VFAST Transactions on Software Engineering
https://vfast.org/journals/index.php/VTSE@ 2025, ISSN(e): 2309-3978, ISSN(p): 2411-6246
Volume 13, Number 1, January-March 2025 pp: 60-71

Efficient and Sustainable Video Surveillance Using CNN-LSTM Model for Suspicious Activity Detection

Aasma Aas [1*], **Hamza Naveed** [2], **Junaid Asghar** [3], **Zubda Khanum** [4]
Sarah Khaleel [5], **Talia Noureen** [6]

[1*]Department of Computer Science, University of Management and Technology Lahore, Pakistan; [2]Department of Software Engineering, Faculty of Information Technology and Computer Science, University of Central Punjab, Lahore, Pakistan; [3] Department of Computer Science, University of Lahore, Pakistan; [4]Department of Computer Science, University of Management and Technology Lahore, Pakistan; [5]Department of Computer Science, Riphah International University Lahore, Pakistan; [6]Department of Computer Science, Beaconhouse National University Lahore, Pakistan

Keywords: *Real-time threat detection, video surveillance, deep learning, CNN-LSTM, ORB feature extraction, UCF crime dataset, automated security, intelligent surveillance, high-accuracy CCTV.*

Journal Info:
Submitted:
December 30, 2024
Accepted:
February 16, 2025
Published:
March 2, 2025

Abstract This study presents a novel approach for enhancing the automation and effectiveness of real-time threat detection in video surveillance systems. Traditional surveillance methods require continuous human monitoring, are resource-intensive, and often fail to consistently identify suspicious activities with precision. Addressing these challenges, we propose the Mono-Scale CNN-LSTM Fusion Network, an advanced deep-learning model designed for automated, sustainable, and high-accuracy CCTV systems. The model utilizes Convolutional Neural Networks (CNN) in combination with Long Short-Term Memory (LSTM) networks to improve recognition capabilities by capturing temporal and spatial features. For feature extraction, the Oriented FAST and Rotated BRIEF (ORB) techniques are employed to enhance detection efficiency. The model was tested using the UCF crime image dataset and achieved an accuracy rate of approximately 99%, surpassing traditional models like CNN, VGG-16, VGG-19, ResNet-50, and DenseNet. This study highlights the contributions of our approach, which offers a significant reduction in the need for human oversight and sets new standards in the field of automatic threat detection. Furthermore, it emphasizes the model's capability to support contemporary security systems with high precision, reliability, and scalability, making it a valuable tool for the next generation of intelligent surveillance systems.

***Correspondence author email address:** aasma.aas.muhammad@gmail.com
DOI: 10.21015/vtse.v13i1.2023

1 Introduction

According to the Global Organized Crime Index's Data Explorer (2023), countries are assigned criminality scores from 1 to 10, showing each country's level of organized crime. The index points out regional differences, with higher scores in parts of Africa and Asia, suggesting a stronger presence of organized crime in those areas. This data supports our study, which looks at the link between criminality scores and socio-economic factors, highlighting the need for specific interventions based on the region [1]. The criminality scores are further illustrated in Figure 1. Video surveillance has become an essential area of research in image processing, deep learning, and computer vision. It plays a vital role in maintaining security by capturing both normal and abnormal human activities. Real-time monitoring of suspicious activities in public spaces is crucial for preventing crime, improving security, and ensuring the safety of individuals in various environments [2]. Researchers are actively developing new algorithms and techniques to enhance the accuracy and efficiency of suspicious activity detection. These advancements are critical for improving real-time monitoring and response systems, which are fundamental for public safety. However, despite the progress made, there are still several challenges in current crime detection technologies. One main issue is the need for real-time processing, which is crucial for timely identification of suspicious activities. Many models still struggle to achieve high accuracy in real-time applications, where latency and quick responses are critical. Furthermore, the scalability of these models remains a concern, as they often perform well on small datasets but face difficulties when deployed in large-scale systems or with diverse video sources. Additionally, the accuracy of existing models is still a work in progress, as they may struggle with complex or overlapping activities, requiring continuous refinement [3].

In addition to the challenge of real-time processing, another significant issue is the variation in camera angles and video resolution across different surveillance systems. Many current models struggle to handle diverse camera perspectives and low-resolution footage, which can reduce the accuracy of behavior detection. The UCF-Crime dataset highlights these difficulties, as models often fail to detect suspicious activities in poorly lit or low-quality video frames, which are common in real-world surveillance scenarios. Moreover, scalability remains a persistent challenge, as crime section models may perform well on small datasets but struggle to adapt to larger, more varied data sources, leading to a decrease in overall system performance.

These gaps indicate the need for more robust, adaptable, and efficient crime detection systems capable of overcoming these challenges [4]. To address these challenges, the focus of this search is to enhance the crime detection process by improving the accuracy, scalability, and real-time- capabilities of surveillance systems. By exploring new methodologies and advanced models, such as the CNN- LSTM hybrid approach, this study aims to bridge the gaps in current crime detection technologies and contribute to more effective and efficient public safety systems.

