Towards Intelligent Retail Security: ConvLSTM-Based Shoplifting Detection with Adam Optimization

Kiran Narang[1], Gargi Singh[2], Laxman Singh[3], Deepti Gupta[4], Rekha Kashyap[3], Shashi Tanwar[5], Chitvan Gupta[6]

[1]Dept. of Computer Science and Engineering (AI), KIET Group of Institutions, Ghaziabad, U.P, India,
[2]Dept. of Sciences, Christ University, Delhi NCR, India
[3]Dept. of Computer Science and Engineering (AI-ML), KIET Group of Institutions, Ghaziabad, U.P, India
[4]Department of CSBS, Noida Institute of Engineering & Technology, Greater Noida, U.P., India
[5]Department of CSE, Anangpuria School of Management & Technology, Faridabad (Haryana), India
[6]Dept. of Mathematics & Computing, Noida Institute of Engineering & Technology, Gr. Noida, U.P., India

Abstract—Shoplifting is a major concern for retail businesses, causing significant financial losses worldwide. This study presents a real-time detection framework that integrates Convolutional Long Short-Term Memory (ConvLSTM) networks with the Adam optimizer to enhance surveillance-based anomaly recognition. The framework used deep learning for extracting spatial features from video frames with LSTM units that capture temporal dependencies, allowing the system to identify subtle and complex patterns of suspicious behavior. The Adam optimizer is employed to adaptively adjust learning rates, ensuring faster convergence and stable model training.Experimental evaluation on the UCF-Crime dataset (shoplifting category) and a custom retail surveillance dataset demonstrates that the proposed method achieved an accuracy of 92%, outperforming other traditional state of art approaches. This research study also addressed key challenges such as class imbalance, environmental noise, and variability in recording conditions that helped to enhance the overall results of anomaly recognition. Hence, this research overall contributes towards building intelligent, scalable, and automated retail security systems capable of reducing theft-related losses.

Key Words: Shoplifting Detection, Video surveillance, Deep Learning, Anomaly Detection, ConvLSTM, Retail Security, Real- Time Monitoring, Adam Optimizer.

1. Introduction

Each year, businesses around the world suffer significant financial losses due to retail theft, especially shoplifting. Reports indicate that annual losses from such incidents amount to billions, directly impacting profitability and operational efficiency. Conventional monitoring is highly dependent on human intervention, which introduces limitations such as fatigue and decreased attention. Moreover, manual surveillance can lead to missed incidents, delayed responses and reduced overall security. To mitigate these challenges, retailers have invested in extensive security measures, such as surveillance camera systems [1] [2]. Traditional video surveillance also has many limitations in preventing shoplifting. Shoplifting can occur in blind places or locations that are not captured by cameras. Security professionals may not be continually monitoring the feeds, and they may overlook suspicious activities. Traditional systems frequently require a manual examination of footage to detect incidents, which can be time-consuming. Motion detection may cause false alerts, wasting security resources. These limitations underscore the need for more sophisticated and intelligent monitoring technologies that can successfully prevent and detect shoplifting.

Modern surveillance technology is transforming the way of security, especially in retail settings. With AI-powered analytic, these systems can detect suspicious activity and warn personnel immediately, allowing for quick action to avoid stealing. Advanced motion detection capabilities can generate alarms for unusual activity, reducing false alerts and allowing security personnel to focus on serious threats. Furthermore, facial recognition technology can recognize and alert security to known shoplifters, allowing for proactive loss prevention measures. Using these cutting-edge capabilities, retailers may drastically improve their security posture and secure their assets more effectively [3] [4].

In this research paper, automation of shoplifting detection using the hybrid deep learning model based on Convolutional Long Short- Term memory network is proposed. The ConvLSTM architecture is employed to process frame sequences from surveillance videos to detect patterns that indicate shoplifting, such as hidden goods or suspicious movements. By learning regular shopping behaviour patterns, LSTMs can identify aberrant acts that may imply stealing [5].

