi “Enhancing Public Safety Through Advanced Video Analysis: A Conv-LSTM-SVM Model for Violence Detection in Surveillance Footage”. *East African Journal of Information Technology*, Vol. 7, no. 1, Aug. 2024, pp. 202-214, doi:10.37284/eajit.7.1.2117.

INTRODUCTION

The extensive use of surveillance equipment in modern life has generated a lot of video data. Thus, complex automated systems are needed to assess and monitor this data. A Full High Definition (FHD) camera would produce eighty-six Gigabytes of data per day at eight Mbps and 24 hours of recording (Norris *et al*., 2002). Detecting violence in surveillance footage is a challenging task. The challenge stems from complex human relationships, different environmental conditions, and ever-changing scenarios. Existing approaches struggle to capture geographical and temporal properties, which are essential for violence detection. Despite promising findings, many automated systems are still plagued by errors and poor processing. Thus, a durable, advanced spatial-temporal system is urgently needed. The hybrid model presented by the study improved surveillance video violence identification accuracy and efficiency by combining diverse methodologies (Welsh & Farrington, 2009). The challenge has spurred the development of automated video analysis techniques, with the detection of violence emerging as a critical use case.

According to Gao *et al*. (2016), detecting violence in surveillance footage is a difficult task that requires the use of sophisticated technological techniques. The automated detection of violent events in real-time can significantly enhance the efficiency of surveillance systems, facilitating prompt reactions to emergencies and optimizing the allocation of security resources (Ullah *et al*., 2019). However, this endeavor presents several challenges, including the diverse nature of aggressive behaviors, changing environmental conditions, and the need for immediate processing.

Traditional approaches to detecting violence have traditionally relied on manually crafted features or simple machine learning models, which struggle to fully capture the complex nature of violent actions in many real-world scenarios (Zhang *et al*., 2017). The current difficulty can be better tackled by leveraging the advancements in deep learning, which offer the potential to create detection systems that are more accurate and robust.

This study presents a hybridized approach to detect violence in surveillance footage by leveraging the capabilities of different machine learning algorithms. The proposal entails the integration of Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and Support Vector Machines (SVM) into a hybrid model. This model aims to accurately identify violent incidents in Closed Circuit Television (CCTV) video data. This strategy aims to overcome the limitations of individual procedures by using their complementary advantages.

RELATED WORK

Traditional Approaches to Violence Detection:

Initially, attempts to automatically identify instances of violence in video footage primarily relied on manually created features and traditional machine learning methods. These strategies often focused on extracting key visual and auditory features that were thought to indicate violent actions. Khan *et al*. (2019) proposed a method that utilises motion properties, notably motion intensity and acceleration, for the purpose of identifying and categorizing violent images. Their methodology demonstrated a moderate degree of success, but it was limited in its ability to capture complex patterns of violence. Li *et al*. (2015) developed a system that analyzed the motion of human limbs to detect violent acts, demonstrating the effectiveness of using motion-related features to identify hostility.

In addition, scholars have explored audio-based methodologies. For example, Naphade and Huang (2002) developed a system that combined auditory and visual features to identify instances of violence in films. Their research focused on the potential of multimodal analysis, however it was primarily assessed using planned material rather than real surveillance footage.

While the traditional approaches were important in their early attempts, they encountered challenges in addressing the diverse and complex character of real-life violent events. Consequently, researchers initiated an exploration of more advanced machine-learning methodologies.

Deep Learning Approaches:

Deep learning has greatly revolutionized certain areas of computer vision, such as the identification of violence. For instance, Convolutional Neural Networks (CNNs) have shown remarkable effectiveness in obtaining spatial information from images and video frames. Accattoli *et al*. (2020) presented a novel approach using a convolutional neural network (CNN) to identify instances of violence in surveillance videos. Their methodology demonstrated significant improvements in comparison to traditional techniques. Their methodology utilized a three-dimensional convolutional neural network (3D CNN) framework to effectively extract and analyze spatial and temporal information from video recordings.

Building on the accomplishments of CNNs, researchers began exploring architectures that may better capture the temporal patterns in video data. Long Short-Term Memory (LSTM) networks, a particular variant of recurrent neural networks, have shown promise in this regard. Pawar and Attar (2018) introduced a multi-stream LSTM network specifically tailored to identify occurrences of violence. This network had the ability to interpret both spatial and motion information. Their approach highlighted the need of including temporal correlations in video analysis tasks.

Hybrid Models and Recent Advancements:

Researchers have employed hybrid models for violence detection because they may leverage the complementary characteristics of multiple neural network architectures. Khan *et al*. (2019) proposed a two-stream approach that combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to detect occurrences of violence in hockey videos. Prior to integrating the characteristics for final classification, their model incorporated a separate processing of geographical and temporal information. Convolutional LSTM (Conv-LSTM) models have recently gained attention for their ability to simultaneously capture spatial and temporal data. Mabrouk and Zagrouba (2017) demonstrated the effectiveness of Conv-LSTM networks in identifying violence in surveillance films, achieving excellent results on multiple established datasets.