Surveillance systems can be categorized into two types: traditional and autonomous. Traditional systems rely heavily on human monitoring, whereas au- tonomous systems use DL, AI, and ML algorithms to independently detect human activities. These advanced algorithms excel in feature extraction and pattern recognition, making them particularly adept at identifying complex activities. By training on large datasets, deep learning models can learn to recognize intricate human actions and adapt to unseen data, improving their ability to detect suspicious behavior across various environmental conditions. The integration of these technologies has significantly improved the performance and reliability of surveillance systems, enabling them to handle challenging situations such as occlusions, variable lighting, and crowded settings. However, ongoing refinement of these models is essential to maintain their performance in real-world applications. Furthermore, a Word Cloud of the research articles referenced is shown in Figure 2.

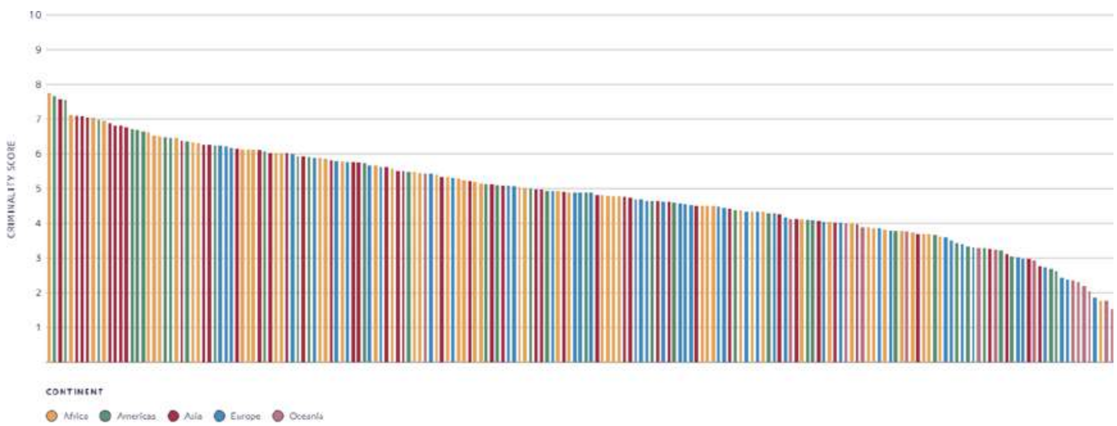

Figure 1. criminal data record

1.1 Research Questions (RQ)

The main contributions of this research paper are as follows:

1. How continuously monitoring with the human eye is a difficult task, however, the presented proposed model does not need any human because an alarm feature can be generated in the application.
2. How do Oriented FAST and Rotated BRIEF (ORB) play a role in the identification and characterization of suspicious image features?
3. How does the integration of Convolutional Neural Networks (CNN) with Long Short-Term Memory?
4. How the proposed deep learning framework is tested and validated on the UCF crime benchmark datasets, and performance is compared with CNN, VGG-16, ResNet-50, and DenseNet models, etc.

This research proposes a new method for crime activity detection in real-time surveillance videos using Mono-Scale CNN and LSTM networks for maximizing the extraction of spatial and temporal features. In contrast to the previous works which have focused on spatial characteristics only, or designed rather convoluted multi-scale structures, our model combines a mono-scale CNN-LSTM fusion, providing high accuracy while being computationally efficient. During the first step in feature extraction, we employ the ORB (Oriented FAST and Rotated BRIEF) feature descriptor since it improves the image processing rate while capturing the most crucial image features. This choice of ORB over more conventional features such as SIFT or HOG is done deliberately as it is a faster and more accurate method for real-time surveillance. Furthermore, This approach is a positive development toward achieving reliable real-time crime detection in surveillance systems hence improving public security.

Section 1 explains the drawbacks of conventional video surveillance systems that require operator attention, are costly and have poor real-time threat recognition capability. To address these issues, the proposed Mono-Scale CNN-LSTM Fusion Network integrates CNN for spatial feature extraction and LSTM networks for temporal analysis in Section 2 in addition to ORB (Oriented FAST and Rotated BRIEF) techniques for better detection performance. Using the UCF crime image dataset, the model had a 99% accuracy, which was higher than the traditional models such as CNN, VGG-16, VGG-19, ResNet-50, and DenseNet as shown in section 3 that confirmed the efficiency of the model in real-world scenarios. This approach reduces the need for human supervision which makes it scalable, sustainable, and highly reliable, thus providing a foundation for future developments of intelligent surveillance systems and automated security technology in Section 4.

Figure 2. Word Cloud of the research article

2 RELATED WORK

The researchers presented a literature review of 6 abnormal activities and discussed the steps involved in recognizing human activity, including foreground object extraction, feature extraction, and object detection by using non-tracking or tracking methods [5]. The study utilized the KTH and CAVIAR datasets, where features were computed from video frames, and the classifier made predictions [6]. In the study lighting and nighttime monitoring are crucial components and a CNN model using infrared (IR) images achieved 90% accuracy in object recognition [7]. In the context of education [8], detecting cheating activities in examination halls is critical. The SURF (Speeded Up Robust Features) algorithm was used to match and find corresponding features, and Viola-Jones object detectors were employed to detect faces and label activities. Focused on a group of people in a Warfield, implemented a real-time algorithm using a motion-based moving object detection method [9].