For optimization of the proposed model, Adam optimizer is used that dynamically adjust the learning rates of the model using first and second moment gradient estimates. This improves convergence speed and stability of the model, making it suitable for real time application [5], [6].

The structure of this paper is as follows: Section II reviews the related work on shoplifting detection and video-based anomaly analysis. Section III discusses the detailed purposed methodology, with ConvLSTM architecture and optimization strategy. Section IV outlines the data set, experimental setup, and the evaluation metrics to check the performance of the model. Section V presents the results and analysis of the framework. Lastly Section VI provides a summary of future research directions.

2. Literature Review

The detection of anomalies in video surveillance, for shoplifting situations, is strongly dependent on studying past methodologies in order to enhance accuracy and efficacy. Various deep learning models, including CNN, RNN, and hybrid architectures, have been used to improve detection accuracy and effectiveness. This section provides an overview of important research works that have contributed to the field of detecting unusual activities.

Deep learning models used for video surveillance improved anomaly detection over early methods that rely on hand-designed features and statistical models. Traditional methods create many problems in dynamic real-world situations which can also be solved using neural networks to extract spatial and temporal patterns and hence improving classification accuracy [7]-[9]. Spatiotemporal analysis is very important for recognizing anomalies in surveillance videos. The Incremental Spatiotemporal Learner (ISTL) model, enhances real-time detection by using learning strategies. In recent deep learning development, CNN-based architectures have been found efficient in handling real-time video data for anomaly classification [10], [11].

Several deep learning architectures have been developed and designed to enhance the anomaly detection performance of the model. Hybrid models, such as CNN-BiLSTM, effectively capture spatial and temporal dependencies, which leads to good classification accuracy. Attention-based mechanisms optimize these models by focusing on critical regions within a video frame. Influence-aware attention networks refine detection in complex and more crowded surveillance scenarios by using location-based and motion- based weighting techniques [4]– [6].

Extracting meaningful features is a very important step in anomaly detection, because it helps in eliminating irrelevant information while containing essential patterns. Techniques such as deep autoencoders and CNN-based models have been widely used for feature extraction in video analysis. Hybrid models, such as CNN-BiLSTM architectures state effective real-time classification of anomalies by safeguarding both spatial and temporal dependencies. Convolutional autoencoders have been used to learn latent representations of normal and abnormal activities, in improving the accuracy of anomaly detection frameworks [1], [8].

Accurately recognizing human actions plays a very important role in anomaly detection for surveillance applications. Several studies have been done on the deep learning-based methodology, such as hybrid deep evolving neural networks that combine ConvLSTM and Long-term Recurring Convolutional Networks (LRCNs), to enhance the classification of human activities. Two-stream convolutional networks, which process spatial and motion information simultaneously, have been widely used to improve anomaly detection accuracy in security-sensitive environments, which is of utmost importance [2] [3].

Table 1 gives a summary of existing research contributions, their datasets, important evaluation parameters, and methodologies used.

Table 1: Summary of Various State of Art methods for Video Anomaly Detection

Authors	Research Contributions	Dataset Used	Performance Metrics
Duong et al. [12] (2023)	Surveyed deep learning methods for video anomaly detection	Multiple public datasets	Accuracy, Precision, Re- call
Nawaratne et al. [13] (2020)	Proposed ISTL framework for real time surveillance	CUHK Avenue Dataset	Detection Rate, False Alarm Rate
Nithesh et al. [14] (2022)	Developed CNN-based deep learning model for anomaly detection	UCF-Crime Dataset	Accuracy, F1-score
Jain et al. [15] (2024)	Introduced CNN model for improved spatiotemporal anomaly recognition	ShanghaiTech Dataset	Precision, Recall, Accu- racy
Muneer et al. [16] (2023)	Created a benchmark dataset and CNN-BiLSTM model	Custom Shoplifting Dataset	Accuracy, Sensitivity, Specificity
Chang et al. [3] (2022)	Designed a contrastive attention module for weakly supervised learning	UCSD Pedestrian Dataset	Precision, ROC-AUC Score
Zhou et al. [8] (2020)	Introduced an attention-driven loss function for improved generalization	UMN Dataset	Detection Accuracy, F1- score
Zhang et al. [9] (2022)	Developed influence-aware attention model for surveillance	Subway Entrance Dataset	Precision, Recall, False Alarm Rate