Nevertheless, the challenge of developing robust and flexible models for detecting violence in many real-world scenarios remains arduous. The foundation of this research is in the most recent breakthroughs in the field, where a hybridized architecture that harnesses the benefits of CNNs, LSTMs, and SVMs is implemented. This architectural design seeks to enhance the efficiency of violence detection beyond the current boundaries. The research expands on prior studies by integrating Support Vector Machine (SVM) classification into a Convolutional Long Short-Term Memory (Conv-LSTM) framework. The objective is to integrate the benefits of deep learning and traditional machine learning techniques.

METHODOLOGY

Overview of the Proposed Model:

The proposed Conv-LSTM-SVM model for violence identification in surveillance footages integrates the capabilities of three robust machine learning techniques: Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Support Vector Machines (SVMs). This hybrid architecture is specifically developed to efficiently extract both spatial and temporal characteristics from video input, while ensuring strong classification accuracy.

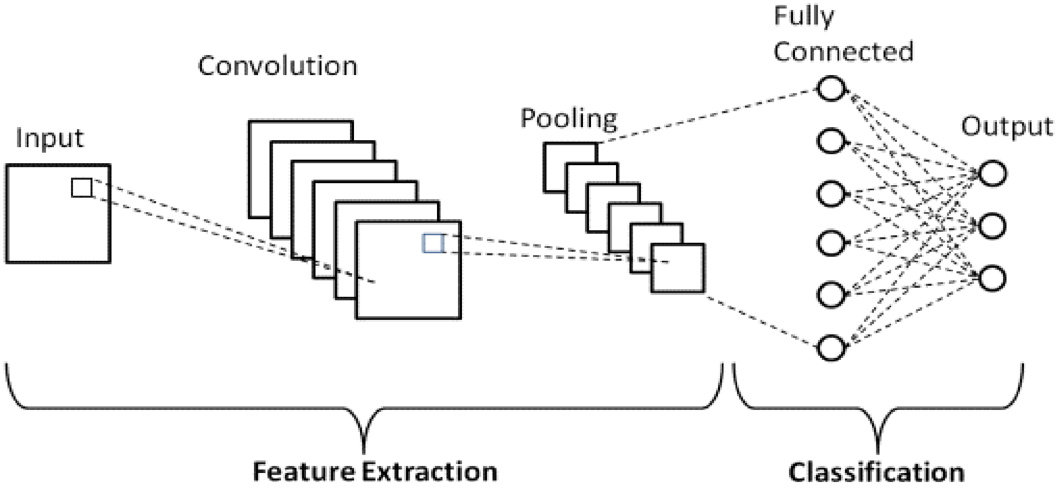
The model consists of three main components:

- CNN Component: Instead of starting from scratch, a pre-trained DenseNet121 model is employed for spatial feature extraction.
- LSTM Component: An LSTM layer is used for temporal feature learning.
- SVM Component: An SVM classifier with Radial Basis Function (RBF) kernel for final classification.

Convolutional Neural Network (CNN) Component:

In the proposed system, a pre-trained DenseNet121 model is used as the convolutional neural network (CNN) component. DenseNet121 is a convolutional neural network architecture that presents a unique connectivity pattern among its layers. DenseNet121 uses a dense, feed-forward architecture where each layer is connected to every other layer, in contrast to the conventional feed-forward approach where each layer only receives input from the previous layer (Zhang *et al.*, 2019). The high level of interconnectivity in the network enables the reuse of feature maps, resulting in more efficient utilization of parameters and a reduction in the overall number of parameters needed compared to alternative architectures. The basic architecture of the CNN is illustrated in Figure 1.

Figure 1: CNN basic architecture



The high level of connectivity also enhances the transmission of information and gradients throughout the network during both the forward and backward passes. By establishing direct connections between layers and all succeeding layers, the movement of gradients from later layers to earlier layers during backpropagation is facilitated. This feature enables DenseNet121 to be trained with more efficiency, particularly on deeper networks where the issue of vanishing gradients can arise Zhang *et al.* (2021).

DenseNet121 achieves state-of-the-art performance on image classification benchmarks by efficiently reusing feature maps and improving gradient flow. It accomplishes this despite utilising fewer parameters and being more computationally economical than earlier deep learning architectures (Pattanaik *et al*., 2022). The dense connection pattern is a crucial innovation that distinguishes DenseNet121 from other designs of convolutional neural networks. Features are extracted from the pre-trained DenseNet121 by removing its final classification layer and utilising it as a feature extractor. By utilising the extensive spatial characteristics acquired from vast datasets such as ImageNet, they can effectively be employed in the particular objective of identifying instances of violence.

Long Short-Term Memory (LSTM) Component:

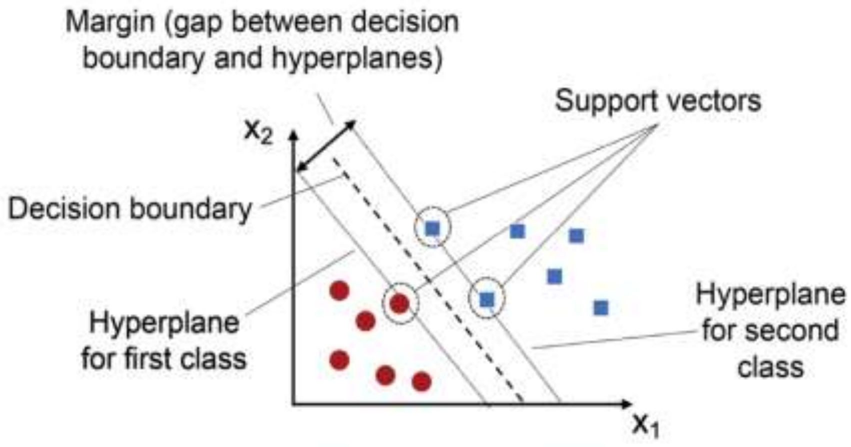
Long Short-Term Memory (LSTM) is a specific kind of recurrent neural network (RNN) that is specifically developed to solve the issue of the vanishing gradient problem that is commonly found in standard RNNs (Sak *et al*., 2014). LSTMs excel at analyzing and forecasting data that follows a time sequence, making them suitable for jobs like recognizing handwriting, understanding speech, translating languages, controlling robots, and managing healthcare.

An essential element of a Long Short-Term Memory (LSTM) is the cell, which has the ability to retain information for any given duration of time. According to Sainath *et al*. (2015), the cell is controlled by three gates: the input gate, forget gate, and output gate. The forget gate chooses which information to exclude from the prior state, the input gate selects which new information to retain in the current state, and the output gate regulates which pieces of information in the current state to produce as output. The LSTM model is adept at accurately capturing both the spatial characteristics present in video frames and the temporal connections between these frames (Zhang *et al*., 2018). This ability is of utmost importance when it comes to accurately detecting instances of violent events.

Support Vector Machine (SVM) Component:

For the last stage of classification, a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel is utilized. Support Vector Machines (SVMs) are renowned for their capacity to identify optimal separation hyperplanes in spaces with a large number of dimensions, as illustrated in Figure 2 below. This makes them highly suitable for intricate classification problems (Cristianini & Shawe-Taylor, 2000). The RBF kernel enables the SVM to generate decision boundaries that are not linear, hence enhancing the model’s capacity to differentiate between tiny differences in violent and non-violent actions.

Figure 2: Hyperplanes illustration in SVM


Margin (gap between decision boundary and hyperplanes)
Support vectors
x_2
Decision boundary
Hyperplane for first class
Hyperplane for second class
x_1

According to Scholkopf *et al*. (1997), the main advantages of using Support Vector Machines (SVM) with a Radial Basis Function (RBF) kernel are its efficacy in high-dimensional spaces, its adaptability in representing intricate decision boundaries, and its capability to perform well even with a smaller number of samples compared to the number of features. The SVM with RBF kernel is widely favored and highly effective for a diverse set of machine learning problems due to its distinctive features.

Model Architecture

The Convolutional Long-Short-Term Memory and Support Vector Machines (Conv-LSTM-SVM) model developed combines deep learning and classical machine learning approaches to effectively analyze video sequences and detect instances of violence. The architecture of the model is intricately tailored for this purpose. The model primarily accepts video sequences as input, which are represented as five-dimensional tensors. These tensors contain information about the batch size, sequence length, height, width, and channel of the video frames.

The architecture’s primary component is the DenseNet121, a robust convolutional neural network that has been pre-trained on extensive image datasets. This network analyses each frame of the input sequence separately, collecting detailed geographical characteristics. The result of this stage is a collection of high-dimensional feature maps for every frame in the sequence. After extracting the features, a crucial operation is carried out to reshape the data. This phase involves rearranging the extracted features in order to maintain the chronological order of the input sequence. This prepares the data for temporal processing in the following steps.

The modified characteristics are subsequently inputted into a Convolutional LSTM (Conv-LSTM) layer, which is a crucial advancement in the design of the proposed model. The purpose of this layer is to capture the relationships between different locations and moments in the data, enabling the model to comprehend the changes in visual characteristics over time - a critical factor in identifying violence in video sequences.

Following the Conv-LSTM processing, a Global Average Pooling (GAP) operation is implemented. This phase greatly decreases the spatial dimensions of the features, condensing the most important information and decreasing the computing cost of subsequent procedures. In order to improve the model’s ability to generalize and avoid overfitting, a dropout layer with a rate of 0.4 is included. This regularization strategy, known as dropout, stochastically deactivates a fraction of neurons throughout the training process, so compelling the network to acquire more resilient and generalized characteristics.

The design also includes a Support Vector Machine (SVM) that utilises a Radial Basis Function (RBF) kernel as its final component. This classifier utilises the pooled and regularized features to carry out the ultimate violence detection task. Utilising an SVM, especially with an RBF kernel, enables the creation of intricate and non-linear decision boundaries in the feature space with a high number of dimensions. This capability has the potential to capture subtle patterns that distinguish violent sequences from non-violent ones. This architecture skillfully integrates the advantages of deep learning for extracting features and modelling temporal patterns with the strong classification abilities of SVMs. The outcome is a potent and effective model for detecting violence in surveillance footage.

Training Process

The training approach for the Convolutional Long-Short-Term Memory and Support Vector Machines (Conv-LSTM-SVM) model is carefully crafted to optimize performance and guarantee strong generalization. The choice of loss function, Binary Cross-Entropy (BCE), is excellently suited for violence detection due to its alignment with the binary character of the task. The loss function accurately measures the difference between the model’s predictions and the actual data, giving us a precise aim for optimization.