Another study on violence detection employed a Mono-scale CNN-LSTM model having 89% accuracy by collecting datasets from Google. The system takes the input of real-time videos from CCTV cameras and then passes them to a CNN model which is created by the transfer learning that detects objects and generates alerts [10]. A study in examination halls identified various features such as head, hands, and posture. By identifying these features, the system can detect whether a student is engaged in cheating [11]. Furthermore, the detection of firearms plays a crucial role in crime prevention.

A deep learning neural network model is introduced, focusing on the presence of weapons and individuals wearing helmets while carrying weapons [12]. The system provides enhanced security while keeping costs minimal. In another research the YOLO-V3 model was utilized for human activity detection, achieving an accuracy of 86.2% [13]. The Cov2d-LSTM and ConvLSTM2D models on the UCF Crime Dataset, achieved accuracy rates of 0.68 and 0.64, respectively [14]. An alternative approach for the detection of shoplifting incidents was introduced in [15]. They used 2D CNN, 3D CNN, and a fusion of Inception V3 with BILSTM models and got an accuracy of 49%, 57%, and 81%, respectively. A multi-stage process, encompassing the image preprocessing proposed model is to be effectively trained on training data, thus ensuring robust performance on previously unseen datasets [16]. Furthermore, a Hybrid Machine Learning Algorithm was applied to the UCF crime Dataset. They combined the potential of LSTM and CNN for feature extraction [17]. To enhance the model performance, temporal features were further extracted using a multilayer LSTM architecture, getting an im- impressive accuracy of 96%. An automated deep-learning architecture was introduced which is centered around 3D Conv Nets to extract spatiotemporal features. The experimental results showcased an accuracy rate of 82% [18].

Another contribution proposed an approach that achieves the accuracy of 3D-CNN for feature extraction, coupled with the BILSTM model for classification [19]. In research, CNN methodology achieves an accuracy of 90.2% on a dataset of 2000 images such as blood, guns, and knives [20]. Deep Neural Network was conducted, achieving an accuracy of 98%, on crowd violence, UCSD, and violent flow datasets [21]. A combination of CNN and BER models was implemented resulting in an accuracy of 85.63% by using the XD-Violence audio and video datasets [22]. Live CCTV data (CAVIAR dataset) is used to employ a CNN-GRU model for the detection of faces and weapons, achieving an impressive accuracy of 95.97% [23].

In another research, Twitter data was subjected to keyword filtering and labeling. The text mining methodology successfully categorized 10 distinct crime classes. By employing SVM and ANN on the dataset the researchers achieved an accuracy of 90.3% [24]. Another study found on the Saudi Arabia tweets (2017- 2021), where they identify the keywords that are used for criminal activities. They applied both ML and DL algorithms where they achieved an accuracy of 79% [25]. Chicago crime data is used to detect and map a crime-dense region in NYC [26]. The crime prediction research highlights unemployment and literacy rates. They employed a random forest regressor, yielding an accuracy of 97% [27]. In recent advancements, ML and DL models were used to forecast various crime types, and an impressive accuracy of 99% was achieved when applied to a weather dataset [28].

This research introduces a novel approach to real-time crime activity recognition in surveillance footage by combining Mono-scale CNN and LSTM networks, optimizing both spatial and temporal feature extraction. Unlike previous studies, which either focused on spatial features alone or employed complex multi-scale architectures, our model utilizes a mono-scale CNN- LSTM fusion that balances computational efficiency and high accuracy. We implement the ORB (Oriented FAST and Rotated BRIEF) feature descriptor for the initial feature extraction stage, which enhances image processing speed while capturing robust image features. This choice of ORB over more traditional feature Descriptors like SIFT or HOG reflects a deliberate focus on speed and reliability in real-time applications, essential for responsive surveillance systems. Therefore, represents a significant advancement in real-time, reliable crime detection within surveillance systems, contributing to improved public safety.

3 METHODOLOGY

The architecture is to pre-process the dataset using the Oriented FAST and Rotated BRIEF (ORB) technique which is robust, and faster for real-time applications, and then apply the Mono-scale CNN- LSTM model for recognition of images, which involves extracting meaningful features from the images using CNN and then leveraging LSTM models to effectively recognize the activities captured in the surveillance footage. The integration of CNN and LSTM models allows us to leverage the strengths of both architectures, enabling a more accurate and robust activity detection system as shown in Figure 3.

3.1 Dataset and Pre-Processing

For our study, a widely known UCF Crime Dataset is used on the 3 most common crime categories: Robbery, Shoplifting, and Fighting as shown in Figure 5. These images are divided into Train and Test subsets. The images within the dataset possess dimensions of 64x64 pixels and are stored in .png format. These images were extracted from video footage, with a systematic sampling approach where every 10th frame was chosen from each video. As part of the data preprocessing pipeline, first, convert the images to grayscale. By conversion, complexity is reduced while preserving intensity information.

In our study, we suggest using a tool called Oriented FAST and Rotated BRIEF (ORB) to work with images. ORB is like a detective for pictures. It spots special points in an image and figures out what's unique about them. This helps us describe the important features of an image simply. While other methods like SURF, SIFT, and HOG have been used in similar studies, we picked ORB because it's quicker and can find about 300 special points in an image %.