Nasaruddin et al. [10] (2020)	Applied spatiotemporal attention mechanism with 3D CNN	VIRAT Dataset	Accuracy, Computational Efficiency
Ullah et al. [11] (2021)	Combined CNN and BiLSTM for real-time video anomaly detection	XD-Violence Dataset	F1-score, Sensitivity
Ribeiro et al. [1] (2020)	Designed convolutional autoencoder for spatiotemporal anomaly detection	UCSD Ped2 Dataset	Precision, Detection Rate
Dasari et al. [2] (2022)	Used ConvLSTM and LRCN for human activity recognition	Kinetics-600 Dataset	Accuracy, Recall, Con- fusion Matrix

3. Research Background

A. ConvLSTM Architecture

The proposed model utilizes a *Convolutional Long Short- Term Memory (ConvLSTM)* network, designed to process spatiotemporal data effectively. Unlike standard LSTM models, ConvLSTM integrates convolutional operations within its gates, preserving the spatial structure of input frames. This architecture is particularly beneficial for video-based anomaly detection, where both *appearance (spatial features)* and *movement (temporal dependencies)* play a crucial role. Figure 1 shows layered architecture of ConvLSTM model.

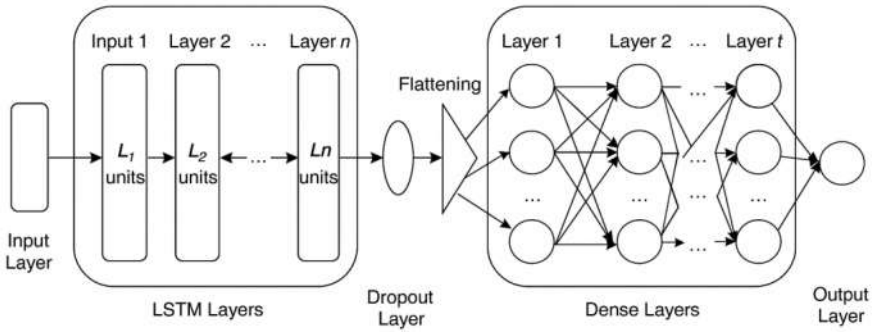

Fig. 1: Block diagram of the LSTM-based architecture.

Following are the key components of the ConvLSTM Architecture:

Input Gate:

It determines how much of the new input should be added to the cell state. Convolution operation of input gate using input X_t and hidden state H_t is given by

$$i_t = \sigma[W_{xf} * X_t + W_{hi} * H_{t-1} + b_i] \quad (1)$$

Forget Gate:

It determines "ft" whatever information is to be forget from previous gates.

$$f_t = \sigma[W_{xf} * X_t + W_{hf} * H_{t-1} + b_f] \quad (2)$$

Cell state:

It Stores long-term memory over time which is updated with the input gate, forget gate, and cell input.

$$C_t = f_t \odot C_{t-1} + I_t \odot \tanh[W_{xc} * X_t + W_{hc} * H_{t-1} + b_c] \quad (3)$$

Output Gate:

It Determines how much of the cell state will be disclosed to the output.

$$O_t = \sigma[W_{xo} * X_t + W_{ho} * H_{t-1} + b_o] \quad (4)$$

Hidden state

It determines hhe final output of the ConvLSTM cell at time t which combines the output gate and the cell state.

$$y = \sigma[W_x + b] \quad (5)$$

where Xt represents the input at time t, Ht is the hidden state, Ct is the cell state, ⊙represents the Hadamard product, and σ is the activation function of the sigmoid. Equations 1,2,3,4 and 5 shows the formula for calculating Input gate, Forget gate, Cell state, Output gate, and hidden State respectively [12]-[15].

