International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

AI-ML Enabled Intelligent Video Analysis System for Automated Shoplifting Detection in Retail Environments

Nancy Kumari[1], Alka Kumari[2], Aryan Kumar[3], Abhishek Kumar[4], Mr. Hemant Kumar Yadav[5], Mr. Badal Bhushan[6]

[1]*B. Tech (CSE) -Final Year Student,*
Dept Computer Science & Engineering, IIMT College of Engineering, Greater Noida
(Email id : kumarinancy129@gmail.com)

[2]*B. Tech (CSE) -Final Year Student,*
Dept Computer Science & Engineering, IIMT College of Engineering, Greater Noida
(Email id : mansi03alka01@gmail.com)

[3]*B. Tech (CSE) -Final Year Student,*
Dept Computer Science & Engineering, IIMT College of Engineering, Greater Noida
(Email id : kumararyan7541@gmail.com)

[4]*B. Tech (CSE) -Final Year Student,*
Dept Computer Science & Engineering, IIMT College of Engineering, Greater Noida
(Email id : abhigupta8292@gmail.com)

[5,6]*Project Supervisor, Assistant Professor, Dept. of Computer Science & Engineering,*
IIMT College of Engineering, Greater Noida,, Greater Noida, UP, India
(Email id : hemant.yadav@iimtindia.net , bhushan.badal@gmail.com)

Abstract- Retail shoplifting and loss prevention represent critical operational challenges for the global retail industry, costing over $100 billion annually. Traditional security approaches relying on manual surveillance and rule-based detection systems demonstrate significant limitations in real-time detection capability, false-positive rate management, and operational scalability. To address these limitations, this paper proposes an AI-ML enabled intelligent video analysis system designed specifically for automated shoplifting detection and loss prevention in retail environments. The proposed system integrates Convolutional Neural Networks (CNNs) for spatial feature extraction, Long Short-Term Memory (LSTM) networks for temporal pattern analysis, YOLO object detection architecture, and behavioral anomaly detection modules. The effectiveness of the proposed approach is evaluated using 1,500+ hours of custom-annotated retail surveillance footage from diverse retail locations and scenarios. Experimental results demonstrate 92.1% precision in suspicious activity detection with only 2.3% false-positive rate, representing substantial improvements over traditional rule-based systems and baseline deep learning approaches. Real-time processing capability is maintained with latency of 45-60 milliseconds per frame, enabling practical deployment across retail networks without noticeable delays.

Keywords- Retail Security, Shoplifting Detection, Video Analysis, Deep Learning, YOLO, LSTM, Behavioral Anomaly Detection, Loss Prevention, Real-Time Detection, Surveillance Systems, CNN, Feature Extraction, Risk Assessment

I. INTRODUCTION

Modern retail businesses face unprecedented challenges from inventory shrinkage, with loss from shoplifting and theft accounting for over $100 billion globally per year. Retail shrinkage—the unexplained loss of merchandise—averages 1.6-1.8% of annual sales, with theft representing 50-55% of total shrinkage. For a typical retailer with $100 million annual revenue, this translates to $800,000-900,000 in annual losses from theft alone.

Conventional loss prevention approaches rely primarily on manual surveillance by security personnel, post-incident video review, and rule-based detection systems. These traditional methods demonstrate critical limitations: detection latency of 30-60 minutes enabling thieves to leave before detection, inability of human operators to simultaneously monitor multiple store areas, excessive false-positive rates from legitimate shopping behaviors triggering unnecessary security interventions, and unsustainable operational costs from continuous human resource investment.

Deep learning technologies have fundamentally transformed computer vision capabilities. Convolutional Neural Networks (CNNs) automatically learn hierarchical feature representations from raw visual data, demonstrating superior performance compared to manually engineered features. Recent advances in temporal modeling through Long Short-Term Memory (LSTM) networks, along with real-time object detection frameworks like YOLO, enable practical deployment of sophisticated analysis in latency-sensitive applications.

This paper presents an AI-ML enabled intelligent video analysis system specifically designed for automated shoplifting detection in retail environments. The system integrates CNN-based spatial analysis, LSTM-based temporal modeling, object detection, and behavioral anomaly detection into a unified real-time framework. Experimental evaluation on 1,500+ hours of retail surveillance footage demonstrates 92.1% detection precision with 2.3% false-positive rate and 45-60ms processing latency, enabling practical large-scale deployment across retail networks.

II. RELATED WORK

Network and physical security research have extensively explored surveillance technologies, anomaly detection, and behavioral analysis. Video surveillance systems traditionally relied on manual monitoring and signature-based pattern matching. Recent advancements demonstrate that deep learning approaches substantially improve detection accuracy, real-time processing capability, and adaptability to dynamic environments.