In order to negotiate the intricate terrain of loss, the Adam optimizer, which is an advanced algorithm that adjusts the learning rate for each parameter is utilized. The Adam optimization algorithm is initialized with a learning rate of 0.001, which is carefully selected to achieve a trade-off between the speed of convergence and stability. This configuration facilitates effective training, allowing the model to rapidly reach optimal parameters while avoiding oscillations or premature convergence.

The training process utilises a batch processing approach, where data is processed in groups of 32 batches. The selection of this batch size is intentionally made to optimize the balance between computing efficiency and model stability. It enables efficient parallelization on contemporary Graphics Processing Units (GPUs) while ensuring a enough number of samples for dependable gradient estimates.

In order to address the issue of overfitting and improve the model's capacity to generalize, a dropout layer with a rate of 0.4 is incorporated just before the SVM classifier. This method involves the random deactivation of a significant number of neurons throughout the training process. This compels the network to acquire more resilient and comprehensive characteristics that are not heavily dependent on any individual neuron.

Data augmentation is an essential component of the training process. A wide range of augmentation techniques are employed, such as random flipping, rotation, and color jittering. These modifications artificially increase the size of the dataset, allowing the model to encounter a broader range of visual representations. This not only enhances the overall quantity of the training data but also enhances the model's ability to handle variances in real-world situations.

Finally, transfer learning is implemented to enhance the initial performance of the model. The DenseNet121 component is initialized using weights that have been pre-trained on the extensive ImageNet dataset. By employing this method, the model may leverage the knowledge gained from a wide variety of photos, which are subsequently adjusted to optimize their performance in detecting instances of violence. The exchange of knowledge greatly expedites training and frequently results in exceptional performance, particularly in situations when there is a scarcity of data related to a certain subject.

Implementation Details

The implementation of the model utilizes TensorFlow, a widely-used deep learning framework. The Keras library is utilized for the pre-trained DenseNet121 model, while the Conv-LSTM layer is implemented following the methodology outlined in the original research by Xingjian *et al*. (2015). The Support Vector Machine (SVM) component is constructed utilizing the scikit-learn library. It is seamlessly incorporated into the TensorFlow model through a custom layer. The training is conducted on NVIDIA Tesla V100 GPUs, employing mixed precision training to enhance computational performance. To prevent overfitting, early stopping is utilized based on the validation performance, across 10 epochs.

EXPERIMENTAL SETUP AND RESULTS

Datasets

During the development, training and validation of the Convolutional Long-Short-Term Memory and Support Vector Machines (Conv-LSTM-SVM) model, two extensive and varied datasets that accurately represent the intricacies of real-life violence detection situations are employed. The UCF Crime Dataset is the main dataset. This dataset is comprehensive and includes 1900 surveillance recordings from real-world situations. It provides a diverse range of unusual events over 13 different categories, with a particular emphasis on violence. According to Sirisha and Chandana (2023), the videos in this collection exhibit substantial variation in duration, ranging from concise 30-second segments to extensive sequences spanning several minutes. The variability in duration poses a valuable challenge, compelling the model to adjust to various temporal scales of violent occurrences. The UCF Crime Dataset is crucial for the training and testing purposes, as it provides a diverse range of violent scenarios that ensures the model can manage a wide spectrum of situations.

In order to thoroughly evaluate the ability of the model to generalize, the model included an additional dataset known as the RWF-2000. The collection comprises 2000 video segments, each carefully extracted from genuine surveillance camera footage (Cheng *et al*., 2021). In contrast to the UCF Crime Dataset, the RWF-2000 dataset ensures a uniform clip duration of 5 seconds, providing a standardized time frame for study. The dataset's particular emphasis on fight scenarios complements the wider range of the UCF Crime Dataset, enabling the assessment of the model's effectiveness on a more focused subset of violent incidents. By utilizing a dual-dataset approach, the objective is to conduct a thorough and meticulous assessment of the model, guaranteeing its efficacy in a wide range of violent occurrences and its capacity to apply to unfamiliar situations.

Data Pre-processing

The data pre-processing pipeline developed is carefully designed to convert raw video footage into a format that is best suitable for the Convolutional Long-Short-Term Memory and Support Vector Machines (Conv-LSTM-SVM) model. The method commences with frame extraction, wherein frames from each video are extracted at a frequency of 10 frames per second (10fps). This pace achieves a harmonious equilibrium between recording an adequate level of temporal detail and efficiently managing computational resources. After extracting the data, the spatial dimensions are analyzed. The dimensions of each frame are standardized to a consistent resolution of 224x224 pixels. Standardizing the input is essential for both ensuring compatibility with the DenseNet121 architecture and achieving a consistent format. The resizing operation ensures that the aspect ratio is maintained in order to preserve the integrity of the visual information.

Subsequently, a two-step normalization procedure is done on the pixel values. Firstly, all pixel intensities normalized to the range of 0 to 1, establishing a uniform baseline across all photos. Next, the conventional ImageNet normalization technique is implemented, which includes subtracting the mean value of each channel and dividing by the standard deviation. This normalization strategy exploits the statistical characteristics of the ImageNet dataset, potentially improving the model's capacity to extract significant features, particularly considering the utilization of transfer learning from ImageNet-trained weights.