In a ConvLSTM-based shoplifting detection architecture, layers such as ConvLSTM2D Layers, MaxPooling3D Layers, and Dense play particular roles in spatial, temporal, and decision-making activities.

ConvLSTM2D Layers: At the core of the model are Con- vLSTM2D layers, responsible for extracting spatio-temporal patterns from video sequences. Each layer processes a sequence of 2D feature maps, ensuring simultaneous spatial and temporal feature

learning. It identifies the Movement/ Features such as Person's posture, Suspicious hand movement, Object (e.g., item being taken) [16].

MaxPooling3D Layers: MaxPooling3D is a down sampling layer that selects the maximum value within a 3D sliding window to reduce the size of 3D feature maps (usually derived from videos or volumetric data). It works with five-dimensional form tensors given by (B, T, H, W, C). Here B indicates batch size, T indicates temporal dimension (number of frames), H, W indicates height and width and C indicates channels. MaxPooling3D layer with Pool_size= (p_t, p_h, p_w) and Stride= (s_t, s_h, s_w).

$$Y[t, i, j, c] = \max\{X[t', i', j', c]\}$$
$$t' \in [t * s_t, t * s_t + p_{t-1}]$$
$$i' \in [i * s_h, i * s_h + p_{h-1}]$$
$$j' \in [j * s_w, j * s_w + p_{w-1}]$$

Here Y is the output tensor, p_t, p_h, p_w be the pooling window size (time, height, width), s_t, s_h, s_w be the strides, c be the channel index (not pooled).

Dense Layer: After feature extraction and pooling, the dense layer performs the final classification. The extracted features are flattened and passed through a fully connected layer with a sigmoid activation function, which determines the probability that a particular instance is a shoplifting or not:

$$y = \sigma[Wx + b]$$

where x is the input feature vector, W is the weight matrix, b is the bias, and σ represents the activation function.

Optimization with Adam

In this work, the ConvLSTM network is optimized using the Adam optimizer, a widely used method in deep learning due to its robustness and efficiency. Adam integrates the benefits of AdaGrad, which adapts learning rates for each parameter, and RMSProp, which adjusts learning rates based on recent gradient magnitudes. Unlike traditional stochastic gradient descent, Adam computes individual adaptive learning rates for different parameters by combining estimates of both the first-order moment (mean of gradients) and the second-order moment (variance of gradients). This mechanism makes it highly effective for complex spatiotemporal models such as ConvLSTMs, where gradient magnitudes may vary across layers. With default hyperparameters (β1=0.9, β2=0.999, ϵ=10−7) and an initial learning rate of 3×10−4, Adam provides fast convergence, numerical stability, and reduced sensitivity to manual tuning. Consequently, the optimizer enhances the ability of the proposed model to detect shoplifting activities by efficiently capturing motion dynamics in video data while mitigating challenges like vanishing and exploding gradients [17]-[19].

Mathematical Formulation of Adam

Adam optimizer combines the benefits of momentum-based updates and adaptive learning rates. The optimization process follows these equations:

First moment estimate (mean):

$$m_t = [\beta_1 * m_{t-1} + (1 - \beta_1) * g_t] \quad (6)$$

Second moment estimate (variance):

$$v_t = [\beta_2 * v_{t-1} + (1 - \beta_2) * g2_t] \quad (7)$$

Bias-corrected first moment estimate:

$$\widehat{m_t} = \left[{m_t}/{(1 - \beta_1)_t}\right] \quad (8)$$

Bias-corrected second moment estimate:

$$\widehat{v_t} = \left[{v_t}/{(1 - \beta_2)_t}\right] \quad (9)$$

Parameter update:

$$\theta_t = \left[\theta_{t-1} - \left(\alpha * {\widehat{m_t}}/{\left(\varepsilon + \sqrt{v_t}\right)}\right)\right] \quad (10)$$