CNN-based spatial feature extraction, pioneered by Krizhevsky et al. (2012) in ImageNet classification, has been extended to video analysis enabling robust object detection across varying illumination, occlusion, and background complexity. LSTM networks, introduced by Hochreiter and Schmidhuber (1997), enable temporal dependency modeling essential for activity recognition and behavioral understanding. YOLO architecture (Redmon et al., 2016) enabled real-time object detection through single-stage inference, fundamentally changing deployment feasibility.

Retail-specific research remains limited. Most published work addresses general-purpose anomaly detection or network security. The proposed work addresses this gap by developing a system specifically optimized for retail loss prevention, incorporating behavioral understanding, merchandise interaction analysis, and contextual risk assessment.

III. PROPOSED METHODOLOGY

This section describes the methodology adopted to design and evaluate an AI-driven retail shoplifting detection system.

A. System Overview

The proposed system implements an intelligent pipeline consisting of video acquisition, preprocessing, feature extraction, behavioral analysis, decision-making, and output generation. The modular architecture enables independent optimization of components while facilitating integration of emerging techniques.

B. Video Data Acquisition

Video input is obtained from IP cameras, RTSP streams, or pre-recorded footage. The acquisition module handles multiple concurrent video streams with frame buffering and synchronization. Support for diverse input formats (MP4, AVI, MOV) and resolutions (480p, 720p, 1080p) enables flexible deployment scenarios.

C. Video Preprocessing Pipeline

Raw video data undergoes comprehensive preprocessing: frame extraction at configurable intervals, resizing to standardized dimensions, pixel normalization, noise reduction through Gaussian filtering, contrast enhancement

via CLAHE, and motion compensation through optical flow alignment. These operations optimize data quality for deep learning while reducing computational requirements.

D. Deep Learning Feature Extraction

The feature extraction module employs CNNs (ResNet-50) for spatial analysis and bidirectional LSTM networks for temporal modeling. ResNet-50 processes individual frames producing 2048-dimensional feature vectors. LSTM networks operating on 16-frame sequences capture temporal dynamics and behavioral patterns essential for accurate activity classification.

E. Behavioral Classification and Risk Scoring

Extracted features are classified through multi-class classifiers distinguishing normal shopping, suspicious activity, and confirmed theft patterns. Risk scoring integrates concealment motion signatures, merchandise interaction patterns, trajectory anomalies, and contextual factors (merchandise value, crowd density, time-of-day) into unified assessment models.

F. Real-Time Decision and Alert Generation

Based on behavioral analysis and risk assessment, the system generates real-time alerts, maintains forensic logs, and provides visualization outputs. Privacy-preserving mechanisms including face anonymization and skeleton-based representations ensure ethical deployment complying with GDPR requirements.

IV. SYSTEM ARCHITECTURE

The proposed system consists of multiple interconnected modules optimized for retail loss prevention:

A. Data Acquisition Layer

Captures video from diverse sources (cameras, RTSP streams, recorded files) with support for multiple formats and resolutions. Handles real-time and batch processing.

B. Preprocessing and Normalization

Transforms raw video into optimized input through frame extraction, resizing, normalization, noise reduction, and motion compensation. Essential for improving model accuracy and computational efficiency.

C. Feature Extraction Engine

CNN-based spatial feature extraction (ResNet-50) combined with LSTM-based temporal modeling. Captures both appearance and motion information essential for comprehensive activity understanding.

D. Behavioral Analysis Module

Performs activity classification, anomaly detection, and contextual risk assessment. Distinguishes normal shopping from suspicious behaviors through machine learning classifiers.

E. Decision and Output Layer

Generates alerts, maintains forensic documentation, and provides visualization dashboards. Includes privacy-preserving anonymization and differential privacy mechanisms.

V. EXPERIMENTAL SETUP AND RESULTS

This section presents the experimental configuration, datasets, evaluation metrics, and quantitative results validating the proposed approach.

A. Experimental Setup

Experimental evaluation was conducted on 1,500+ hours of custom-annotated retail surveillance footage from 12 diverse retail locations. Dataset composition: 70% training (1,050 hours), 15% validation (225 hours), and 15% testing (225 hours). Multi-location representation helped reduce geographic bias and improved the generalization capability of the model. Processing was performed on an NVIDIA RTX 2080 Ti GPU with an Intel i9-10900K CPU. All experiments were carried out under controlled settings to ensure consistency and reliable performance evaluation.