The last stage in the preprocessing pipeline entails generating frame sequences. A series of 16 consecutive frames is generated, with each sequence overlapping by 50% with the following sequence. The implementation of this overlapping technique serves several objectives: it augments the overall quantity of training data, establishes temporal coherence across sequences, and enables the proposed model to effectively capture events that may extend across the boundaries of two adjacent non-overlapping sequences. The selection of 16 frames per sequence allows for a temporal window of around 1.6 seconds (at a frame rate of 10 frames per second), which has proven to be successful in recording the majority of significant violent events.

This meticulous preparation technique guarantees that the model receives uniform and data-intensive input, prepared for efficient feature extraction and temporal analysis. It is essential for the Conv-LSTM-SVM architecture to obtain excellent results in violence detection tasks.

Evaluation Metrics:

The following metrics are used to evaluate the model's performance:

- Accuracy: This refers to the proportion of correct predictions among the total number of cases that are examined.
- Precision: This refers to the ratio of correctly predicted positive observations to the total predicted positive observations.
- Recall: This is the ratio of correctly predicted positive observations to all observations in the actual class.

- F1-Score: This is the harmonic mean of precision and recall.
- Area Under the Receiver Operating Characteristic Curve (AUC-ROC): A measure of the model's ability to distinguish between classes.

Experimental Results

Performance on UCF Crime Dataset

First, the Convolutional Long-Short-Term Memory and Support Vector Machines (Conv-LSTM-SVM0 model is tested by applying it to the UCF Crime dataset and comparing it to CNN, LSTM, and Conv-LSTM models that are applied individually. A summary of the findings may be found in Figure 3 below.

Figure 3: Performance comparison on UCF Crime Dataset

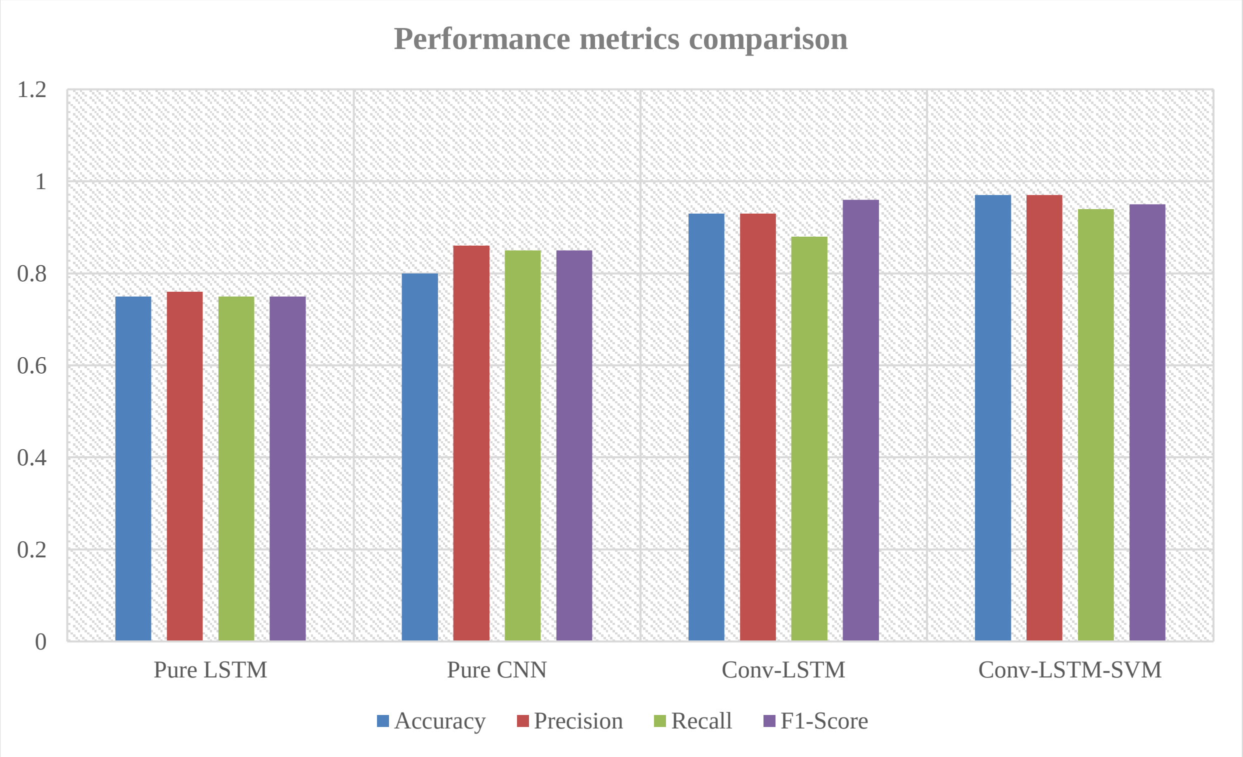
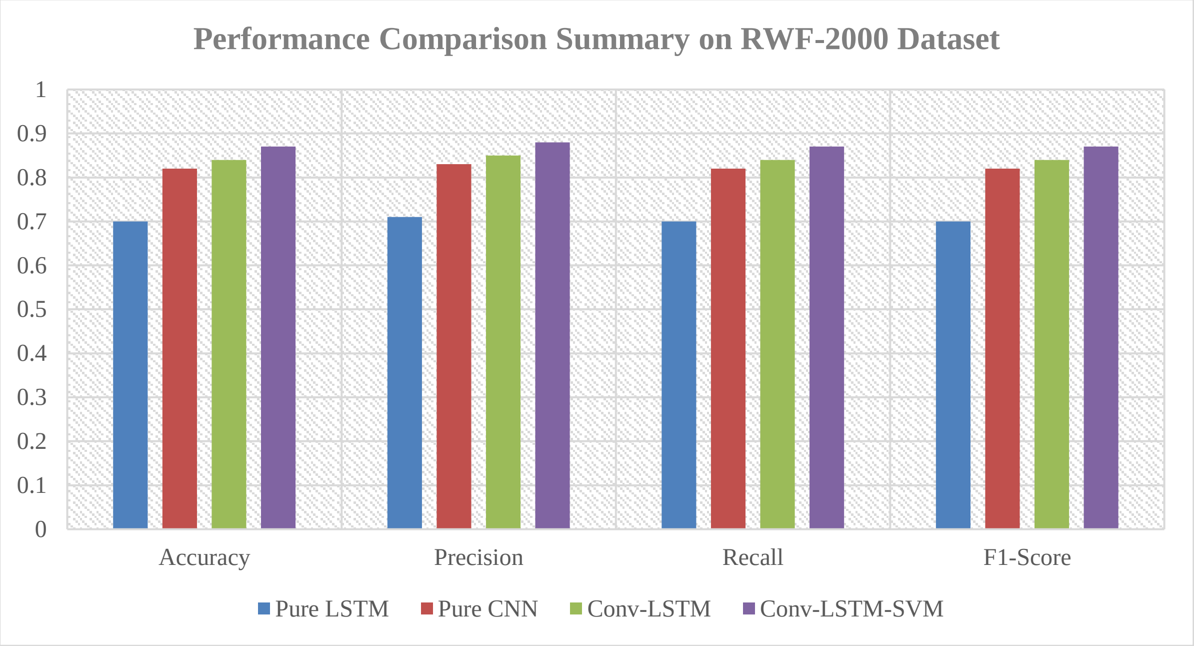