Here Equation (6) defines the first moment estimate, where an exponential moving average of the gradients is calculated to capture momentum. Equation (7) shows the second moment estimate, which tracks the average of squared gradients to measure their magnitude. Since both of these estimates start from zero, Equation (8) and Equation (9) apply bias correction to obtain more accurate values of the first and second moments during the initial iterations. Finally, Equation (10) gives the parameter update rule, where the corrected first moment is normalized by the square root of the corrected second moment, scaled by the learning rate, and adjusted with a small constant to avoid division by zero. Together, these equations describe how Adam adapts the learning rate for each parameter, making training more stable and efficient. Additionally, m_t and v_t are the first and second moment estimates, g_t is the gradient at time step t. β_1 and β_2 are the exponential decay rates for the moment estimates. α is the learning rate, ε is a small constant for numerical stability.

By integrating momentum and adaptive learning rates, Adam achieves faster convergence compared to classical optimizers, making it well-suited for training the proposed ConvLSTM-based shoplifting detection framework.

Dataset Used

For our experiments, we utilize the UCF-Crime dataset, specifically focusing on the shoplifting category, combined with a custom-collected retail surveillance dataset. The surveillance dataset was carefully annotated into two categories: shoplifting and normal behaviour clips. To ensure balanced analysis, all videos were pre-processed into fixed-length segments, resized to 128×128 pixels, and normalized. This hybrid dataset setup not only provides diversity in shoplifting scenarios but also enhances the robustness of the model by including real-world variations such as lighting changes, crowd density, and camera angles. Finally, the dataset was split into the ratio of 70%, 15%, and 15% for training, validation, and testing, ensuring that no overlap of video clips occurred across splits.

4. Proposed Method

The proposed framework for shoplifting detection employs Convolutional LSTM (ConvLSTM) networks, optimized with Adam, to jointly capture spatial features from video frames and temporal patterns of human actions. Figure 2 illustrates the overall structure of the ConvLSTM- based shoplifting detection model, demonstrating the flow of information through its various components. The system is organized into the following stages:

Data Acquisition and Preprocessing

Surveillance videos are divided into overlapping clips of 16–32 frames using a sliding window. Each frame is resized to 128×128 pixels and normalized. To improve generalization, data augmentation techniques such as horizontal flipping, brightness adjustment, and random cropping are applied. Since shoplifting events are much fewer than normal activities, oversampling and focal loss are used to balance the dataset and handle class imbalance effectively.

Spatial Feature Extraction

Each frame is passed through a Time Distributed CNN encoder composed of Conv2D and MaxPooling layers. This encoder extracts key spatial features such as shelves, objects, and hand gestures, while preserving positional information. The encoder produces a sequence of spatial feature maps for every frame.

Temporal Modeling with ConvLSTM

The extracted feature maps are processed through ConvLSTM2D layers, which simultaneously model spatial locality and temporal dependencies. This enables the detection of behavioural patterns such as repetitive reaching, concealment of items, or unusual object interactions. The ConvLSTM output represents a spatiotemporal embedding of the video clip.

Classification Layer

The ConvLSTM output is passed through a MaxPooling3D layer to reduce dimensionality and highlight the most informative spatiotemporal features. A dropout layer is applied to avoid overfitting. Finally, a fully connected Dense layer with sigmoid activation provides the probability of shoplifting versus normal activity.

Optimization with Adam

Training is conducted using the Adam optimizer with a learning rate of 3×10-4, $\beta 1=0.9$, $\beta 2=0.999$, $\epsilon=10-7$. Adam is selected because it adaptively adjusts the learning rate for each parameter, ensuring stable convergence and faster optimization in complex ConvLSTM models.

Inference Strategy

During real-time operation, the video stream is divided into successive clips, each processed through the model. To minimize false positives, a temporal smoothing mechanism (such as majority voting or probability thresholding across consecutive clips) is used. An alert is triggered only when the predicted probability of shoplifting remains consistently above a threshold (e.g., 0.8) for multiple clips.