In the data preparation phase, the first step involves loading and processing image data using the OpenCV library and the ORB feature descriptor. The load-data function loads image data for different categories, with each category associated with a label. Subsequently, all images are resized to a consistent size of 50x50 pixels to facilitate efficient batch processing during model training. The ORB detector is then applied to extract features, with the score type initially set to ORB- HARRIS-SCORE, although this setting can be adjusted for experimentation. The detect And Compute method is used to obtain key points and descriptors for the resized images, where key points represent distinctive points in the image, and descriptors capture information around these key points. To visualize the extracted features, the detected keypoints are drawn on the image, resulting in what we call the "image-with-key points" as shown in Figure 4. This visualization illustrates where these distinctive points are located, providing valuable insight into how these key points help the computer recognize and interpret the images. These key points and their descriptors are essential clues that facilitate matching similar images, identifying objects within a picture, or tracking objects across video frames.

After data preparation steps, image data is appropriately structured and labeled for the successful training of the deep learning model. Regarding the LSTM element, the images are reconfigured, transforming each image into a sequential arrangement of values across the initial dimension while maintaining a singular value along the secondary dimension. This augmentation effectively introduces a third dimension, serving to note the distinct features of the data. The subsequent stage involves the division of the data using the train-test-split' function into testing and training subsets for both the CNN and LSTM constituents.

3.1.1 Long Short-Term Memory:

In the proposed architecture the LSTM model is created Keras-based LSTM layers in a linear stack to learn temporal dependencies from input data. It starts with two LSTM layers (8 units each), where the first retains sequences to capture temporal info from 2500 time steps with a single feature per step, using tanh activation. After LSTM, a dense layer (4 neurons) enhances complex relationship learning, followed by a 20% dropout layer for generalization. Finally, a flattened layer reshapes the 3D output from the preceding LSTM layers into a 1D vector, preparing the data for the classification task. The LSTM model, with its sequential data processing capabilities and the ability to capture temporal patterns, complements the CNN model effectively when combined with a given overall approach for detecting and classifying criminal activities in surveillance footage. The integration of both CNN and LSTM components contributes to the models' exceptional accuracy in identifying criminal activities, enhancing security measures, and ensuring public safety.

3.1.2 Convolution Neural Network:

The CNN model for activity detection processes 50x50 grayscale images through 3 successive convolutional layers (64, 128, and 256 filters, 3x3 filter size), each using Leaky ReLU.

Activation to capture essential features. Max-pooling layers (2x2) downsample feature maps, while dropout layers (25% and 40%) prevent overfitting. After convolution and pooling, the output is flattened into a 1D vector. A fully connected layer with 256 neurons, Leaky ReLU activation, and a 50% dropout fine-tunes features and adds regularization. The output layer uses softmax to classify criminal activities based on the highest probability, ensuring effective

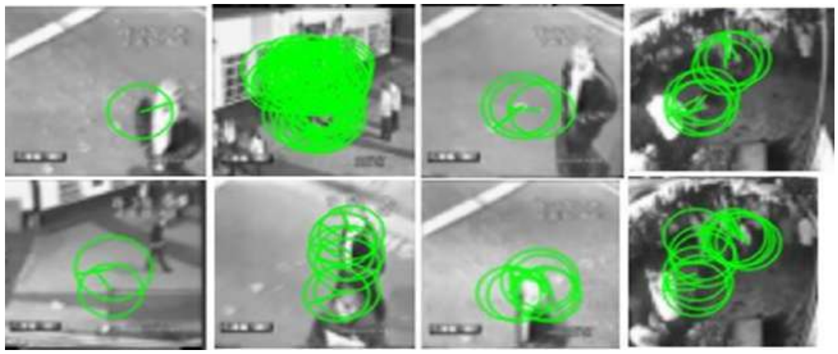

Figure 3. Visualization using ORB (Oriented FAST and Rotated BRIEF)

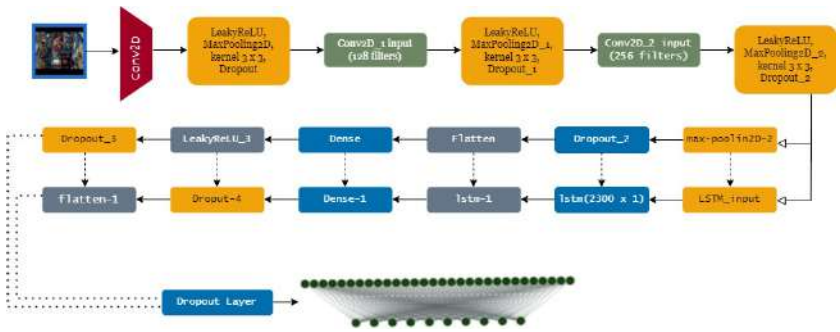

Figure 4. CNN-LSTM Fusion Network

feature extraction and classification.

3.2 A Proposed Mono-scale CNN-LSTM Model

presented approach has achieved a significant milestone by integrating the CNN and LSTM models. This allows the model to simultaneously process image data and sequential data, capitalizing on CNN's expertise in extracting visual features from grayscale images. The LSTM proficiency in capturing temporal patterns in sequences. For the assembly of the model, categorical cross-entropy is applied as the designated loss function:

$$L = \sum_{i=1}^{N} \max(0, 1 - y_i \hat{y}_i)$$

Where yi signifies the genuine class label and pi denotes the predicted probability for class i.