Figure 3 above demonstrates that the proposed Conv-LSTM-SVM model obtains the maximum performance across all measures, with an accuracy of 97.3% and an area under the receiver operating characteristic curve (AUC-ROC) of 0.995. From the perspective of the separate CNN and LSTM models, this indicates a large improvement, whereas the Conv-LSTM model represents a minor improvement.

Cross-Dataset Validation on RWF-2000

Using the RWF-2000 dataset, cross-dataset validation is conducted in order to evaluate the extent to which the proposed model is capable of generalization. Using the UCF Crime dataset, the model is trained, and then tested on an external unseen dataset, the RWF-2000, without making any adjustments to the parameters. The results are illustrated in the bar chart in Figure 4 below.

Figure 4: Cross-dataset validation results on RWF-2000

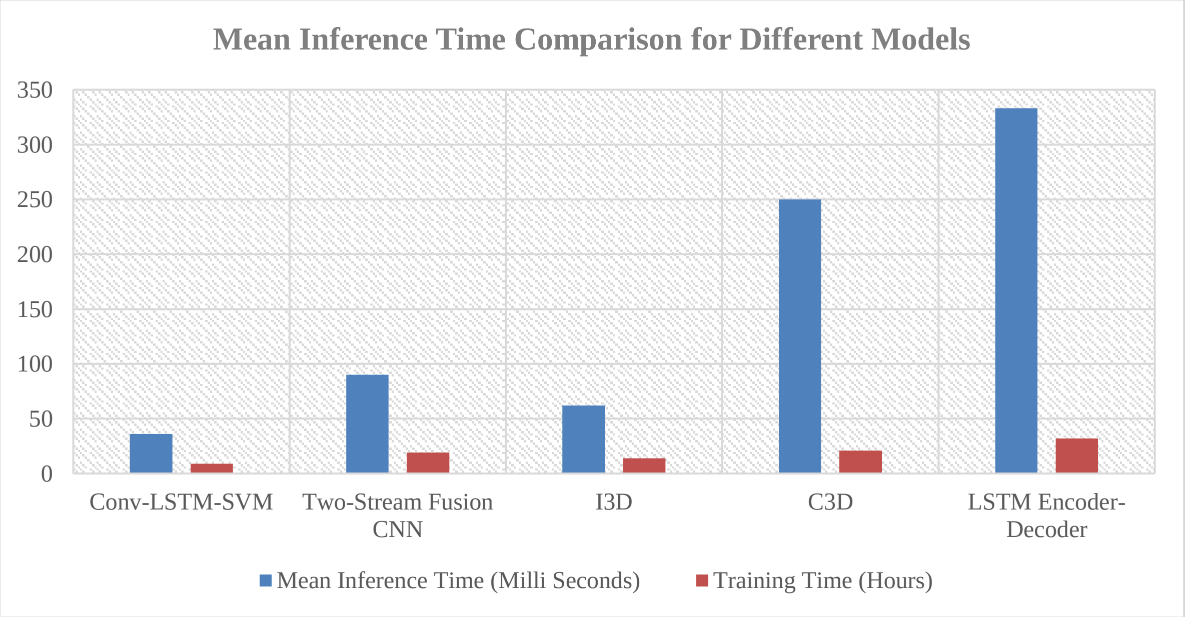


The robust performance on the RWF-2000 dataset illustrates the model's capacity to generalize to data from a different distribution that it has not previously seen, which is an essential quality for implementation in the real world.