Figure 2 shows block diagram of the complete workflow of the proposed shoplifting detection system. The process begins with the collection of video data, which is carefully preprocessed by resizing each frame to a fixed resolution, applying augmentation techniques to improve data diversity, and normalizing pixel values for uniformity. These prepared frames are first analyzed by a Convolutional Neural Network (CNN), which extracts important spatial characteristics such as shapes, textures, and object details. The spatial features are then passed to a Convolutional Long Short-

Term Memory (ConvLSTM) network, which learns the sequence of movements and temporal patterns present across consecutive frames. To refine the output, a global average pooling layer condenses the feature maps, while a dropout layer is introduced to avoid overfitting. The final stage involves a fully connected dense layer that classifies the activity into two categories: shoplifting or normal behavior. During training, the Adam optimizer plays a crucial role by adjusting the model parameters

based on the computed loss, ensuring faster convergence and stable learning throughout the network. This structured flow enables the system to effectively combine spatial and temporal information for reliable detection of shoplifting incidents.

5. Results and Discussion

The performance of the ConvLSTM-based shoplifting detection model has been calculated using standard classification metrics, including *accuracy, precision, recall, and F1 score*. The proposed ConvLSTM model achieves the highest performance across all metrics, with 92% accuracy, 90% precision, 91% recall, and 91% F1-score. In comparison, the CNN-only model achieved an accuracy of 82%, while the LSTM-only model lagged behind at 79%. The hybrid CNN+LSTM improved performance to 86%, but still fell short of the proposed ConvLSTM.

Comparison with Baseline Models

The proposed ConvLSTM model outperformed traditional CNN and LSTM-based models by leveraging both spatial and temporal feature extraction. Unlike standalone CNN models, which primarily capture spatial details, or LSTMs, which process sequential dependencies, ConvLSTM effectively combines both aspects, leading to improved classification performance see table 2.

However, despite its advantages, challenges remain in identifying subtle shoplifting activities. Future work should explore further refinements, including hybrid deep learning approaches and real-time adaptation techniques, to improve the effectiveness of the model.

Table 2: Comparative Analysis of the proposed ConvLSTM model with other models.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN-only	82	80	78	79
LSTM-only	79	77	75	76
CNN+LSTM	86	84	85	84
Proposed ConvLSTM	**92**	**90**	**91**	**91**

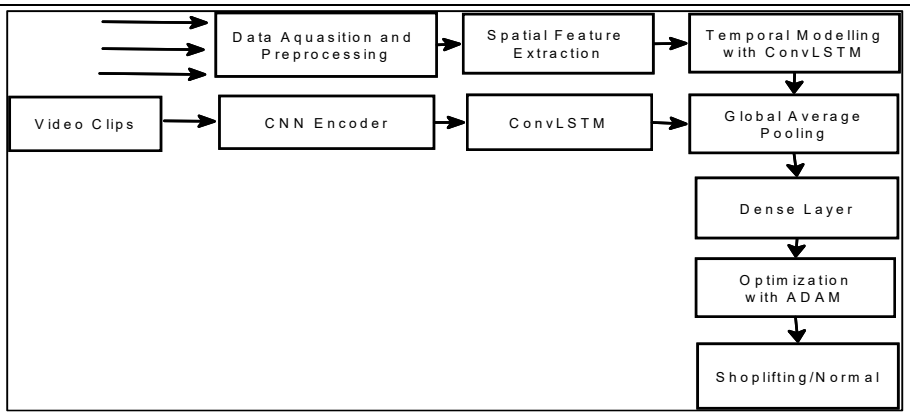

Fig. 2: Block diagram of the proposed shoplifting detection system.