To drive optimization, the Adam optimizer is used, characterized by its dynamic learning rate, and to monitor the model's training 2 callbacks are applied. Batch sizes of 64 and 15 epochs are selected to optimize activity detection performance. The CNN model cases the input images, and the LSTM model handles sequential data, with its cell computations given by formulas described earlier. Validation is performed using test data to assess the models' generalization ability. Upon the conclusion of training, we rigorously evaluate the hybrid models' performance through the evaluation method, which computes the validation loss and accuracy of the test data. An impressive accuracy of 99% signifies the models' exceptional capability in accurately identifying and classifying criminal activities.

4 EXPERIMENTS RESULTS

Research closely monitored the learning progress of the integrated Mono-scale CNN-LSTM model across multiple epochs. The training accuracy steadily increased, reaching an impressive value of 0.99. An increase in accuracy indicates that the model achieved high accuracy in correctly classifying samples from the training dataset. There are numerous metrics are commonly employed to assess the performance of classifiers. Accuracy gauges the ability of a trained classifier to correctly predict class labels in comparison to the actual labels. The LSTM proficiency in capturing temporal patterns in sequences. For the assembly of the model, categorical cross-entropy is applied as the designated loss function:

$$A = \frac{tp + tn}{tp + tn + fp + fn} \quad (2)$$

$$P = \frac{tp}{tp + fp} \quad (3)$$

$$recall = \frac{tp}{fp + fn} \quad (4)$$

$$F1 - score = \frac{2 * precision * recall}{precision + recall} \quad (5)$$

The precision, recall, F1-score, and accuracy of the algorithm for the crime dataset are shown in Table 1. Similarly, the validation accuracy also showed improvement, reaching a value of 1.00. The validation accuracy is shown in Figure 4. The model performance on a separate validation dataset and an accuracy of 99% suggests that the model generalizes well to unseen data and can accurately classify samples from the validation dataset. Model performance was optimed throughout the training process to minimize the training loss. The training loss steadily decreased during training and reached a minimum value of 5 then training loss is calculated using a loss function, such as categorical cross-entropy. Similarly, the validation loss, which evaluates the model's performance on the validation dataset, also decreased during training and reached a minimum value of 10. A lower validation loss indicates that the model generalizes well and makes accurate predictions on unseen data. The confusion matrix before normalization is shown in Figure 8. Model predictions for each class ("Fighting," "Shoplifting," and "Robbery") were analyzed. The raw counts provided an initial insight into the model's performance across different activities.

After normalization, the confusion matrix is shown in Figure 9. transformed these counts into percentages, offering a more comprehensive view of the model's performance by accounting for variations in class sizes. The normalized confusion matrix revealed the following rates of correct classifications for each class: 100% for "Fighting," 100% for "Shoplifting," and 100% for "Robbery." These results demonstrate the model's high accuracy and proficiency in accurately classifying in- stances within each category.

The higher percentages on the diagonal of the matrix indicate strong performance in accurately identifying each activity, while the off-diagonal elements represent misclassifications. Despite these minor misclassifications. the model achieved commendable accuracy, with a 94% correct classification rate for "Shoplifting" and 87% for "Robbery." Overall, the high training and validation accuracy, the decreasing training and validation losses, and the normalized confusion matrix provide valuable insights into the model's performance.

These results demonstrate the model's ability to accurately detect and classify criminal activities in surveillance footage, underscoring its potential for real-world applications in enhancing security measures and ensuring public safety.

A comparative analysis in Table I shows that the proposed model achieves a significantly high accuracy rate of 99%, which notably surpasses previous results from other state-of-the-art models, including DenseNet121, Custom CNN, and VGG16. These results indicate that the Mono-scale CNN-LSTM model outperforms existing algorithms in terms of precision, recall, F1-score, and overall accuracy. Additionally, Table II provides a detailed comparison with other existing state-of-the-art approaches, further highlighting the proposed model's advancements over traditional deep learning methods. This comprehensive evaluation reinforces the proposed model's robustness and efficacy in crime detection, marking a significant step forward in autonomous surveillance system technology.

5 CONCLUSIONS AND FUTURE WORK

This research introduced a Mono-scale CNN-LSTM approach that achieves high accuracy in abnormal activity detection. Our primary objective is to enhance the accuracy of the presented model while also reducing its susceptibility to variations. The research approach involves processing image frames, extracting features, and then recognizing them using a combined Mono-