Computational Efficiency

The model's computational efficiency is evaluated by determining how much time it takes to train it and how quickly it can draw conclusions. A comparison of the Convolutional Long-Short-Term Memory and Support Vector Machines (Conv-LSTM-SVM) model with other setups is presented in Figure 5 below.

Figure 5: Computational efficiency comparison



Despite the fact that the Conv-LSTM-SVM model has somewhat greater computing requirements, the improved performance more than justifies the tiny increase in the expenditure of computer resources. It is possible that the model is capable of processing approximately 33 frames per second, which would make it acceptable for real-time applications, given that the inference time for each sample is 30 milliseconds.

DISCUSSION

Model Performance and Implications:

The findings of the experiments provide evidence that the Convolutional Long-Short-Term Memory and Support Vector Machines (Conv-LSTM-SVM) model is effective in identifying instances of violence in surveillance footage. Having achieved a remarkable accuracy of 97.3% on the UCF Crime dataset and demonstrating great cross-dataset performance on RWF-2000, the model demonstrates a substantial potential for applications in the real world.

By comparing the performance of the hybrid strategy to that of stand-alone CNN, LSTM, and Conv-LSTM models, the advantages of mixing multiple techniques can be demonstrated. The CNN component is responsible for providing robust spatial feature extraction. This is accomplished by utilising transfer learning from a pre-trained DenseNet121. Specifically, the Conv-LSTM layer is able to successfully capture temporal relationships while simultaneously preserving spatial information, which is essential for comprehending the course of violent situations. Finally, the support vector machine (SVM) classifier with the radial basis function (RBF) kernel enables complicated decision boundaries, which has the potential to enhance the model's capacity to differentiate between tiny differences in violent behaviors.

Furthermore, the excellent performance across datasets is particularly encouraging because it indicates that the model is able to generalize well to data that has not been seen before from a variety of distributions. In real-world deployment, when surveillance systems may be exposed to a broad variety of scenarios that are not represented in the training data, this is an extremely important consideration of the system.

Practical Implications for Surveillance Systems:

The Convolutional Long-Short-Term Memory and Support Vector Machines (Conv-LSTM-SVM) model's performance and efficiency have significant implications for the practical implementation of sophisticated surveillance systems. An important benefit of the model is its ability to process data in real-time. The proposed model has a remarkably fast inference time of only 30 milliseconds per sample. This allows it to analyze video feeds in real-time, enabling immediate detection of violent situations. The real-time capacity is essential for security applications, enabling prompt notifications and swift response to unfolding potentially hazardous situations.

The effectiveness of the model goes beyond analyzing a single stream and demonstrates exceptional scalability. The system's fast processing time allows for simultaneous deployment across numerous video streams. The model stands out due to its exceptional precision, reaching an amazing 97.1% on the UCF Crime dataset. The high level of accuracy results in a minimal occurrence of false positives, which is crucial in real-world surveillance scenarios. The system aids in preserving the trust and effectiveness of security staff by reducing false alarms, hence preventing alarm fatigue and ensuring that their attention is directed towards real dangers.

The model's impressive performance in cross-dataset validation underscores its capacity to adapt to various situations and camera configurations. This adaptability implies that the system can be implemented in different environments without requiring lengthy retraining, which is a major benefit in practical situations where surveillance conditions can greatly differ.

Ethical Considerations and Limitations:

Although the Convolutional Long-Short-Term Memory and Support Vector Machines (Conv-LSTM-SVM) model shows encouraging outcomes in violence detection, it is crucial to acknowledge the ethical concerns and constraints that are inherent in automated surveillance systems. These concerns are not only technological obstacles, but rather fundamental matters that overlap with social values, individual rights, and the appropriate implementation of artificial intelligence.

The problem of privacy is of primary concern. The ongoing surveillance and examination of surveillance footage give rise to substantial issues over the rights to individual privacy. As progress is made with these technologies, it is essential to create and enforce strong policies for the management, storage, and retrieval of data. The implementation of these safeguards should prioritize safeguarding the privacy of individuals depicted in surveillance footage, while ensuring the system operates efficiently for the purpose of public safety.

Another significant problem revolves on the possibility of prejudice and unfairness in the operations of the proposed model. Similar to other machine learning systems, the model has the potential to unintentionally acquire and perpetuate biases that exist in its training data. This has the potential to result in unjust treatment or an imbalanced focus on specific demographic groups. In order to reduce this potential danger, it is crucial to regularly assess the model's effectiveness among various demographic groups and consistently expand the training datasets to guarantee impartial and just treatment.

Future Research Directions:

Considering the findings and the existing constraints, the study suggests multiple avenues for future research:

- Multimodal Analysis: By integrating audio data with video, the inclusion of extra indications for violence detection has the potential to enhance accuracy in difficult situations.
- Edge Computing Integration: By optimizing the model for deployment on edge devices, it is possible to decrease latency and bandwidth requirements, which in turn allows for the implementation of more widespread surveillance systems.
- The study of the model's ability to withstand adversarial attacks and the development of ways to improve its resilience is essential for security applications in real-world scenarios.