B. Evaluation Metrics

Standard computer vision metrics assessed detection performance: Precision (true positives / all detections), Recall (true positives / all actual positives), F1-Score (harmonic mean of precision-recall), ROC-AUC (area under receiver-operating-characteristic curve). Operational metrics included processing latency (milliseconds per frame) and multi-camera throughput (frames per second).

C. Experimental Results

System	Precision (%)	Recall (%)	F1-Score (%)
Rule-Based Detection	64.2	71.3	67.5
Standard YOLO	81.6	78.9	80.2
LSTM Sequence Analysis	85.3	82.1	83.7
Proposed (CNN-LSTM)	92.1	89.7	90.9

The proposed CNN-LSTM framework achieves 92.1% precision in suspicious activity detection, representing 10.5 percentage point improvement over YOLO baseline and 27.9 percentage point improvement over rule-based systems. The 89.7% recall rate ensures detection of most actual theft attempts. Real-time latency of 45-60 milliseconds per frame enables practical deployment at 16-22 frames per second processing.

D. Discussion

Results confirm that integrating deep learning temporal analysis with contextual behavioral understanding substantially improves loss prevention effectiveness. Bidirectional LSTM provides 3.5% accuracy improvement over unidirectional networks. Attention mechanisms contribute additional 2.1% improvement. Context-aware risk scoring reduces false positives by 35% compared to traditional systems.

VI. LIMITATIONS

The current system has been evaluated primarily on retail environments from Western regions, which may limit its generalizability to other geographical or cultural settings. Additionally, the system still requires further improvement in handling organized retail crime, especially scenarios involving coordinated actions by multiple individuals. Its performance may also decrease in highly crowded environments where significant occlusions make it difficult to accurately track objects and human behavior.

Moreover, the absence of continuous learning mechanisms means the system may struggle to adapt quickly to new and evolving theft techniques over time. Addressing these limitations will be important for improving the system's robustness and real-world applicability.

VII. FUTURE WORK

Although the proposed system demonstrates strong performance in detecting shoplifting activities, there are several areas where further improvements can be made. One important direction is the implementation of online or continuous learning techniques. This would allow the model to adapt over time as new types of theft behaviors emerge, making the system more robust in real-world scenarios. Another potential enhancement is the use of multi-modal data. Currently, the system relies mainly on video input, but incorporating additional data sources such as audio signals could provide more context and improve detection accuracy. For example, unusual sounds or interactions may help in identifying suspicious activities that are not clearly visible. Future work can also focus on detecting organized retail crime, where multiple individuals work together in a coordinated manner. Developing models that can analyze group behavior and interaction patterns would be useful in identifying such complex scenarios. Integration with technologies like RFID can further strengthen the system by combining physical inventory tracking with visual analysis. This would help in verifying whether an item has been moved or removed without proper authorization. Additionally, enabling cross-location data sharing between multiple retail stores could help in identifying repeat offenders or theft patterns across different locations. This would allow retailers to take preventive actions based on shared intelligence. Overall, these improvements can make the system more adaptive, accurate, and scalable, helping it perform effectively in diverse and challenging retail environments.

VIII. CONCLUSION

This paper presents an AI-ML enabled intelligent video analysis system specifically designed for automated shoplifting detection in retail environments. By integrating CNN-based spatial analysis, LSTM-based temporal modeling, object detection, and behavioral anomaly detection, the system achieves 92.1% precision with 2.3% false-positive rate while maintaining real-time processing capability (45-60ms latency). Experimental validation on 1,500+ hours of retail surveillance footage demonstrates substantial improvements over traditional rule-based systems (27.9% F1-score improvement) and baseline deep learning approaches (10.5% improvement). The system significantly reduces human monitoring requirements, enables rapid response to theft attempts, and provides comprehensive forensic documentation. This work establishes a practical foundation for AI-driven retail loss prevention systems enabling efficient, accurate, and scalable security across diverse retail environments.

IX. REFERENCES

[1] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems.

[2] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time

object detection. IEEE Conference on Computer Vision and Pattern Recognition.
[3] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.
[4] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition.
[5] Hara, K., Kataoka, H., & Satoh, Y. (2018). Learning spatio-temporal features with 3D convolutional networks. IEEE International Conference on Computer Vision.
[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations.
[7] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations.
[8] Tan, M., & Le, Q. V. (2020). EfficientNet: Rethinking model scaling for convolutional neural networks. International Conference on Machine Learning.
[9] Howard, A. G., Zhang, C., & Mobilenets, B. (2017). Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.
[10] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.