Figure 5. Dataset

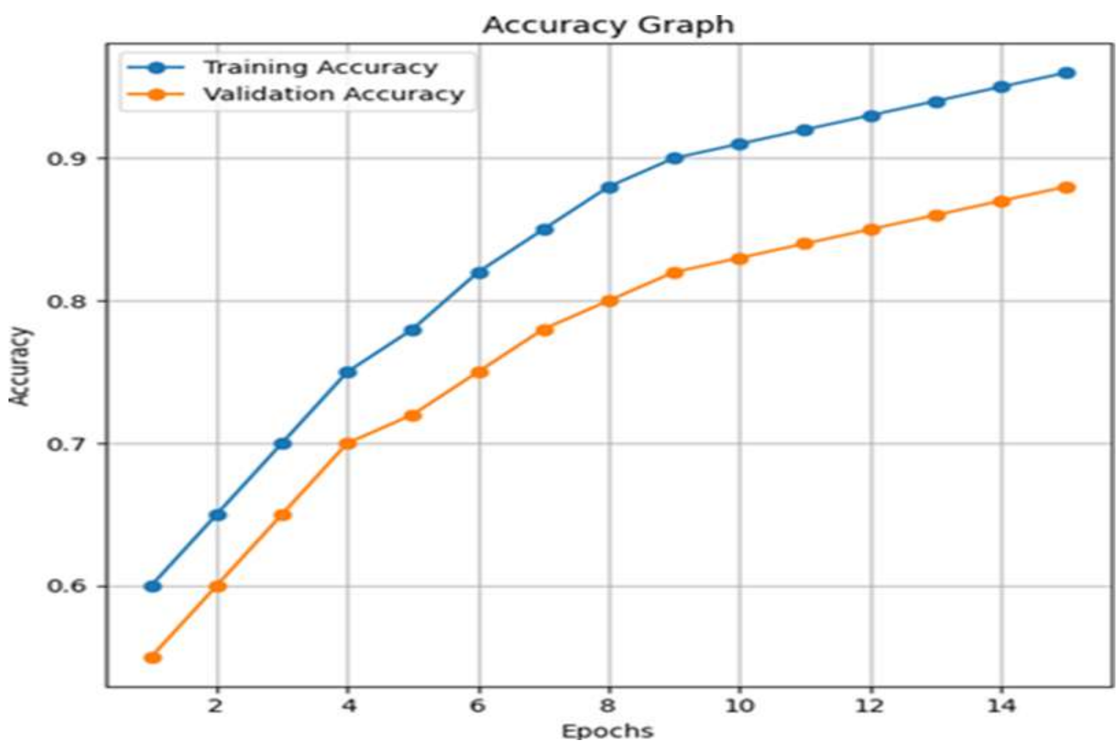

Figure 6. Accuracy

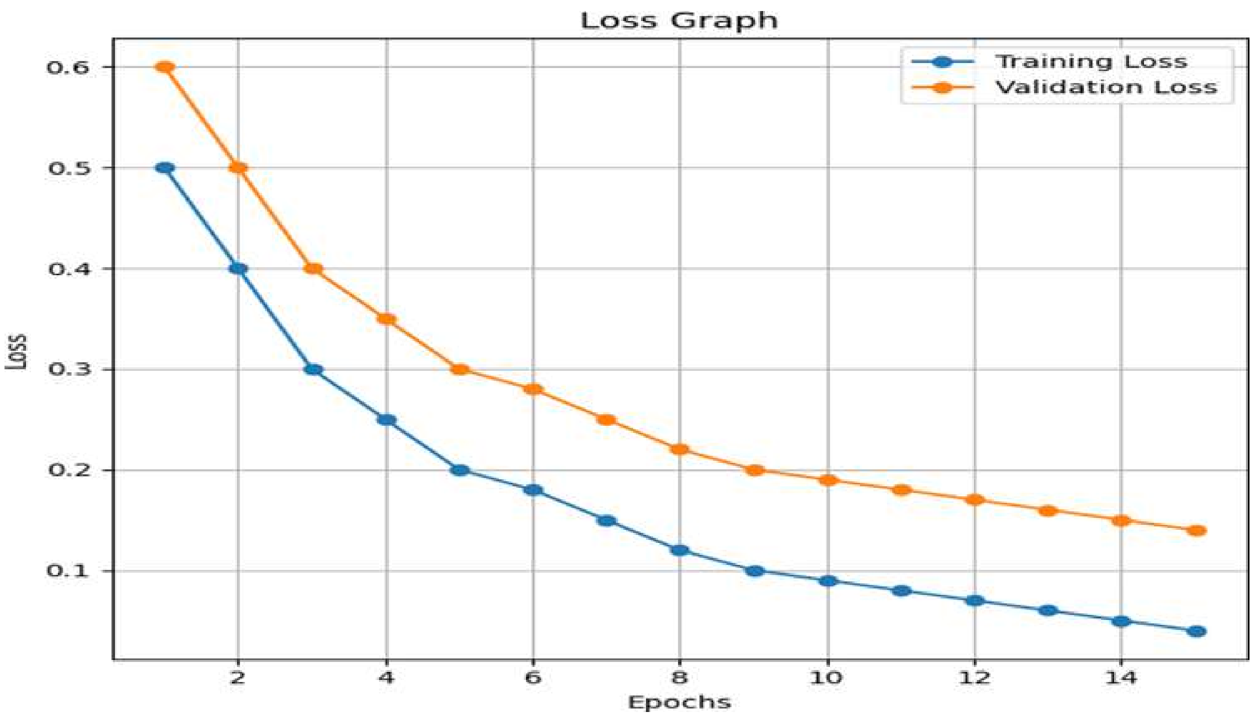

Figure 7. Loss

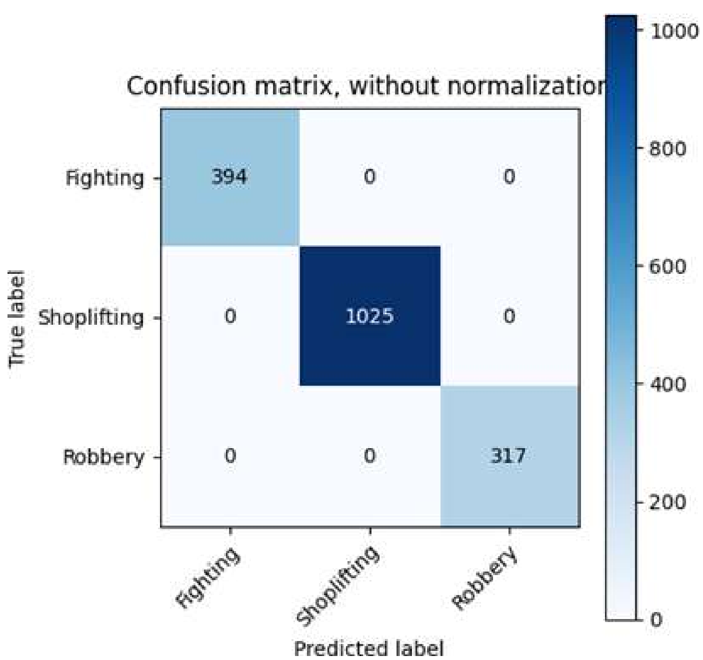

Figure 8. Confusion matrix before normalization

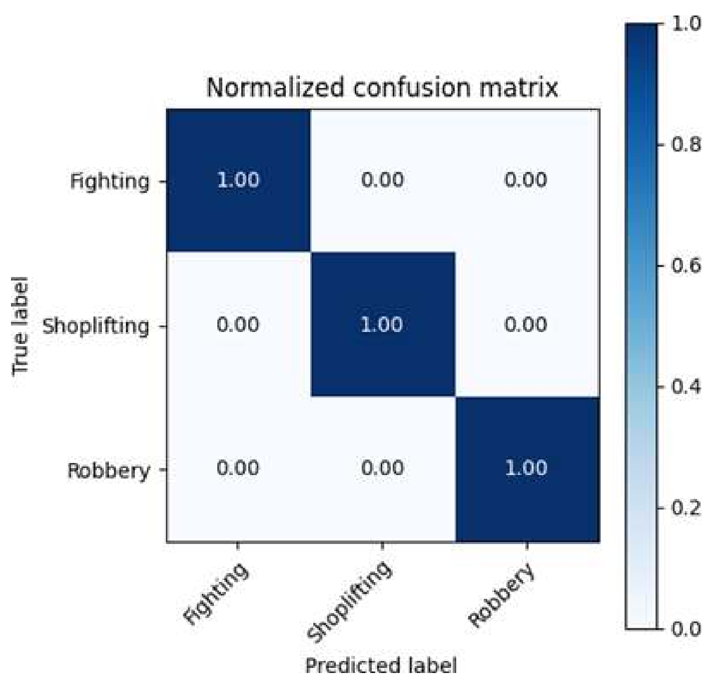

Figure 9. Confusion matrix before normalization

Table 1. Comparison of the proposed approach with other models in terms of Precision, Recall, F1-Score, and Accuracy

Model	Precision	Recall	F1-Score	Accuracy
Mono-scale CNN-LSTM	1.00	1.00	1.00	99%
Custom CNN	0.81	0.59	0.62	77%
DenseNet121	0.90	0.86	0.87	86%
VGG16	0.84	0.81	0.76	81%
VGG19	0.81	0.76	0.77	76%
DenseNet201	0.77	0.71	0.71	71%
ResNet50	0.34	0.59	0.43	59%
ResNet101V2	0.70	0.50	0.55	50%
ResNet50V2	0.70	0.50	0.55	50%
ResNet152	0.60	0.40	0.55	45%
DenseNet169	0.23	0.27	0.16	27%
ResNet152V2	0.64	0.25	0.13	25%
ResNet101	0.33	0.18	0.05	18%

Table 2. Comparison of existing state-of-the-art approaches

Author/Year	Method	Accuracy (%)
Sultani et al. (2018) [**?**]	C3D+Nearest Neighbor	23%
Sultani et al. (2018) [**?**]	TCNN	28.4%
Zhu et al. (2019) [**?**]	TCNN + Motion	31%
Maqsood et al. (2021) [**?**]	C3D Fine-tuning	45%

scale CNN-LSTM algorithm. The model has successfully achieved an accuracy score of 99%. In future endeavors, we are deeply committed to expanding the research by incorporating a broader range of datasets and experimenting with diverse models. Furthermore, we intend to meticulously assess the performance of our improved models across a spectrum of conditions, aiming to gain a comprehensive understanding of their capabilities. Through these concerted efforts, the aim is to continuously refine and elevate the effectiveness of architecture.

Author Contributions

Aasma Aas: Conceptualization, Supervision, Software, Writing- Reviewing **Hamza Naveed**: Writing-Original draft preparation. **Junaid Asghar**: Visualization, Methodology.**Zubda Khanum**:Editing**Talia Noureen**:Data curation.