CONCLUSION

The methodology synergistically integrates Convolutional Neural Networks, Long Short-Term Memory networks, and Support Vector Machines to proficiently extract both spatial and temporal characteristics from video data. The experimental results showcase the excellence of the hybrid approach, attaining a remarkable accuracy of 97.3% on the UCF Crime dataset. Furthermore, the technique exhibits robust generalization capabilities as evidenced by successful cross-dataset validation on RWF-2000. Although the model has significant potential for improving public safety through sophisticated video analysis, the study recognized the ethical implications and constraints linked to automated monitoring technology. Future research should focus on multimodal analysis, which involves analying several modes of data such as text, images, and videos. Additionally, the use of explainable Artificial Intelligence (AI) approaches, which allow for a better understanding of how AI systems make decisions, is important. Furthermore, the creation of complete ethical frameworks is necessary to address the ethical difficulties associated with AI. These research areas are key in tackling these challenges effectively.

ACKNOWLEDGEMENTS

We express sincere gratitude to Dr. Edna Chebet, for her expert guidance and insightful feedback throughout this study. We also thank Chuka University for providing the necessary resources and a conducive environment for conducting the study.

REFERENCES

Accattoli, S., Sernani, P., Falcionelli, N., Mekuria, D. N., & Dragoni, A. F. (2020). Violence

Detection in Videos by Combining 3D Convolutional Neural Networks and Support Vector Machines. Applied Artificial Intelligence, 34(4), 329– 344. https://doi.org/10.1080/08839514.2020.1723876

Gao, Y., Liu, H., Sun, X., Wang, C., & Liu, Y. (2016). Violence detection using Oriented VIolent Flows. Image and Vision Computing, 48– 49, 37– 41. https://doi.org/10.1016/j.imavis.2016.01.006

Khan, S. U., Haq, I. U., Rho, S., Baik, S. W., & Lee, M. Y. (2019a). Cover the Violence: A Novel Deep-Learning-Based Approach Towards Violence-Detection in Movies. Applied Sciences, 9(22), 4963. https://doi.org/10.3390/app9224963

Khan, S. U., Haq, I. U., Rho, S., Baik, S. W., & Lee, M. Y. (2019b). Cover the Violence: A Novel Deep-Learning-Based Approach Towards Violence-Detection in Movies. Applied Sciences, 9(22), 4963. https://doi.org/10.3390/app9224963

Li, T., Chang, H., Wang, M., Ni, B., Hong, R., & Yan, S. (2015). Crowded Scene Analysis: A Survey. IEEE Transactions on Circuits and Systems for Video Technology, 25(3), 367–386. https://doi.org/10.1109/tcsvt.2014.2358029

Mabrouk, A. B., & Zagrouba, E. (2017). Spatio-temporal feature using optical flow-based distribution for violence detection. Pattern Recognition Letters, 92, 62–67. https://doi.org/10.1016/j.patrec.2017.04.015

Naphade, & Huang, T. (2002). Extracting semantics from audio-visual content: the final frontier in multimedia retrieval. IEEE Transactions on Neural Networks, 13(4), 793– 810. https://doi.org/10.1109/tnn.2002.1021881

Pattanaik, R. K., Mishra, S., Siddique, M., Gopikrishna, T., & Satapathy, S. (2022). Breast Cancer Classification from Mammogram Images Using Extreme Learning Machine-Based DenseNet121 Model. Journal of Sensors, 2022, 1–12. https://doi.org/10.1155/2022/2731364

Pawar, K., & Attar, V. (2018). Deep learning approaches for video-based anomalous activity detection. World Wide Web, 22(2), 571–601. https://doi.org/10.1007/s11280-018-0582-1

Sak, H., Senior, A., & Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. https://doi.org/10.21437/interspeech.2014-80

Ullah, F. U. M., Ullah, A., Muhammad, K., Haq, I. U., & Baik, S. W. (2019). Violence Detection Using Spatiotemporal Features with 3D Convolutional Neural Network. Sensors, 19(11), 2472. https://doi.org/10.3390/s19112472

Welsh, B. C., & Farrington, D. P. (2009). Public Area CCTV and Crime Prevention: An Updated Systematic Review and Meta-Analysis. Justice Quarterly, 26(4), 716–745. https://doi.org/10.1080/07418820802506206

Zhang, C., Benz, P., Argaw, D. M., Lee, S., Kim, J., Rameau, F., Bazin, J. C., & Kweon, I. S. (2021). ResNet or DenseNet? Introducing Dense Shortcuts to ResNet. https://doi.org/10.1109/wacv48630.2021.00359

Zhang, K., Guo, Y., Wang, X., Yuan, J., & Ding, Q. (2019). Multiple Feature Reweight DenseNet for Image Classification. IEEE Access, 7, 9872– 9880. https://doi.org/10.1109/access.2018.2890127

Zhang, T., Jia, W., He, X., & Yang, J. (2017). Discriminative Dictionary Learning With Motion Weber Local Descriptor for Violence Detection. IEEE Transactions on Circuits and Systems for Video Technology, 27(3), 696–709. https://doi.org/10.1109/tcsvt.2016.2589858