6. Discussion

A. Strengths

The proposed framework demonstrates several strengths that make it highly effective for real-world shoplifting detection. By integrating convolutional layers with ConvLSTM units, the system successfully captures both spatial details and temporal patterns, allowing it to recognize complex behaviours beyond static appearances. The use of the Adam optimizer further strengthens the model by ensuring faster and more stable convergence during training, even with large and high-dimensional data. Robust preprocessing steps, including normalization, augmentation, and class balancing, enhance adaptability to variations in lighting, camera angles, and class imbalance issues. Importantly, the model is validated on

both public and custom surveillance datasets, making it directly relevant to practical retail security scenarios. With higher accuracy and reduced false alarms compared to conventional methods, the framework shows strong potential for real-time deployment and scalability in modern retail environments.

B. Limitations

Although the proposed ConvLSTM framework shows promising results for detecting shoplifting, certain limitations remain. First, the system heavily depends on the quality of surveillance footage. Low-resolution videos, poor lighting, or crowded scenes can reduce the accuracy of feature extraction and make it difficult to distinguish between normal shopping gestures and suspicious actions. Second, the model requires large amounts of annotated training data, particularly for the minority class (shoplifting instances). Since real shoplifting events are relatively rare and often not publicly available, the model may face challenges of class imbalance, leading to potential bias toward normal behavior. Third, while ConvLSTM is effective in modeling spatiotemporal features, it is computationally intensive. Real-time deployment in large retail environments may require high-end GPUs or optimized hardware to process continuous streams efficiently. Finally, the system may generate false positives in scenarios where legitimate customer actions resemble suspicious movements, which could lead to unnecessary alerts. Addressing these issues through improved dataset collection, lightweight model design, and the integration of additional contextual information remains a direction for future work.

C. Future Directions

For the effectiveness of the proposed model in this research paper, the following future improvements are given:

Expanding dataset diversity: Increasing the number of shoplifting cases to reduce the class imbalance between shoplifting and non-shoplifting cases and improve the model's Overall generalization for detection.

Integrating architectures: Exploring more models, to solve the problem faced in detection models, can be combining ConvLSTM with object detection techniques for enhanced performance and giving good result.

Optimizing real-time deployment: Refine computational efficiency like accuracy and recalls for real surveillance applications which can be use in the market.

7. Conclusion

This research presents a ConvLSTM-based framework, optimized with the Adam optimizer, for effective shoplifting detection in retail environments. A major strength of this work lies in the model's ability to achieve **high accuracy (92%)**, outperforming conventional techniques in identifying suspicious behaviors. By combining spatial and temporal analysis, the system can capture even subtle patterns of theft that might otherwise go unnoticed. The integration of data augmentation, oversampling, and dropout layers enhances robustness, reducing errors and ensuring consistent performance across varied conditions such as lighting changes and camera angles. The accurate results obtained in experiments highlight the practical reliability of this framework, making it suitable for real-time retail surveillance. Overall, the study demonstrates how intelligent deep learning approaches can significantly improve security systems, minimize financial losses, and build safer retail spaces.

References

[1] M. Ribeiro, M. Romero, A. E. Lazzaretti, and H. S. Lopes, "Learning Spatio-Temporal Features for Detecting Anomalies in Videos using Convolutional Autoencoder," *Anais do 14. Congresso Brasileiro de In- telige˙ncia Computacional*, 2020. Available: https://api.semanticscholar. org/CorpusID:211196812.

[2] P. Dasari, L. Zhang, Y. Yu, H. Huang, and R. Gao, "Human Action Recognition Using Hybrid Deep Evolving Neural Networks," in *Proc. 2022 Int. Joint Conf. Neural Netw. (IJCNN)*, 2022, pp. 1-8. Available: https://api.semanticscholar.org/CorpusID:252625561.

[3] Chen, Y., & Wang, Y. (2020). Application of AI in retail security. Journal of Retailing and Consumer Services, 57, 102224.

[4] Hollinger, R. C., & Davis, J. L. (2006). Employee theft and shoplifting in retail stores. Security Journal, 19(3), 161-173.

[5] Asif, U., Hassan, M., Jawad, M., Khan, M. Z., & Iqbal, J. (2021). Shoplifting detection using hybrid neural network CNN-BiLSTM and development of benchmark dataset. In 2021 International Conference on Digital Futures and Transformative Technologies (ICDFTT). IEEE.