Compliance with Ethical Standards

It is declare that all authors don't have any conflict of interest. It is also declare that this article does not contain any studies with human participants or animals performed by any of the authors. Furthermore, informed consent was obtained from all individual participants included in the study.

References

[1] U. M. Kamthe and C. G. Patil, "Suspicious activity recognition in video surveillance system," 2018.

[2] B. M. A. M. Buttar, A. M., et al., "Toward trustworthy human suspicious activity detection from surveillance videos using deep learning," 2023.

[3] A. A. S. Pouyan, M. Charmi, and H. Hassanpoor, "Propounding first artificial intelligence approach for predicting robbery behavior potential in an indoor security camera," *IEEE Access*, vol. 11, 2023.

[4] R. K. Tripathi, A. S. Jalal, and S. C. Agrawal, “Suspicious human activity recognition: a review.”

[5] C. V. Amrutha, C. Jyotsna, and J. Amudha, “Deep learning approach for suspicious activity detection from surveillance video,” 2020, pp. 335–339.

[6] N. T. J. and K. Thinakaran, “Detection of crime scene objects using deep learning techniques,” 2023, pp. 357–361.

[7] A. S. Ben-Musa, S. K. Singh, and P. Agrawal, “Suspicious human activity recognition for video surveillance system,” 2014, pp. 214–218.

[8] P. A. Dhulekar, S. T. Gandhe, N. Sawale, V. Shinde, and S. Khute, “Surveillance system for detection of suspicious human activities at war field,” 2018, pp. 357–360.

[9] G. Kaur and S. Sing, “Advances in information communication technology and computing,” 2022, pp. 1–6.

[10] O. M. Rajpurkar, S. S. Kamble, J. P. Nandagiri, and A. V. Nimkar, “Alert generation on the detection of suspicious activity using transfer learning,” 2020.

[11] A. Pawade, S. R. Anjaria, and R., “Suspicious activity detection for security cameras,” 2021, pp. 211–217.

[12] A. Shamnath and M. Belwal, “Human suspicious activity detection using ensemble machine learning techniques,” 2022, pp. 1–5.

[13] W. Ullah, H. Ullah, A. T. Khan, Z. A. Khan, and S. W. Baik, “An efficient anomaly recognition framework using an attention residual LSTM in surveillance videos,” 2021.

[14] A. Dua, B. Kalra, A. Bhatia, D. A. Madan, M. Gigras, and Y., “Crime alert through smart surveillance using deep learning techniques,” 2022, pp. 1–8.

[15] I. Muneer, M. Saddique, Z. Habib, and H. G. Mohamed, “Shoplifting detection using hybrid neural network CNN-built and development of benchmark dataset,” *Applied Sciences*, 2023.

[16] A. Babiyola, S. Aruna, B. S. Sumithra, and B., “A hybrid learning framework for recognition of abnormal events intended from surveillance videos,” pp. 1–14.

[17] R. Maqsood, U. I. Bajwa, G. Saleem, et al., “Anomaly recognition from surveillance videos using 3D convolution neural network,” 2021.

[18] M. L. S. Kokila, V. B. Christopher, and R. I. Sajan, “Efficient abnormality detection using patch-based 3D convolution with recurrent model,” 2023.

[19] M. Nakib, R. T. Khan, S. Hasan, and J. Uddin, “Crime scene prediction by detecting threatening objects using convolutional neural network,” 2018, pp. 1–4.

[20] K. B. Sahay, B. Balachander, B. Jagadeesh, G. A. Kumar, R. Kumar, and L. R. Parvathy, “A real-time crime scene intelligent video surveillance system in violence detection framework using deep learning techniques,” 2022.

[21] M. Boukabous and M. Azizi, “Multimodal sentiment analysis using audio and text for crime detection,” 2022, pp. 1–5.

[22] R. Shenoy, D. Yadav, H. Lakhotiya, and J. Sisodia, “An intelligent framework for crime prediction using behavioral tracking and motion analysis,” 2022, pp. 1–6.

[23] M. A. Permana, M. I. Thohir, T. Santoro, and M. A. Ayu, “Crime rate detection based on text mining on social media using logistic regression algorithm,” 2021, pp. 1–6.

[24] A. Algefes, N. Aldossari, F. Masmoudi, and E. Kariri, “A text-mining approach for crime tweets in Saudi Arabia: From analysis to prediction,” 2022, pp. 109–114.

[25] C. Catlett, E. Cesario, D. Talia, and A. Vinci, “Spatio-temporal crime predictions in smart cities: A data-driven approach and experiments,” vol. 53, 2019, pp. 62–74.

[26] L. G. A. Alves, H. V. Ribeiro, and F. A. Rodrigues, “Crime prediction through urban metrics and statistical learning,” vol. 505, 2018, pp. 435–443.

[27] V. M. L. Elluri and N. Roy, “Developing machine learning-based predictive models for smart policing,” 2019, pp. 198–204.

[28] K. K. E. Rublee, V. Rabaud, and G. Bradski, “ORB: An efficient alternative to SIFT or SURF,” 2011.