[6] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Montreal, Canada, 2014, pp. 568-576.

[7] S. Chang, Y. Li, S. Shen, J. Feng, and Z. Zhou, "Contrastive Attention for Video Anomaly Detection", *IEEE Trans. Multimedia*, vol. 24, pp. 4067- 4076, 2022. doi: 10.1109/TMM.2021.3112814. Available: https://github. com/changsn/Contrastive-Attention-for-Video-Anomaly-Detection.

[8] J. T. Zhou, L. Zhang, Z. Fang, J. Du, X. Peng, and Y. Xiao, "Attention- Driven Loss for Anomaly Detection in Video Surveillance," *IEEE Trans. Cir. Syst. Video Technol.*, vol. 30, no. 12, pp. 4639-4647, 2020. doi: 10.1109/TCSVT.2019.2962229.

[9] S. Zhang, M. Gong, Y. Xie, A. K. Qin, H. Li, Y. Gao, and Y.-S. Ong, "Influence-Aware Attention Networks for Anomaly Detection in Surveillance Videos," *IEEE Trans. Cir. Syst. Video Technol.*, vol. 32, no. 8, pp. 5427-5437, 2022. doi: 10.1109/TCSVT.2022.3148392.

[10] N. Nasaruddin, K. Muchtar, A. Afdhal, and A. P. J. Dwiyantoro, "Deep anomaly detection through visual attention in surveillance videos", *J. Big Data*, vol. 7, no. 1, p. 87, 2020. doi: 10.1186/s40537-020-00365-y.

[11] W. Ullah, A. Ullah, I. U. Haq, K. Muhammad, M. Sajjad, and S. W. Baik, "CNN Features with Bi-Directional LSTM for Real-Time Anomaly Detection in Surveillance Networks," *Multimedia Tools Appl.*, vol. 80, no. 11, pp. 16979-16995, 2021. doi: 10.1007/s11042-020-09406-3.

[12] H.-T. Duong, V.-T. Le, and V. T. Hoang, "Deep Learning-Based Anomaly Detection in Video Surveillance: A Survey," Sensors, vol. 23, no. 11, article 5024, 2023. doi: 10.3390/s23115024. Available: https://www.mdpi.com/1424-8220/23/11/5024.

[13] R. Nawaratne, D. Alahakoon, D. De Silva, and X. Yu, "Spatiotemporal Anomaly Detection Using Deep Learning for Real-Time Video Surveil- lance," IEEE Trans. Ind. Inform., vol. 16, no. 1, pp. 393-402, Jan. 2020. doi: 10.1109/TII.2019.2938527.

[14] K. Nithesh, N. Tabassum, D. D. Geetha, and R. D. A. Kumari, "Anomaly Detection in Surveillance Videos Using Deep Learning," in Proc. Int. Conf. Knowledge Eng. Commun. Syst. (ICKES), Dec. 2022, pp. 1-6. doi: 10.1109/ICKECS56523.2022.10059844.

[15] S. Jain and N. Choudhary, "AI Techniques for Anomaly Detection in Video Surveillance Using Deep Learning Method", in Proc. I*nt. Conf. Artif. Intell. Internet Things (AIIoT)*, May 2024, pp. 1-6. doi: 10.1109/AIIoT58432.2024.10574643.
[16] I. Muneer, M. Saddique, Z. Habib, and H. G. Mohamed, "Shoplifting Detection Using Hybrid Neural Network CNN-BiLSMT and Devel-opment of Benchmark Dataset", Appl. Sci., vol. 13, no. 14, article 8341, 2023. doi: 10.3390/app13148341. Available: https://www.mdpi. com/2076-3417/13/14/8341.
[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Proc. Int. Conf. Learn. Representations (ICLR), San Diego, CA, USA, 2015.
[18] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
[19] S. Ruder, "An overview of gradient descent optimization algorithms", arXiv preprint arXiv:1609.04747, 2016.